ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

The Encyclopædia Britannica
is published with the editorial advice
of the faculties of the University of Chicago;
a committee of persons holding
academic appointments at the universities
of Oxford, Cambridge, London, and Edinburgh;
a committee at the University of Toronto;
and committees drawn from members of the faculties
of the University of Tokyo
and the Australian National University.



THE UNIVERSITY OF CHICAGO

"Let knowledge grow from more to more
and thus be human life enriched."

The New

Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 5

Knowledge in Depth

FOUNDED 1768
15TH EDITION



Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974
Chicago/London/Toronto/Geneva/Sydney/Tokyo/Manila/Seoul

©1979
by Encyclopædia Britannica, Inc.
Copyright under International Copyright Union
All rights reserved under Pan American and
Universal Copyright Conventions
by Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 77-94292
International Standard Book Number: 0-85229-339-9

Conifer

Conifers are woody plants so named because they bear their seeds upon hard or papery scales arranged in spirals or whorls around an axis, forming a cone. Typically, they are evergreen, upright trees and shrubs, but not all members of this botanical order, Coniferales, meet that description. Some, such as various creeping junipers, never grow more than 30 centimetres tall (about one foot). Others, such as the larches and the cypress, lose their needlelike leaves annually with the approach of autumn. Even the feature of cone bearing seems to be violated by the plum-yews and podocarps, which bear seeds in olive-like "fruits" rather than in cones, and the true yews, which enclose their seeds in a fleshy cup, called an aril. All conifers, however, bear their pollen in smaller, thin-scaled cones, each scale bearing at least two pollen sacs.

General features

Conifers have been of interest to man for several centuries, because many genera are of value as lumber-producing trees. They are important also as subjects for landscaping, for producing forests that hold high recreational value in many parts of the world, and for protecting soil from excessive erosion. The smaller species are of little monetary value except as garden specimens, but many of them are efficient in reducing erosion.

Early in human history, coniferous trees became objects of admiration because of their symmetrical growth and general beauty and important as sources of fuel and of material for construction of shelters. Various peoples held certain trees or groves sacred to their deities. The early inhabitants of India worshipped the deodar cedar and considered their groves the sacred abode of saints, sages, and prophets. The Aztecs venerated various large cypress trees in Mexico, and other peoples similarly regarded local conifers.

The wood of conifers is relatively soft, straight-grained, of even texture, readily worked, and strong under stress. It is, therefore, suitable for many purposes, from general construction, cabinetwork and interior finishing to the manufacture of boxes, crates, and scores of wooden items.

DIVERSITY

Among the smallest conifers known is *Dacrydium laxifolium*, a native of the mountains of New Zealand; mature fruiting specimens of this plant have been found that were scarcely eight centimetres (three inches) tall. Many other conifers are trailing or creeping, such as some junipers, which have been selected and horticulturally developed as ground covers less than 30 centimetres tall. Some species are shrubby, with several stems arising from the root crown, and rarely exceed eight metres (26 feet) in height. Giants of the conifer forest include *Dacrydium cupressinum*, at 55 metres tall (180 feet); redwood (*Sequoia sempervirens*), many up to 90 metres tall (300 feet) with trunks up to six metres (20 feet) or more in diameter; and big tree (*Sequoiadendron giganteum*), rarely reaching 90 metres tall but with trunks up to nine metres (30 feet) in diameter.

The tallest conifers

Most coniferous trees are pyramidal or conical when young, becoming spirelike and then rounded or flat-topped in old age. Many have columnar trunks free of limbs for considerable heights. A few can send up new shoots from stumps after the tree is cut, but most of them are killed and unable to sprout after logging or severe fires have cleared an area.

Massive buttresses support a few conifers; *e.g.*, the swamp cypress (*Taxodium distichum*) and some of the larger species of *Chamaecyparis* and *Cryptomeria*. Such buttressed trees may have a circumference fully three times as great at ground level as at two to three metres (seven to ten feet) higher on the same tree. Others have no apparent swelling at the base but may have radiating roots that anchor the trees against strong winds.

Young trees of the Lawson cypress (*Chamaecyparis lawsoniana*) and several other species have downwardly sweeping branches, the lower ones often resting on the ground at their tips. Most conifers, however, have spreading to ascending branches until the tree is quite old, at which time the top is flat or rounded and the branches, confined to the uppermost parts, mostly spread horizontally. Conifers growing near the timberline, along windswept coastal headlands, and on exposed ridges are often grotesquely twisted by the wind or by the weight of winter-long snow covers. Such stunted Alpine forests are known as elfinwood, or krummholz.

DISTRIBUTION AND ABUNDANCE

Among conifers the pine family (Pinaceae) forms vast forests in the North Temperate Zone in both the Old World and the New World. The pines themselves (*Pinus*) occur mainly in a broad band along the northern, cooler part of the Northern Hemisphere. Their southern extensions occur chiefly along mountain ranges, where they grow generally at progressively higher altitudes as they approach the tropics. Lowland pine forests occur in the southeastern parts of the United States and along parts of coastal areas of the Mediterranean.

The swamp cypress grows near coasts and along rivers and lakeshores inland throughout much of the southeastern United States. In wet or swampy areas, this tree has buttresses and upward extensions from its roots called knees; its wood is remarkably resistant to decay, which accounts for the name "wood everlasting," often applied in the lumber trade.

"Wood everlasting"

Extensive coniferous forests occur in the Rocky Mountains and in ranges paralleling the Pacific coast. Douglas fir (*Pseudotsuga menziesii*) is a valuable timber tree of this area; from northern California well into British Columbia it forms magnificent forests and constitutes the most valuable softwood tree on the continent. This area also produces trees of high economic importance among the firs (*Abies*), larches (*Larix*), spruces (*Picea*), pines (*Pinus*), hemlocks (*Tsuga*), and junipers (*Juniperus*; commonly called cedar). A considerable extent of these forested areas in the Pacific Northwest is managed and harvested by lumber companies on a continuous-yield basis. In arid to semidesert regions flanking the Rockies are vast stands of shrubby or treelike junipers which bind the soil against erosion. Nearer the Pacific Ocean, where rainfall is higher, conifers include incense cedar (*Calocedrus*), Lawson and Nootka cypresses (*Chamaecyparis*), redwood, and white and western red cedars (*Thuja*); all are valuable timber trees and several are planted as ornamentals. The California big tree grows naturally only along the western flank of the Sierra Nevada. All groves of this tree are under government protection, and their cutting is strictly prohibited.

Conifers ranging southward into Latin America include the genera *Abies, Calocedrus, Cupressus, Juniperus, Pinus, Picea,* and *Pseudotsuga* from the Rocky Moun-

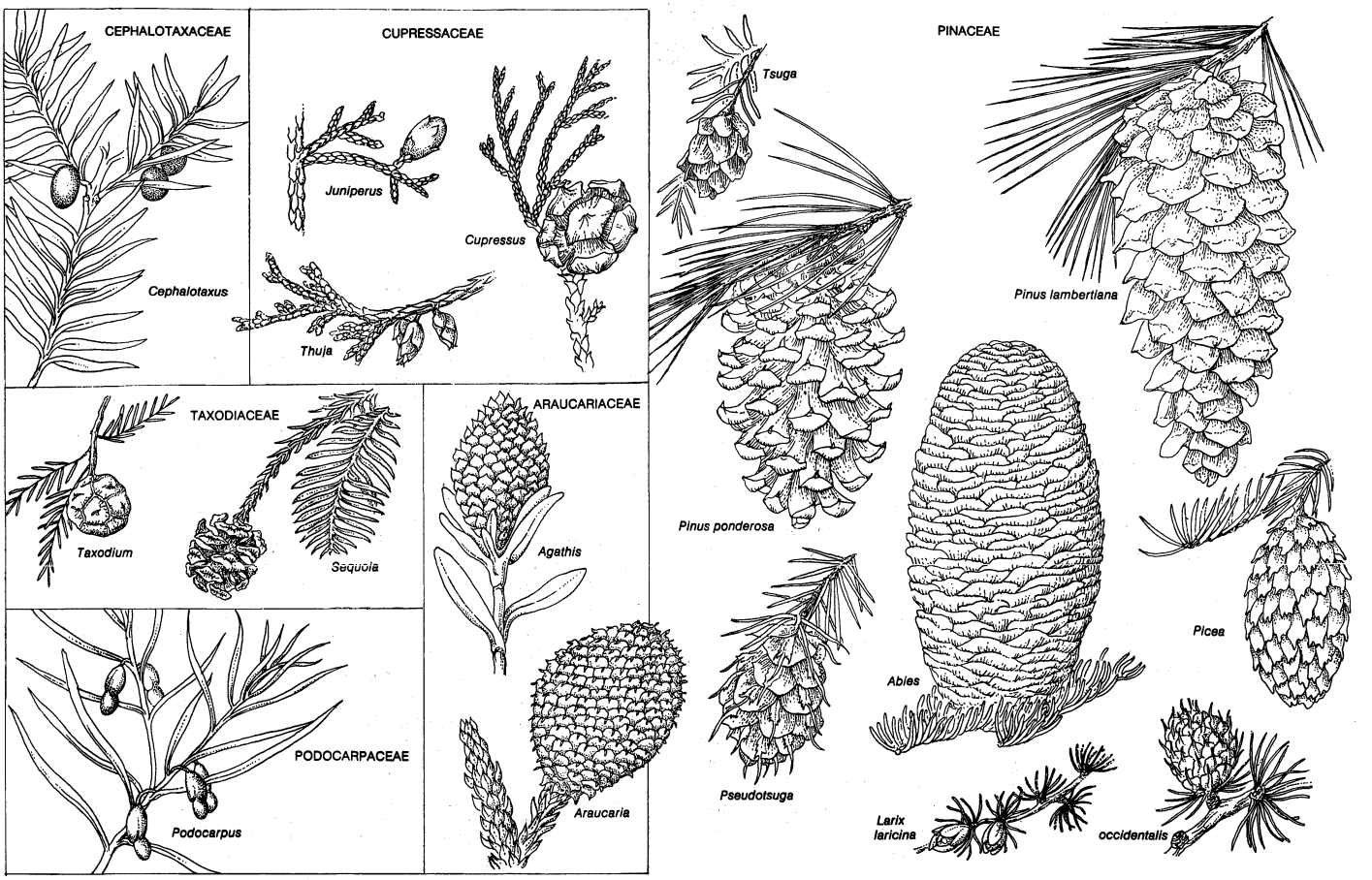


Figure 1: Conifer cones.

Drawing by M. Pahl based on *(Cephalotaxus, Podocarpus, Araucaria) Taxonomy of Vascular Plants* by G.M.H. Lawrence, Copyright 1951 by The Macmillan Company, reprinted with permission of Macmillan Publishing Co., Inc.; *(Larix laricina, Picea, Abies, Taxodium, Cupressus) Trees of North America—A Golden Field Guide* by C. Frank Brockman, illustrations by Rebecca Merrilees © Copyright 1968 by Western Publishing Company, Inc.; *(Sequoia) Illustrated Flora of the Pacific States* by LeRoy Abrams (1923), Stanford University Press; *(Pinus lambertiana, Pseudotsuga) Trees—A Golden Nature Guide* by Herbert S. Zim and Alexander C. Martin, illustrations by Dorothea and Sy Barlowe © Copyright 1956, 1952 by Western Publishing Company, Inc.

tains area and the ahuehuete (*Taxodium mucronatum*).

Genera of conifers now surviving only in Eurasia and northern Africa include *Cedrus* (the mountains of northwestern Africa. Asia Minor, and the Himalayas); *Cephalotaxus* (from Korea to India); *Cryptomeria* (China and Japan); *Cathaya, Fokienia, Metasequoia,* and *Pseudolarix* (China); *Sciadopitys* and *Thujopsis* (Japan); and *Taiwania* (China and Taiwan). *Tetraclinis* is native only in the mountains of North Africa, Malta, and in a small colony in Spain. The dawn redwood (*Metasequoia*) is an especially noteworthy conifer in that it was presumed extinct and described from fossil material in the early 1940s; surprisingly, it was later found in China.

Coniferous genera common to both the Old World and the New World are *Abies, Calocedrus, Chamaecyparis, Cupressus, Juniperus, Larix, Pinus, Pseudotsuga, Thuja,* and *Tsuga.*

The largest genus of conifers

The *Podocarpus* genus, which has its centre of distribution and probably its origin in the New Zealand–Australia–New Caledonia region, has spread widely to Africa, East Asia, South America, Mexico, and the Caribbean. It is the coniferous genus with the largest number of species—more than 110—and with the only known parasitic conifer, *P. ustus,* a small shrub up to one metre (three feet) tall, which lives on the roots of another conifer, *Dacrydium taxoides.*

Callitris, confined to Tasmania and Australia, provides important timber trees and shrubs, such as *C. endlicheri,* favoured for reforestation because it grows on stony ground unfit for agriculture; even trees too small for sawlogs provide a valuable source of tannin in their bark.

Conifers are encountered less frequently in pure stands in the Southern Hemisphere. Most of them are scattered among hardwood forests or as small patches in restricted habitats. Just as *Pinus* and *Pseudotsuga* furnish a high percentage of the timber harvested in the north, kauri (*Agathis australis*), rimu (*Dacrydium cupressinum*), and several species of *Podocarpus* furnish most of the softwood lumber in Australia, New Zealand, the South Pacific insular areas, and Africa.

Australia, New Zealand, New Caledonia, and neighbouring islands support several genera of conifers of interest because of their evolutionary significance, aesthetic value, or unusual adaptations to environmental conditions; such genera are *Acmopyle, Actinostrobus, Araucaria, Athrotaxis, Callitris, Diselma, Microcachrys, Microstrobus, Neocallitropsis, Papuacedrus,* and *Phyllocladus.* Several of the above-mentioned genera produce timber trees, some in such rugged terrain that it is uneconomical to transport logs to sawmills.

South America and Africa are relatively poor in conifers. South America has two species of *Araucaria* and one species each of *Austrocedrus* and *Fitzroya,* both genera producing trees suitable for lumber in Argentina and Chile. Two other monotypic (or single species) genera, *Pilgerodendron* (Chile) and *Saxegothaea* (Chile and Patagonia), are valued, the first as timber trees, the second because it is a connecting link between the families Podocarpaceae and Araucariaceae.

Taiwan's coniferous stands

A remarkable concentration of conifers occurs on the island of Taiwan, which has native representatives from five of the six coniferous families occurring in nearly pure stands at altitudes between 5,900 and 7,300 feet. Most of the lumber-producing trees occur on rugged mountain slopes. A species of conifer discovered on the island in 1906 was named *Taiwania.*

ECONOMIC IMPORTANCE

The softwood timbers provide almost 75 percent of the commercial lumber used for general construction, mine

timbers, fence posts, poles, boxes and crates, and lesser articles. During an average year, Canada, for example, produces from its coniferous forests sawed lumber valued at almost $500,000,000, pulpwood worth $650,000,000, and paper for newsprint valued at another $600,500,000. The United States produces considerably more cut timber but less pulpwood. Europe's production is somewhat lower. Finland, Norway, and Sweden produce large quantities of coniferous logs for pulping annually without seriously depleting their forests because of their forest-management practice of cutting trees of specified age classes on a rotating schedule.

In addition to timber used in construction, the world's nations consume huge quantities of coniferous wood as fuel and in the manufacture of cellulose products, plywood, and veneers.

Coniferous trees also are the sources of valuable resins, volatile oils, turpentine, tars, and pharmaceutical products. Many of these products are used in the manufacture of varnishes, paints, greases, and soap. *Pinus* is the chief source of raw pitch, which is treated by steam distillation to yield turpentine and resin. Kauri gum, used chiefly in fine varnishes and linoleum, is obtained from accumulations on large limbs, around the bases of the trunks of *Agathis australis*, and by digging fossil gum from beneath the soil in boggy areas (these trees are native only on New Zealand's North Island, where tapping for gum has been made illegal). A similar resin, Manila copal, comes from *A. alba*, native in the Philippines (there, too, tapping or cutting of the trees is now strictly forbidden).

Natural history

LIFE CYCLE

Alternation of generations

The conifers, like all seed plants, exhibit an alternation of generations (see Figure 2) between an asexual phase (sporophyte) and a sexual phase (gametophyte). The familiar coniferous tree is the dominant, conspicuous sporophyte generation, which bears two kinds of cones, both kinds on the same tree (and thus is termed monoecious) in most conifers, on separate trees (dioecious) in a few. The smaller of the cones produces the pollen grains, technically called the male gametophytes, or microgametophytes. The larger, often woody cones bear female gametophytes, or megagametophytes, on the upper side of the cone scales. The nucleus of each living cell in a sporophyte contains a double set of chromosomes—the bodies that transmit genetic characteristics from parent to offspring—and the sporophytes are thus known as the diploid ($2n$) generation in the life cycle. Cells called spore mother cells—in pollen sacs on the male, or staminate, cones and in ovules on the female, or ovulate, cones—undergo a special cell division called meiosis, whereby two successive divisions of a spore mother cell produce four nuclei, each with half as many chromosomes as the spore-mother-cell nucleus, the haploid (n) number. Thus, the spore mother cell is the last diploid member of the sporophyte phase in the alternation of generations, and the microspores and megaspores are the first cells of the alternate, or gametophyte, phase.

In the staminate cone the microspores separate, each developing a spore coat consisting of an outer (exine) and an inner (intine) layer. In many conifers the exine balloons out to form two or three hollow wings, which increase the buoyancy of the pollen grain (microgametophyte) and facilitate transportation by air currents. Simultaneously with the formation of the wings, a series of divisions begins, which sooner or later produces two male gametes, or sex cells, ready to fertilize the egg nucleus when the latter is receptive. When the pollen is shed, each grain is usually made up of two or more prothallial cells (flattened cells with degenerate nuclei, the only remaining vegetative cells of the male gametophyte generation), a tube nucleus, and a generative nucleus. In some genera no prothallial cells form, and tube and generative nuclei develop after the pollen grain lodges in the opening (micropyle) of an ovule. Pollen grains usually are released in the spring and float long distances on air currents. Only a minute fraction of them lodge between the scales of ovulate cones; those that do may germinate and ultimately fertilize the egg nucleus. In *Pinus* and other genera that require two years from appearance of young ovulate cones to produce ripe seeds, the pollen lodges on the ovulate scales one spring, and the pollen tube grows part way toward the egg nucleus, then remains quiescent through the winter; it fertilizes the egg nucleus in the spring of the following year. In those that produce seed in one year, the processes are telescoped, and all stages from pollination to ripe seed occur in one spring–summer–autumn period.

The female strobilus (ovule-bearing structure) is an obvious cone in all members of the families Araucariaceae, Cupressaceae, Pinaceae, and Taxodiaceae, in spite of the different appearances among them. The nature of the ovule-bearing structures in *Cephalotaxus* and in the family Podocarpaceae is less apparent, but they are greatly reduced cones that have lost most of their scales and undergone fusion of remaining parts.

Conifers with less obvious cones

The ovule is that part of the ovulate cone that produces the female gametophyte (megagametophyte). The number of ovules per scale varies; a single functional one appears on each fertile scale in the Araucariaceae, Cephalotaxaceae, and Podocarpaceae. Two ovules occur on each scale in the Pinaceae, and from two to nine are found among the Cupressaceae and Taxodiaceae.

In the early stages of development, an ovule consists of a spherical or ellipsoidal mass of cells on the upper face of a cone scale. This mass is made up of food-storage tissue (nucellus) that occupies the bulk of the ovule and is surrounded by a coat, or integument, two to several cells thick. The integument is free from but lies closely against the nucellus, except at the apex, where a small opening (the micropyle) leads to the ovule.

Drawing by M. Pahl based on (A, B, C, D, E, H) B. Lloyd, *Handbook of Botanical Diagrams* (1962), University of London Press Ltd.; and (F, G) Harold C. Bold, *The Plant Kingdom*, 2nd ed., © 1964, reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

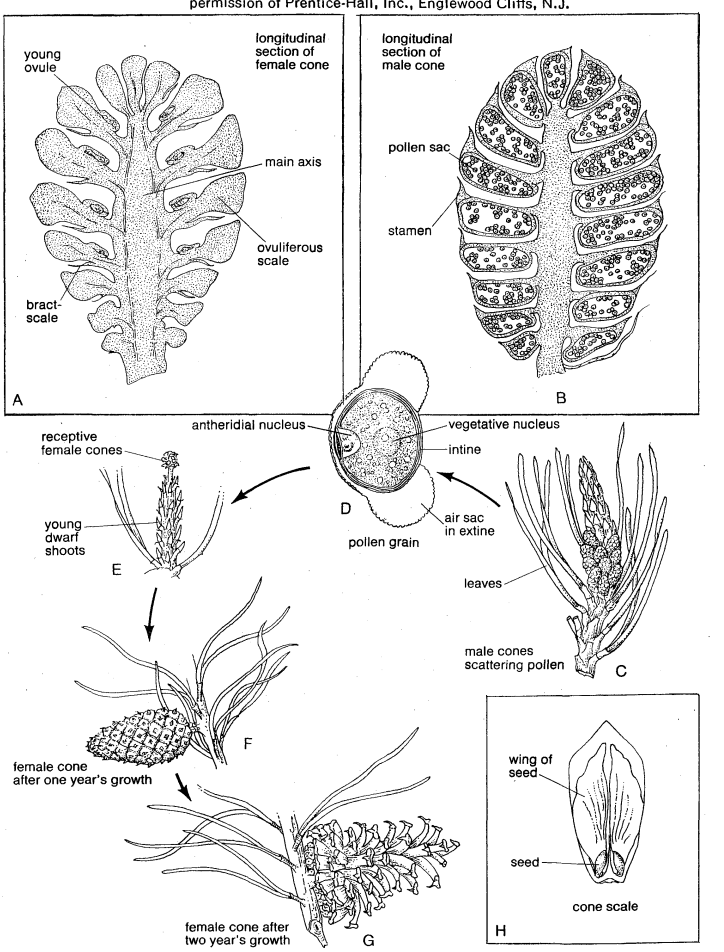


Figure 2: Reproductive structures and events in male and female cones.

Ovulate cones begin to develop in the fall, remain quiescent during winter, and have ovules ready for pollination in late spring. At this time the ovulate cones spread the tips of their scales apart, allowing pollen grains wafting on the breeze to sift among them. The ovules often have a small drop of liquid at the mouth of the micropyle at this time; pollen grains entrapped in the droplet are drawn toward the top of the nucellus as the liquid evaporates. The pollen tube then digests its way through the nucellus toward the archegonium, a flask-shaped structure in which the egg cell is developing. The pollen tube forces its way into the archegonium and extrudes its contents into the egg cell. A male-gamete nucleus soon fuses with the egg nucleus. The fusion of the male and female gametes is a complex process that produces two diploid nuclei in the earliest stage of the embryo. The two nuclei constitute the first cell generation of the next sporophyte generation in the alternation cycle.

SEED DEVELOPMENT

The rate of development accelerates markedly after fusion of the gamete nuclei. A complicated set of events follows, and several embryos are established. This condition, called polyembryony, is universal among conifers, but competition among the several to numerous embryos generally results in elimination of all but one; only rarely do two or more viable embryos develop in a single seed. The seed may be shed soon after maturity or, in closed-cone pines, may remain inside the cones on the tree from several to many years (as long as 30 years in *Pinus radiata*, *P. attenuata*, and *P. muricata*). Most conifers shed their seed at the end of the summer in which fertilization occurred.

Seeds that are shed in the fall lie dormant through the winter, and only about 20 percent of an average seed crop germinates with the arrival of spring.

Seedling mortality

Seedling mortality is high among conifers—approaching 100 percent for some species. Although thousands of seedlings often appear in early summer on land recently burned-over or cleared in construction projects, only a minute percentage of them live through the first summer. The greatest loss is attributable to drought during early growth stages, to excessive heating when in full sun, to the depredations of browsing animals, and to attacks from insect and fungus pests. Competition for space eliminates others, particularly when the early stand is dense.

ECOLOGY

As indicated earlier, conifers may often occur as extensive forests of a single species, but they are also found in ecological associations with other conifers or in a mixture of conifers and deciduous, broad-leaved hardwoods. Some occupy arid slopes, as do the piñon pines (*Pinus edulis*, *P. monophylla*, and *P. quadrifolia*); others grow in dense swamps or boggy land, as do the swamp cypress and the black spruce (*Picea mariana*). *Dacrydium araucarioides*, which grows only in New Caledonia, generally occupies dry situations on serpentine rocks.

Certain species of *Podocarpus* grow only in Alpine habitats in the Southern Hemisphere at elevations of 5,000 feet (1,500 metres) or more above sea level: *P. andinus* in the South American Andes and *P. macrophyllus* in Yünnan, China, are adapted to such habitats. Others occupy rain-drenched slopes and canyons at moderate levels, while a few grow near sea level only. One of the lowland species, *Podocarpus dacrydioides*, is well adapted to swampy ground and holds promise for use in reforesting New Zealand swamplands.

The closed-cone pines listed earlier have become adjusted to long dry periods and irregularly spaced fires sweeping through the forests. Under such conditions a tree may die following a particularly hot fire, but its seeds, protected from excessive heat by the closely appressed, woody cone scales, remain unscathed and are released in large numbers a few weeks after the fire, thereby reseeding the ash-enriched terrain.

Other conifers are tolerant of very low temperatures. Those that grow at timberline and near the Arctic Circle are undamaged by temperatures that drop to about −45° C (−49° F). In contrast, many conifers in the Southern Hemisphere—*e.g.*, tender species of *Araucaria*, *Dacrydium*, and *Podocarpus*—cannot survive even light frosts.

Hazards to conifers

Coniferous trees, like other trees, suffer from violent storms, destructive fires, and man-caused disturbances such as stream diversion, dam building, wasteful logging, and ill-advised agricultural exploitation. In recent years serious injuries have been inflicted by air pollution, in many areas bringing about the death of pure stands of such conifers as white pine. Smelter fumes are particularly injurious and have caused denudation of large areas of conifer forest. Areas cleared to provide agricultural land sometimes result in excessive soil erosion, so the land as well as the forest is destroyed.

Numerous insects and fungi may be extremely destructive of conifers. The white-pine blister rust is a serious menace in North America, and attempts to control it by eradicating the alternate hosts, the gooseberry and currant bushes (*Ribes* species generally), have been only partly successful. Bud-boring insects damage shoots, and other insects attack needles, the bark, and the wood. Under constant, often virulent attacks, conifer forests—especially pure stands, because of the ease with which disease and insects spread through such stands—have suffered greatly.

Form and function

THE ROOT

Many conifer seedlings develop a strong taproot after germination and retain it throughout the life of the tree. Others, especially trees growing in waterlogged soil, send out shallow lateral roots, the taproot dying shortly after the spreading roots are established. All conifers develop short roots from their main laterals or taproots. These short roots, which branch profusely, are the absorptive organs; they are abundantly supplied with mycorrhiza, fungi in a biological relationship with the root cells, which assist in absorbing dissolved minerals.

The internal anatomy of conifer roots is fairly uniform. No bark-forming tissue is apparent in the young roots of conifers. A cross section a short distance back of the tip of a young root shows an epidermis one cell thick, from some cells of which stubby root hairs grow outward. Interior to the epidermis is a layer of cortex, several cells thick, with many of the cells containing mycorrhiza, while others may contain starch grains, oil droplets, and other food reserves. A central strand of tissue comprises the vascular elements of the root, and its outermost sheath of cells, the endodermis, separates the vascular bundles from the cortex. Resin ducts are often seen in the roots.

Soon after the initial tissues become established—that is, the phloem, the tissue that functions in transporting foods manufactured in photosynthetic regions to other parts of the plant, and xylem, the woody parts of the water vascular system—arcs of cambium (dividing) cells become differentiated in the general ground tissue (parenchyma) and add secondary xylem cells to the woody part of the root and new phloem cells exterior to the cambial arcs. This activity soon fills the regions between the arms of protoxylem and produces a rodlike body of xylem consisting mostly of tracheids—long, slender cells with pointed ends and with perforations (bordered pits) in their thickened walls. The tracheids soon become empty conducting elements, having lost their contents during the deposition of thickenings on their walls. These empty elements serve two functions: mechanical strength and transportation of water with its dissolved mineral salts upward from the roots.

Functions of the tracheids

Vascular rays, present in roots in which secondary xylem has been deposited, consist of cells usually arranged in thin plates or bands one cell thick and several to many cells high that run radially from the xylem outward through the phloem. In many conifers the rays consist of thin-walled parenchyma cells only. In other conifers some rays have modified tracheids along the upper and

lower margin, with the parenchyma cells constituting the central part of the strip. The number of rays increases as the stem or root grows in diameter.

THE SHOOT

The seedling shoot (hypocotyl) has the vascular tissues arranged in the same manner as in a root at its lower end, then has a transition region in which parenchyma cells come to occupy the centre of the vascular complex and above which the vascular elements are arranged as in stems. Stomates ("breathing holes") occur in the epidermis of the hypocotyl but are fewer than on leaves.

The growing tip of a conifer stem differs from that of its root in lacking a protective cap, lacking root hairs, and in possessing cells that are potential buds a short distance back of the apex.

A cross-section slice of a two-year-old stem shows an epidermis, usually ruptured at intervals by slit openings called lenticels, which permit the exchange of gases between the interior and the air outside. The centre of a cross section of a young stem consists of thin-walled unspecialized cells, the pith, with patches of xylem (usually three to five) constituting the primary xylem. Exterior to the outermost xylem elements is secondary xylem, which at first consists of discrete bundles but very shortly forms a cylinder of tracheids interrupted only by the medullary rays—rays that run from the pith outward. New vascular rays develop as the stem grows in diameter. Outside the ring of xylem is the cambium, which functions exactly as in roots, cutting off xylem elements to the interior and phloem elements to the exterior.

Resin ducts

Scattered through the xylem in several conifer genera, particularly in *Agathis*, *Larix*, *Picea*, *Pinus*, and *Pseudotsuga*, are resin ducts running longitudinally in the secondary xylem, each surrounded by small, thin-walled cells that exude resin (pitch) into the duct; it is these ducts that supply resin when the tree is tapped or the bark injured. Also present in the xylem of many conifers are scattered parenchyma cells that retain their contents for some time after the adjacent tracheid elements are empty. Parenchyma cells are much shorter than tracheids, have blunt ends, and have simpler perforations in their walls. The vascular and medullary rays in conifers are usually one cell wide and a few cells high (when viewing them in tangential sections of the stem). All conifers that normally have resin ducts in their xylem have occasional fusiform rays; *i.e.*, rays that are spindle shaped in tangential view, several cells wide and three to six times as high as wide. Running longitudinally through the central part of many fusiform rays is a resin duct similar to those parallel to the tracheids. As in roots, the vascular and medullary rays in many conifers are composed of parenchyma cells only. A few, such as *Sequoia*, at times develop ray tracheids along the upper or lower margins or in both positions. Some conifers that normally lack resin ducts—*e.g.*, *Abies*—may develop some in the vicinity of injuries.

LEAVES

Most conifer leaves are relatively narrow (several times longer than wide), usually have stomates arranged in longitudinal rows along either side of the single vein, or vascular bundle, and have a protective waxy cuticle. The vein, in all except a few genera, consists of one or two centrally located strands of vascular cells surrounded by a few parenchyma cells, outside of which is the endodermis. Between the endodermis and the epidermis are parenchyma cells, special transport cells (forming transfusion tissue) along the flanks of the strand, and, exterior to these latter cells, the palisade cells packed with the green-pigment bodies, or chloroplasts, that carry on photosynthesis. Some conifers have cylindrical palisade cells; others, as in *Pinus*, have almost spherical ones with deep folds in their walls, thus increasing the surface through which exchange of gases takes place. Immediately under the epidermis are strands of varying size and thickness composed of rigid, thick-walled cells; the strands constitute the hypodermis, a tissue that adds mechanical strength and stiffness to the leaf. Resin ducts, usually numbering from one to five in most species, vary in length and location, depending upon the species.

Needle-like and scale-like leaves

The leaves of many conifers—especially pines, firs, and spruces—are long and stiff and hence are called needles. Leaves of others—cypresses, cedars, *Calocedrus*, *Chamaecyparis*, and *Cryptomeria*, for example—are smaller and scalelike and have much of their length attached firmly to the twig, with only their tips free. Scale leaves are arranged along the branch in pairs or threes, with alternate sets at right angles to each other so as to form four to six distinct longitudinal rows. In some species the scale leaves take two forms: a longer pair set in one plane and a shorter pair, less spreading and appressed to the twig, set in the other plane, an arrangement that often results in fanlike, flattened branchlets, as in the cedars. In contrast, species with scale leaves all of similar size usually have leafy twigs that are circular in cross section.

Pinus bears two kinds of needles: simple, solitary ones set spirally on the branch of very young seedlings and longer, stiffer ones borne in bundles of two to five on a short stub, or spur branch, with each such bundle ensheathed at its base by papery scales. Each bundle persists from two to 20 years, depending on the species; then the entire spur branch falls, carrying the needles with it.

One conifer, *Phyllocladus*, carries on photosynthesis in leaflike flattened branchlets, its true leaves being minute scales that fall soon after they appear. *Sciadopitys* has two anatomically identical leaves united side by side and borne at the tip of a minute spur that is formed in the angle of a temporary scale leaf and the twig.

Leaves of most species of *Podocarpus*, those of *Cephalotaxus* and *Cunninghamia*, and some species of *Araucaria* are intermediate between needles and scale leaves. They usually are oblong to broadly linear, stiff and leathery, and often have sharp ends. They are spirally arranged, but some of them have twisted bases so that they appear to grow on opposite sides of the twig. Some have several veins running from base to apex rather than a single, middle vein. Several genera in the family Taxodiaceae bear leaves on lateral branchlets that function through one to several seasons, then shed the entire branchlet. *Taxodium* and *Metasequoia* shed such branchlets annually, leaving the trees bare during the dormant season; while *Sequoia* and *Sequoiadendron* retain the branchlets several years and are evergreen.

In some conifers, stomates are confined to the lower surface of the leaves, but in others they may occur on all surfaces with a concentration on the undersurface.

GROWTH AND BIOLOGICAL PRODUCTION

Conifers require a much lower concentration of mineral nutrients than is needed by food crops. Minute quantities of iron are essential, and traces of zinc and several other minerals are needed. (Zinc deficiency caused considerable loss among plantations of Monterey pine [*Pinus radiata*] during the early phases of reforestation in Australia.) Potassium-deficiency symptoms appear in young conifers only when concentration of that mineral falls below four parts per million, and calcium deficiency occurs only when that mineral drops below three parts per million. Mineral deficiencies sometimes produce no apparent symptoms other than reduction in rate of growth, but changes in colour are often indicative of a low level of vital nutrients. Nitrogen deficiency—as might occur in low wetlands—stunts seedlings and causes their leaves to turn yellow; phosphorus-deficient plants turn purple; and potassium and iron deficiencies cause loss of chlorophyll in conifer leaves.

Nearly 50 species of fungi enter into symbiotic relationships with the roots of conifers as mycorrhiza. Mycorrhiza may afford some protection against pathogenic fungi, and they definitely assist the conifer host in drawing up certain minerals and increase the rate of absorption of water from relatively dry soils.

Most conifers are able to carry on photosynthesis in low light intensity; thus, young plants can survive under rather shady conditions and in light near the blue end

of the spectrum. Many conifers carry on photosynthesis more efficiently than broad-leaved hardwoods because their needles have a relatively higher concentration of chlorophyll. Conifers in cool areas continue photosynthetic processes at temperatures as low as −6° C (about 21° F); few flowering plants can approach such a performance. Conifer seedlings grow at maximum efficiency under long-day photoperiods—light for 15–20 hours daily—but cease growing when the 10-hour days of autumn begin. Of course, different photoperiod reactions occur, depending upon the latitude and on species adaptation to environmental factors normal to their respective habitats. Furthermore, growth of conifers is most rapid under conditions of warm days and cool nights, with the greatest growth occurring among those subjected to the greatest differences, within their tolerances, between day and night temperatures.

Conditions for maximum growth

Not only is there an annual growth rhythm, resulting in annual growth rings, but there is also a strong daily rhythm, with more shoot elongation occurring at night than during daylight hours. In semi-arid regions, where precipitation during the growing season often is in the form of brief but violent rainstorms, several growth rings may form in a conifer during a single summer. Such alternate growth surges and low activity are directly correlated with the intermittent rains but lag appreciably behind the actual occurrence of the storms. Where the water supply is uniformly high and the temperatures uniformly warm, however, conifers produce almost no visible growth rings, since growth is fairly uniform.

Physiological factors associated with the initiation of cone production are puzzling, but genetic factors are definitely involved. The average age of first cone formation of 60 conifer species is 5.2 years. A few species produce some ovulate cones at two years of age; others, such as the sugar pine (*Pinus lambertiana*), bear no cones until 25 years old. A number of species have a nonproductive period of several years between the first crop of cones and the next. Among many conifers seasons of high seed production alternate with seasons of low production.

Longevity and senescence are equally difficult to understand. The ability to root from cuttings appears to decline with age. The terminal shoot of a tree four or five years old roots easily when cut and placed in a rooting medium. In contrast, a cutting from the top of a pine tree 200 years old will not root, even if treated with potent growth-promoting substances. The conifers with the longest life-span discovered to date—and the oldest plants in existence—are the bristlecone pines (*Pinus aristata*), growing at 11,000 feet (3,400 metres) above sea level in the White Mountains of California and similar areas in Nevada, Utah, Colorado, and northern Arizona and New Mexico. Ages of almost 5,000 years have been determined by counting growth rings in borings taken from the thick trunks of these gnarled and twisted trees. Conifers that produce tall, sturdy trunks rarely attain ages one-half as great.

The oldest plants

Seeds of conifers vary greatly in viability (ability to germinate) and in longevity. The seeds of pines in the subgenus *Haploxylon*—typified by white pines—lose their viability very rapidly if stored at room temperature, much faster than do those of the subgenus *Diploxylon*, or yellow pines, under similar conditions. Seeds of all conifers that grow in temperate climates or in sub-Arctic areas retain their viability best if stored at a temperature of about 5° C (41° F). Seeds of closed-cone pines have remained viable up to 30 years when contained within the cones held on the trees; they deteriorate rapidly, however, after they are released from the protective cones. Some conifer seeds germinate more rapidly when subjected to at least a short period of illumination, others do best if left in light throughout each day, and a few react favourably to short periods of illumination with red light.

CHEMICAL COMPOSITION

The chemistry of pines is more thoroughly known than that of any other conifer because of the great amount of investigation conducted by industries that utilize pine products.

Cellulose is a component of cell walls in all higher plants, including conifers; lignin occurs in xylem, or wood, and is more complex in its structure. Coniferous lignin differs from the lignin of broad-leaved angiospermous trees in giving no reaction to a test for the chemical syringaldehyde.

The hemicellulose present in conifers can be made to yield such sugars as xylose, mannose, arabinose, galactose, glucose, and rhamnose and the urinelike uronic acids. Conifer wood yields only 15 to 20 percent of hemicelluloses, compared with 20 to 30 percent in angiosperm wood.

The sugar pine exudes a sugary substance that the American Indians used as food and medicine. First named pinite, later pinitol, it has been found in the wood of six species of pines in the subgenus *Haploxylon* but not in the wood of any of the pines of *Diploxylon*.

Polyphenolic compounds occur in the wood of many conifers, but percentages of separate phenols vary greatly from sample to sample. Some polyphenols are abundantly present in the heartwood, or nonliving core, of conifers, with only minimal fractions of them in the sapwood, bark, needles, cones, and pollen grains.

As much as 50 percent of the weight of conifer seeds is fat, which constitutes the main food reserve for the embryo; the fats are mainly triglycerides of unsaturated fatty acids such as oleic, linoleic, and linolenic, predominantly the second. Seeds of pines growing in northern areas contain somewhat higher percentages of unsaturated fats than do those of trees from southern, warmer areas.

Waxes present in the bark, on needles and twigs, and in the pollen of some conifers often impart a whitish or bluish "bloom." These waxes generally are classified as estolides and are made up of long chains of polyesters of hydroxy acids.

Volatile oils occur in the leaves, twigs, wood, and to a lesser degree in the bark of conifers. They usually constitute only about 0.5 percent of the fresh weight of the material processed. The chemical composition and relative volatility of these oils vary greatly, that obtained from needles being different from that from the twigs and older stem wood in the same tree.

Volatile oils

Tall oil, a mixture of substances obtained as a by-product of paper-pulp manufacture, is used chiefly in the manufacture of soaps and greases.

Oleoresins present in many conifers form the basis of the naval-stores industry in several parts of the world. Droplets of oleoresins occur chiefly in cells surrounding the resin ducts in the sapwood. When the resin canals are severed or ruptured, the pitch seeps through the wound and may be collected in a suitable receptacle. Raw oleoresin can be separated into two main components, turpentine and rosin, the relative percentages of which vary considerably among the pitch-producing pines and other conifers. Oleoresins also collect in pockets in the bark of several species of *Abies*, although the wood of this genus lacks resin ducts. Each "blister" must be tapped individually, and the product, Canada balsam, is produced in small quantities and has limited uses.

Venetian turpentine, used mainly in artwork, is an oleoresin taken from the heartwood of the European larch (*Larix decidua*) by boring a hole in the lower trunk, allowing the pitch to accumulate, and collecting it periodically. The annual production is very low, usually not over two ounces of pitch per tree per year.

Oleoresins obtained from stumps and collected by tapping living trees yield rosin generally composed of about 90 percent diterpene resin and acids and up to 10 percent of nonacid compounds. The steam-volatile fraction of oleoresins from pines, called turpentine, is a complex of about 30 chemical constituents.

Evolution and paleontology

ANCESTRAL CONIFERS

The geological history of the conifers began in the Carboniferous period (about 345,000,000 years ago). The

earliest conifers represented in the fossil record have been lumped, for convenience, in the family Voltziaceae, of uncertain relationship to still earlier groups of plants (see Figure 3). Within the Voltziaceae several "form genera" (those based on a general type of organ, such as a leaf, a twig, unidentifiable beyond these general lines) have been included in a newly proposed genus, *Lebachia*, and one older genus, *Walchia*, the latter retained to include all leaf-bearing twigs in which stomatal characters cannot be determined.

"Form genera" of the Voltziaceae

Lebachia, a straight, upright tree with a slender trunk of unknown height, appeared first in the late Carboniferous time (about 300,000,000 years ago) and became commoner and more widespread later. Both staminate and ovulate cones of *Lebachia* were cylindrical and were borne singly at the tips of twigs on the same tree.

In the Permian Period (from 225,000,000 to 280,000,000 years ago) occurred *Ernestiodendron filiciforme*, vegetatively similar to *Lebachia* but evolutionarily advanced over *Lebachia*. A further evolutionary advance is shown by *Pseudovoltzia liebeana*. It had two kinds of leaves, spirally arranged, those at the ends of the younger branches being long, linear, and flattened, while those on the older branches were shorter, more slender, and incurved. Scales of the ovuliferous cones were arranged spirally but had a flattened, five-lobed appendage axillary to each bract except the basal two or three. These appendages have been interpreted as having arisen by fusion of five ovuliferous scales such as those in *Lebachia*, their terminal parts still free, and a loss of the ovule from at least two of the five scales. Continued reduction, or fusion, through a greater part of the scales, plus loss of one more ovule, could have resulted in an ovulate cone scale similar to those now possessed by many modern conifers.

The foregoing series of fossils presents a clear line of development from ancestral forms to the immediate predecessors of modern conifers. This evolutionary line apparently developed rapidly during the Mesozoic Era (from 65,000,000 to 225,000,000 years ago), but few fossils of that age are adequately preserved, so it is difficult to follow advances in later differentiation.

From C. Arnold, *An Introduction to Paleobotany*, copyright 1947, used with permission of McGraw-Hill Book Company

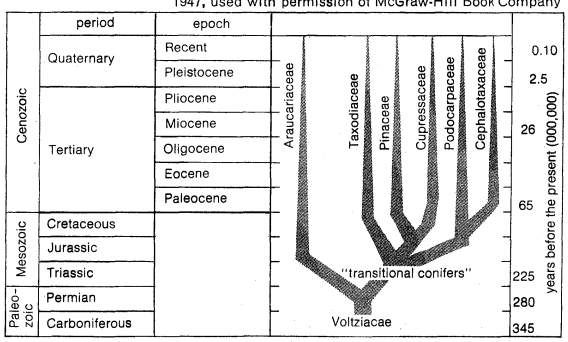


Figure 3: Conifer dendrogram.

FOSSILS OF CONTEMPORARY CONIFERS

It is difficult to identify any modern coniferous genera with fossils older than those from the Cretaceous Period (from 65,000,000 to 136,000,000 years ago).

Early in the 20th century, some botanists argued that the family Pinaceae represented the most primitive stock of the coniferous evolutionary sequence and were followed by the family Araucariaceae. The general opinion has become that both families are descendants of an ancient stock—possibly going back to the *Pityeae*—that they represent two lines that became differentiated in the Middle and Late Mesozoic and that the present segregation of the two—and other families of the Coniferales—is chiefly the result of the extinction of intermediate, Mesozoic types. The conifers as a group reached their zenith in the Middle Cretaceous and have been declining ever since. The angiosperms have been more successful in the never-ending competition among organisms and have come to dominate the floristic scene.

The decline of the conifers

The diagram in Figure 3 of the theoretical lines of descent among the gymnosperms shows the presumed general routes of evolution followed by the different branches of the groups in geologic time.

The oldest fossil referred to the family Araucariaceae with reasonable assurance is *Araucarites delafondii* from the Permian of France. It consists only of triangular cone scales that had separated from the cone axis, with each scale bearing a single seed on its undersurface.

Extensive deposits containing *Araucarioxylon* and *Woodworthia* wood occur in North America, the most famous being those in the Petrified Forest National Park in Arizona. No cones are attached to the petrified logs, but the tracheids have pit characteristics of *Araucaria*, and there is no doubt about the relationship. Several species of *Araucarites* occur in Cretaceous sediments in Alabama, New Jersey, North Carolina, South Dakota, and Wyoming.

Excellent fossil cones of *Proaraucaria mirabilis* came to light in volcanic ash, probably Eocene in age (from 38,000,000 to 54,000,000 years ago), in Patagonia. These cones differ from modern *Araucaria* cones in having a deep cleft between the tips of the bract and the ovuliferous scale, a condition similar to that found in modern members of the Pinaceae. Only one seed appears on a scale, and the structures of both scale and seed strongly resemble those in the *Eutacta* section of *Araucaria*.

Cone scales of *Agathis* are fairly common in Cretaceous and Tertiary sediments. The first specimens found were *A. borealis* from Arctic deposits, but subsequently others were found in Cretaceous rocks along the North American Atlantic coastal plain. Cones of *Protodammara speciosa*, of similar age but slightly smaller than *Agathis borealis*, were obtained in New York.

The plum-yew family (Cephalotaxaceae) is represented by fossils of foliage and seeds that resemble those of *Cephalotaxus*, in Lower Cretaceous beds in Alaska, South Dakota, and Virginia.

A fairly recent family, the cypresses (Cupressaceae), is represented by undisputed fossil specimens of several genera in Upper Cretaceous and Tertiary sediments, including *Callitris*, *Thujopsis*, *Thuja*, *Juniperus*, and *Chamaecyparis*.

Fossils of the pine family (Pinaceae) may extend back into early Mesozoic times, but definite relationships of the genera are hard to establish. *Prepinus*, consisting of needlelike leaves on short shoots, were found in Lower Cretaceous formations in New York. Internal structure of these leaves suggests that *Prepinus* is an ancestral form of *Pinus*, possibly linking the pines, true cedars, and larches. Some undisputed *Pinus* cones and seeds are known from Cretaceous rocks of Greenland, Maryland, South Dakota, Virginia, and western Canada, and cones were found in Lower Cretaceous deposits in Europe. Fossils of various pine structures, including pollen, first appear in Eocene rocks and become abundant in Miocene (about 7,000,000 to 26,000,000 years ago) and Pliocene (from 2,500,000 to 7,000,000 years ago) formations from widely scattered localities.

Fossil remains belonging to *Abies*, *Cedrus*, *Keteleeria*, *Pseudotsuga*, and *Tsuga* have been found in Miocene and Pliocene beds. *Picea* occurred in the Lower Cretaceous of Belgium but did not appear in North American fossil floras until Miocene time, and *Larix* seems not to have appeared in the fossil record until the Pleistocene Epoch, about 2,500,000 years ago. Leafy shoots that might have been precursors of either *Cedrus* or *Larix* came from Jurassic rocks (of 136,000,000 to 190,000,000 years ago) and were called *Pitycladus*.

Three different types of wood with characteristics of the Pinaceae, but without attached cones or leaves, have been found in abundance and now are generally assigned to three form genera—*Cedaroxylon*, *Piceoxylon*, or *Pinuxylon*—which replace earlier names. Altogether, about 200 different "species" presumed by their authors to belong to the Pinaceae have been described.

The earliest known representative of the family Podocarpaceae was found in the Lower Jurassic (about 190,000,000 years ago) of New Zealand and in India. The

Podocarpaceae and Araucariaceae competed for dominance in the Southern Hemisphere at this time, and fossils belonging to *Dacrydium* and *Podocarpus* lines have been found in Middle Jurassic rocks of Antarctica. Other fossil representatives of these genera occurred throughout Cretaceous and Tertiary times in Argentina, Chile, South Africa, India, Australia, and Tasmania. *Phyllocladus asplenioides*, found in Eocene deposits in New South Wales, and other fossils resembling *Phyllocladus* have been found in Cretaceous rocks in Nebraska, New York, and Greenland.

The earliest known podocarps

Some Mesozoic fossils named *Elatides* were first placed in the Pinaceae but now are considered closer to the family Taxodiaceae. *Elatides williamsonis*, from Middle Jurassic strata in Yorkshire, is the most ancient of the unquestioned Taxodiaceae. Its staminate cones had scales with pronounced stalks and broadened, triangular terminal expansions set almost at right angles to the stalk. The ovulate cone scales bore two to five small ovules near the juncture of the stalk and the expanded terminal part. This fossil probably was closely related to *Cunninghamia*, or it might have been intermediate between *Cryptomeria* and *Sequoiadendron*.

The fossil predecessors of *Sciadopitys* go back to Lower Jurassic beds of New Jersey and Norway, to the Lower Jurassic of Sweden, and from Jurassic to Cretaceous in Greenland.

Fossils of *Sequoia*, *Metasequoia*, and *Taxodium* have been hard to differentiate when cones were not among the leaves and twigs. Much of the *Sequoia* and *Taxodium* material of Eocene age is now known to be *Metasequoia*. It is possible that some petrified tree trunks still standing in the Yellowstone National Park may be those of *Metasequoia*.

Genuine *Taxodium* fossils are abundant in Tertiary deposits of North America, Europe, and Asia; fossils uncertainly referred to as *Glyptostrobus* occur in the Lower Cretaceous of Greenland, and more positively determined ones come from Eocene rocks in North Dakota, the Miocene of Oregon, and from several other Tertiary localities; *Cryptomeria* has been found in Tertiary beds in Great Britain and from Upper Cretaceous strata in Japan; *Athrotaxis* has been recorded from the Lower Cretaceous of Patagonia and questionably from Czechoslovakia; *Cunninghamia* has come from the Upper Cretaceous of Japan, and pollen grains of this genus were found in the Green River shales of Eocene age in Colorado and Utah.

Classification

DISTINGUISHING TAXONOMIC FEATURES

The foremost identifying feature of the conifers is the cone, which has already been discussed in earlier sections. Another equally important character resides in the chromosomes. Not only the number but also some of the structural characteristics of conifer chromosomes are remarkably stable and may serve as taxonomic guides. Three basic series of chromosome number occur among the conifers: an 11 series in the Cupressaceae and Taxodiaceae, a 12 series in the Cephalotaxaceae and Pinaceae, and a 13 series in the Araucariaceae. The Podocarpaceae, obviously more evolutionarily versatile than the other families, has no fewer than four basic numbers in its various genera: 12, 13, 19, and 20 having been determined. (The last two numbers in this diverse series probably arose through the fusion of one or two chromosomes, followed by duplication of the entire set.) The chemical characteristics of various resins, turpentines, and other substances are important factors in the classification of conifers. It is already known that terpenoid characteristics of the heartwood of the Cupressaceae are quite different from those of the Pinaceae and that those of the Araucariaceae and Podocarpaceae have distinctive chemical divergences. Phenolic compounds in the haploxylon pines are markedly different from those of the diploxylon pines.

Arrangement of families, genera, and species according to morphological, anatomical, chemical, and ecological characteristics is not too difficult. But construction of a phylogenetic classification that shows lines of evolutionary derivation is a far more difficult task because of the incomplete nature of the fossil record.

ANNOTATED CLASSIFICATION

The taxonomic scheme given below is based upon characters, mostly visible, that are used in formulating what is called a key, or horizontal, rather than phylogenetic, classification.

ORDER CONIFERALES

Creeping shrubs to giant trees, all of which bear staminate, or pollen-producing, cones and most of which bear ovulate, or seed-producing, cones (exceptions: Cephalotaxaceae and Podocarpaceae). The largest order of gymnosperms, with about 50 genera and more than 500 species, found in all parts of the world.

Family Araucariaceae

Shrubs and trees with cone scales flattened, overlapping, and usually numerous and bearing bracts, at least in young cones. A single seed on each ovulate scale. Leaves are mostly leathery or their bases leathery and the free part curved.

The Araucariaceae consist of 2 genera only, *Agathis*, with 21 species, and *Araucaria*, containing 14 species distributed among 3 sections of that genus, based on characters of leaves, cones, seeds, and methods of germination. Section *Colymbea* has broad, flattened, rigid leaves, large cones about 20 cm long, seeds with only 2 cotyledons, and hypogeal germination (*i.e.*, the seed and cotyledons remain under the soil throughout germination and establishment of the seedlings). Section *Eutacta* has awl-shaped, less rigid, smaller, curved leaves, smaller cones (usually less than 10 cm long), seeds with 2 to 4 cotyledons, and epigeal germination (*i.e.*, the cotyledons are pushed above the surface of the soil upon germination). Section *Intermedia* has awl-shaped, only slightly rigid juvenile leaves but broad, flat, rigid adult leaves, cones intermediate in size, 2 to 4 cotyledons, and epigeal germination. The family is completely limited to the Southern Hemisphere in its natural distribution.

Family Cupressaceae (cypresses)

Prostrate to upright shrubs and trees with cone scales usually expanded at their tips and meeting at their edges (not overlapping), numbering usually 3 to 12 per cone; bracts not visible externally. Ovules, containing the seeds, are borne erect. Leaves are scalelike, in pairs or threes; they may be spirally arranged in juvenile forms of certain species.

The cypress family consists of 18 genera and 130 species: *Juniperus* leads with about 60 species, distributed widely in the Northern Hemisphere, *Cupressus* is second with 21 species likewise widely distributed, and the rest of the genera contain from 1 to 13 species each and range less widely.

Family Cephalotaxaceae (plum-yews)

Shrubs or small trees having single seeds borne at the tips of dwarf branches. No cone scales normally visible (one or two sterile ones may appear at base of the seed). Staminate cones are globular or ovoid, pollen sacs 3 to 9 per scale, and the pollen grains are winged.

The Cephalotaxaceae family has 1 genus and 7 species confined to eastern Asia from Assam to Korea and Japan. They have no economic value except as ornamentals

Family Pinaceae (pines, firs, spruces, etc.)

Shrubs to tall trees with cone scales flattened, overlapping, usually numerous, and bracts evident. Two seeds on each ovulate scale. Leaves are needlelike.

The Pinaceae is the most diverse of all the living coniferous families: it includes 10 genera and more than 200 species; *Pinus* alone has about 100 species (some botanists recognize only about 80) divided between 2 subgenera, *Haploxylon* and *Diploxylon*. *Haploxylon* pines, in which each needle has only 1 vascular bundle and usually 5 needles on a dwarf shoot, are called white pines; their wood is usually softer than that of the other subgenus. *Diploxylon* pines, whose needles have 2 vascular strands side by side within the endodermal sheath and mostly in bundles of 3 (3 on a dwarf shoot), have wood that contains more resin than does that of the white pines and is harder, the heartwood being slightly yellowish, hence the common name yellow pines. Many pines bear large, edible seeds that have high nutritive value and are used commercially. The smaller edible seeds of other pines are consumed by animals in large quantities.

Among the other genera of the Pinaceae, *Abies* has almost 50 species, widely distributed in the Northern Hemisphere. *Cedrus*, with 4 species in the Atlas Mountains of North Africa, Asia Minor, and the Himalayas, and *Cathaya* and *Keteleeria*, with 2 each, are both native in China. *Larix* has 10 species in the Northern Hemisphere. *Pseudolarix* contains 1 species, in eastern China. *Pseudotsuga*, with 5 species and an additional

variety, occurs in western North America, Japan, and China. *Tsuga* has about 10 species, in the Northern Hemisphere. *Picea*, with 31 species, is widely distributed in temperate regions of the Northern Hemisphere.

Family Podocarpaceae

Prostrate shrubs to tall trees having single olive-like seeds borne at the tips of dwarf branches. No cone scales normally visible. Staminate cones cylindrical, with two pollen sacs per scale, and the pollen grains not winged.

This predominantly Southern Hemisphere family consists of 7 genera collectively possessing about 150 species. *Podocarpus* surpasses *Pinus*, having more than 110 species distributed throughout most of the Southern Hemisphere and extending well into the Northern Hemisphere in Asia and as far as Mexico and the Caribbean in North America. *Dacrydium* is second in the family, with about 20 species, centred in the New Zealand–New Caledonia area but extending to Chile, the Philippines, and Malaya. Two genera, *Microcachrys* and *Saxegothaea*, each consists of a single species of limited range; the former native in Tasmania and the latter native in southern Chile and western Patagonia. The other 3 genera are: *Acmopyle*, with 3 species in New Caledonia and the Fijis; *Microstrobus*, with 2 species confined to New South Wales and Tasmania; and *Phyllocladus*, with 6 species in New Zealand, New Guinea, Tasmania, Borneo, and the Philippine Islands.

Family Taxodiaceae

Shrubs and tall trees with cone scales usually expanded at their tips and meeting at their edges (not overlapping), numbering usually 3 to 12 per cone; bracts not visible externally. The ovules are inverted (in contrast to the condition in the Cupressaceae). Leaves are linear to awl-shaped (scalelike on adult branches of *Sequoiadendron*) and spirally arranged on branchlets.

The Taxodiaceae include 10 genera, 6 of which have only one species. *Taxodium* and *Athrotaxis* consist of 3 species each, and *Cunninghamia* and *Taiwania* each have 2. It seems probable that the 6 monotypic genera are approaching extinction. They are *Cryptomeria* (Japan); *Glyptostrobus* (south China); *Metasequoia* (interior China); *Sciadopitys* (Japan); *Sequoia* (western coastal U.S.); and *Sequoiadendron* (western flanks of the Sierra Nevada in California).

CRITICAL APPRAISAL

Advances in knowledge of conifer relationships

Classification of the conifers has been debated vigorously for well over a century and is still being studied critically. Significant advances have been made in the understanding of the lines of descent and familial relationships among the conifers. One of the most important of these was that of a Swedish botanist, Rudolf Florin, who, in 1951, clarified the line of descent from *Cordaites*, a gymnosperm extinct since about the end of the Permian, to *Lebachia*, thence through *Ernestiodendron*, *Pseudovoltzia*, and *Ullmannia* to the earliest genera of the conifers, and who detected the homologies between Cordaitean inflorescences and the ovulate strobili of modern conifers. As a result, it is no longer doubted that the one-seeded structures of Cephalotaxaceae and Podocarpaceae are homologous with the apparently vastly different woody cones of *Araucaria* and *Pinus*.

A particularly vexing problem is the proper disposition of the yews (family Taxaceae). Earlier they were considered, along with the pines (Pinaceae), as the only two groups within the order of conifers. Later investigations, however, have led to the separation of the yews as a distinct order, Taxales, equivalent in rank to the Coniferales (see GYMNOSPERM).

The problems encountered in aligning, relating, and separating the genera and species included within the six families of conifers are extremely complex. Such details can be found only by consulting a fairly large number of research papers and books. Such a task would require using interdisciplinary approaches, data derived from microscopic as well as macroscopic features, and the consultation of chemical, physiological, and ecological sources. The brief bibliography below holds clues to more extensive literature dealing with such specialized and divergent subjects.

BIBLIOGRAPHY. L.H. BAILEY, *The Cultivated Conifers in North America, Comprising the Pine Family and the Taxads* (1933), descriptions of native and introduced conifers, with data on habitats, range, cultural needs (U.S. and Canada), pests, and uses; W. DALLIMORE and A.B. JACKSON, *A Handbook of Conifers and Ginkgoaceae*, rev. by S.C. HARRISON (1967), an up-to-date manual covering living conifers, with descriptive text, lists of cultivated forms, ranges of native habitats, and data on distribution, suitability for planting (England), and some indication of uses, with a list of important references consulted; R. FLORIN, "Evolution in Cordaites and Conifers," *Acta Horti Bergiani*, 15:285–388 (1951), a technical paper reviewing the work of earlier and contemporary botanists, with concise coverage of nearly 30 years of research by the author, including a comprehensive bibliography; HUI LIN-LI, "Present Distribution and Habitats of Conifers and Taxads," *Evolution*, 7:245–261 (1953), a review of distribution of these groups, with selected references; N.T. MIROV, *The Genus Pinus* (1967), an outstanding treatment of the pines of the world, giving geographic distribution, general information in ecology, chemistry, physiology, genetics, evolution, by-products, and uses, with references at the end of each chapter.

(I.L.W.)

Connecticut

An overview of the state

One of the 13 original English colonies on America's Eastern Seaboard, Connecticut is one of the six New England states located in the northeastern corner of the United States. Its 5,009 square miles (12,973 square kilometres) make it the third smallest state in the country, but the 3,032,217 residents counted in the 1970 census make it among the most densely populated. It lies athwart the great urban–industrial complex spreading down America's Atlantic Coast, with Massachusetts on the north, Rhode Island on the east, Long Island Sound (an arm of the Atlantic) on the south, and New York on the west.

Connecticut represents a special blending of modern urban life with a deep sense of its environment and its history—beaches and harbours, forest-clad hills, and quiet village greens surrounded by houses dating from the 17th and 18th centuries. At the same time, the Connecticut of the 1970s, incorporating these historic features into its life, is a highly industrialized state with the highest per capita income and value added by manufacture in the nation. The strength of its economy lies in a skilled working force, much of it fabricating products that have been manufactured in Connecticut since they were invented.

As might be expected, the population is heavily urban. The state has no single large city, however, and the intense crowding characteristic of many urban areas is not found in Connecticut. On a national scale, it continues its long tradition of being a prosperous state, with a steady immigration attracted by the good employment opportunities, excellent educational facilities, and pleasant living conditions for the great majority of its people. (For related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA.)

THE HISTORY OF CONNECTICUT

Colonization. Although earlier trading posts had been established on the Connecticut River by the Dutch from New Amsterdam and by the Plymouth Colony, permanent white settlement in Connecticut began as an overflow of people from the Massachusetts Bay Colony in 1633–35. The first settlers chose the fertile lands of the middle Connecticut Valley around Hartford. Later settlers located along the coast, at Saybrook in 1635 and at New Haven in 1638. In 1665 the Connecticut River settlements and the New Haven Colony were united, and the general outline of the state emerged, although its borders were not finally demarcated until 1881. The original Connecticut charter gave residents a claim to a strip of land as broad as the state itself—50 miles—and westward to the Pacific. This not unnaturally conflicted with Pennsylvanian claims, but after some troubles the dispute was settled in Pennsylvania's favour in 1782. West of Pennsylvania, Connecticut established its claim more fully to what came to be called the Western Reserve in northern Ohio, surrounding present-day Cleveland, and immigrants from Connecticut were among the first settlers there following the Revolution.

Land and border disputes

Political, economic, and social maturation. The political development of the colony began with the Funda-

mental Orders of Connecticut (1639), a civil covenant by the settlers establishing the system by which the river towns of Windsor, Hartford (the present capital city), and Wethersfield agreed to govern themselves. The orders created an annual assembly of legislators and provided for the election of a governor. This was superseded by the royal charter of 1662, a liberal document that provided for virtual self-government by the propertied men of orthodox faith in the colony. It served Connecticut well until replaced by the state constitution in 1818, a document that after many amendments was replaced by a constitution adopted in 1965, reflecting the more complex needs of contemporary government. The Congregational Church was disestablished in 1818.

Connecticut remained an agricultural region of farms with only a few small urban areas—Hartford, New Haven, New London, and Middletown—until the early 19th century. The economy began to change after 1800 when textile factories were established, and by 1850 employment in manufacturing outnumbered that in agriculture. The shift to manufacturing had been aided by the inventive genius of a number of Connecticut men. Eli Whitney, inventor of the cotton gin, developed the idea of machine-made parts for guns. An order for muskets from the federal government enabled him to build a musket factory in Hamden. The name and location have changed, but the factory manufactures guns today. The principle of interchangeable parts, adapted to clock manufacturing by Eli Terry of Plymouth in 1802, rapidly became basic to all manufacturing.

The economic, social, and political innovations that emerged in the 19th and 20th centuries were often resisted at first, but eventually were accepted. Slavery, first attacked by legislation in 1784, was not abolished completely until 1848. The constitution of 1818 granted manhood suffrage with certain property qualifications, but women's suffrage came only through federal enactment in 1920.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Surface features.* Essentially an oblong in shape, 100 miles west to east and 50 miles north to south, Connecticut covers the southern portion of the New England Upland. It contains three major regions: the Western Upland, the Central Lowland (Connecticut Valley), and the Eastern Upland. The Western Upland, often called the Berkshire Hills, contains the highest elevations in the state, about 2,300 feet (700 metres) in the northwest corner. It is drained by one major river, the Housatonic, and numerous tributaries.

Physiographic regions

The Central Lowland is different in character, being a downfaulted block of land, some 20 miles wide at the Massachusetts border and narrowing toward the sea, which it meets at New Haven. It is filled with sedimentary materials that have hardened into sandstone and shale. During periodic volcanic activity immense quantities of molten rocks were pushed up onto the surface and cooled, producing the rocky deposits of the central valley. These once level layers of sandstones and trap-rock have been faulted, broken, and tipped so that today there are numerous small ridges, some reaching 1,000 feet above their valleys. Within the lowland, the Connecticut and other rivers have eroded broad valleys into the softer sandstones.

The Eastern Upland resembles the Western in being a hilly region drained by numerous rivers. Their valleys come together to form the Thames River, which reaches Long Island Sound at New London. Elevations in this area rarely reach above 1,300 feet. In both uplands the hilltops tend to be level and have been cleared for agriculture; the first settlements were located there.

Climate. In Connecticut's moderate climate, winters usually average slightly below freezing (32° F, or 0° C) and the state receives from three to five feet of snow each year. Snow may remain on the ground until March, but more commonly mild spells and rains during the winter melt it so that the ground is bare. Summers average between 70° and 75° F (21° to 24° C), with occasional heat waves driving daytime temperatures above 90° F (32° C). Precipitation, averaging three to four inches (75–100 millimetres) per month, is quite evenly distributed. The coastal portions have somewhat warmer winters and cooler summers than the interior, while the northwestern uplands are high enough to have cooler and longer winters with heavier snowfall. Perhaps the most marked characteristic of Connecticut's weather is its changeability. Cold waves and heat waves, storms and fine weather, alternate weekly or even daily. The statement of Hartford resident Mark Twain "If you don't like Connecticut weather, wait a minute" has become a widely appropriated and adapted proposition.

Vegetation and animal life. Originally Connecticut was a forested region. The few Indian clearings, the swampy flood plains, and the tidal marshes accounted for only about 5 percent of the total. It is part of the mixed deciduous and coniferous forest of the eastern United States. The southern two-thirds is largely an oak forest. The northern border belongs to the northern hardwood region of birch, beech, maple, and hemlock. A few higher elevations and some sandy sections support a coniferous forest. Virtually all of the primeval forest has been cut, and the current woodland that covers two-thirds of the state is a mixed forest.

The animal life when the first settlers arrived included deer, bear, wolves, foxes, and numerous smaller species, such as raccoon, muskrat, porcupines, weasels, and beaver. Deer are still found in the less densely settled regions, but in general the larger animals have been severely decimated. Most birds are migratory, but chickadees, blue jays, and the immigrant English sparrows are year-round residents.

Patterns of human use. Most regions in Connecticut are not clearly defined, although Fairfield County in the southwest section is uniquely oriented toward New York City, serving as a major "bedroom suburb" for many of the most affluent commuters to that centre. It includes, nonetheless, two of the largest cities, Stamford and Bridgeport. The northwestern and northeastern quarters are less densely populated areas. They have some agriculture, but most residents here, as elsewhere in the state, work at urban occupations in the manufacturing cities and towns along the rivers.

Rural and urban settlement

Typical of New England, Connecticut's small towns represent a territorial concept that is equivalent to a township in other parts of the country. Within each town, a town centre is surrounded by the town hall, schools, churches, usually a village green, a number of houses, and often a tiny business district with several stores. Elsewhere within the town, other hamlets may contain similar communal gatherings. If the hamlet is on a stream, the houses often cluster around a red-brick factory erected in the 19th century to run its machinery from a waterwheel in the river. Such mill villages are located throughout the state, although many of the factories have been abandoned. Farmsteads and cultivated fields once lay between such small population nodes, but today the roads connecting these villages are sparsely lined with rural, non-farm homes.

City status in Connecticut is determined not by population but by vote of the residents to change their governmental system from a town meeting to a city form. Most of the larger communities have done this, but in the 1970 census there still remained four towns with over 50,000 people and two cities with fewer than 10,000. All of the larger towns and cities are manufacturing centres today, some of which originated as mill towns and grew with their factories. The power source may have changed first to steam and later to electricity, and often the products manufactured may have changed to fill the needs of a new economic and social structure, but each city and town prides itself on the uniqueness that often is associated with its products.

THE PEOPLE OF CONNECTICUT

The ethnic mix. The Algonkin Indians, the original occupants of Connecticut, comprised about 16 separate tribes with some 5,000 to 7,000 members. The first white settlers were English, coming directly from England or

by way of Massachusetts Bay Colony. During the first two centuries, population growth occurred primarily through an excess of births over deaths; immigrants, mainly from the British Isles, came in at a rather slow rate. At the time of the first census, in 1790, Connecticut had a homogeneous population, about 90 percent of which was of English ancestry. Blacks were a minor element in the population from the beginning, accounting for only 2 percent in 1790.

Patterns of immigration

The immigration of the Irish, beginning in the 1840s, and of French Canadians after the Civil War, continued throughout the century. Later in the century the primary sources of foreign immigration shifted to southern and eastern Europe—Italy, Poland, the Austro-Hungarian Empire, and Russia. Each immigrant group tended to congregate in certain parts of the state. Thus New Haven and its suburbs today are populated largely by descendants of Italian immigrants; Poles concentrated in the Naugatuck Valley; the French Canadians are in the northeast. The immigration of blacks into Connecticut after World War II showed the same tendency. In 1970 the black population of the state had risen to 6 percent, 74 percent of whom were clustered in the five largest cities, and New Haven and Hartford were more than 44 percent black. Puerto Ricans have moved out from New York, especially into Stamford and Bridgeport.

Demography. Connecticut's population grew more rapidly during the 1960s than did the nation as a whole, a rate that is a phenomenon of the 20th century. From 1790 to 1840 the state's growth rate hovered between 4 and 8 percent per decade: Connecticut was—considering its small size and its limited agricultural resources—quite adequately filled. Thousands of Connecticut residents, especially the young, migrated to the better agricultural lands in the American West. Their places were taken by newcomers from Europe, and the state's population growth passed the national rate in 1900 and has remained equal to or above it since.

Over 300 years the distribution of Connecticut's people has reflected the region's changing economy and resources of the land. Settlement began in the middle Connecticut Valley, where the soils were good, and on the coast, where maritime activities, trading, and fishing supplemented the living that the settlers derived from the land. The upland areas were not fully occupied until the end of the 18th century, yet by 1790 the population was fairly evenly distributed. Towns with somewhat better agricultural lands or with other resources—marine or mineral—had denser populations. During the 19th century the rise of water-powered manufacturing attracted young people from the agricultural upland towns to the growing mill towns, and virtually all of the upland towns lost population. Towns with better assets for manufacturing grew rapidly.

The movement of people and industry into the cities dominated the population movements until 1920, since which time Bridgeport, Hartford, and New Haven, the three largest cities, have experienced the general movement from the cities to the suburbs and beyond them to the former agricultural hill towns. The place of this older population has been taken by the newer immigrants, the blacks and the Puerto Ricans.

THE STATE'S ECONOMY

Sources of income. The foundation of Connecticut's economy is manufacturing, which employs some 40 percent of the state's work force. In addition to such military products as helicopters, submarines, aircraft engines, guns, and ammunition, it makes thousands of items that are sold on a worldwide basis. Among the items that have been manufactured in Connecticut by long tradition are pins, clocks, silverware, sewing machines, Winchester rifles, and many brass products. The few iron mines closed early in the 20th century, and the state's top ranking in value added by manufacture is due mainly to the necessary import of nearly all raw materials. Only sand, gravel, and stone remain native.

An economic overview

Since 1870, agriculture has declined in importance until today it is a minor element in the economy. Connecticut's farms produce substantial quantities of milk, eggs, poultry, and vegetables for local consumption and one export crop, shade-grown tobacco, mainly for cigars. Fishing has never been very important except for the oyster industry, which in recent years has been devastated by natural enemies and the pollution of the coastal waters containing the oyster beds.

Insurance as a business enterprise did not originate in Connecticut, but today the state often is referred to as the American centre of this activity and Hartford as The Insurance City. Marine insurance was the first concern of Connecticut companies, and eventually the coverages offered expanded to many forms of casualty insurance. Today, three of the largest insurance companies of the United States are based in Connecticut.

Economic management. To correct abuses in the free-enterprise system, Connecticut has had to enact numerous regulations. The first child-labour law was passed in 1842, but it was ineffectual; for 30 years after its passage hundreds of children continued to work long hours in the textile mills. A labour department was set up by the state government in 1873, and since then hundreds of laws and regulations have been enacted to control working conditions. The length of the working day, minimum wage rates, equal pay for equal work, and similar protective regulations have been passed. State departments supervise banks, insurance companies, and the public utilities, and in 1959 the Department of Consumer Protection was organized to consolidate several existing agencies. Labour unions are strong and may be given partial credit for the high wages and good working conditions characteristic of most factories. There is also an active association of manufacturers.

Labour legislation

Transportation. Connecticut's railroad network is a basic link in the Boston–New York transportation pattern, which is chronically bankrupt and, in operation, a bane to New York City commuters. The first railroads were constructed to bring to Connecticut ports the produce of the agricultural interior. Each of the larger river valleys—the Housatonic, Naugatuck, Connecticut, Willimantic, and Quinebaug—supported its own railroad. The present line along the shore was completed in 1852. Until 1930 the railroads flourished with the expanding Connecticut economy, but highway competition for passengers and products reduced railroad traffic severely. Today, most of the river lines have dropped passenger service; freight service continues on some, but on others it has been abandoned. Limited-access highways crisscross the state, but they are concentrated in the densely settled coastal and Connecticut Valley regions. Connecticut pioneered this new type of road. The Merritt Parkway, from New York state to New Haven, its first section, opened in 1938, often is acclaimed as one of the most scenic and best designed of these roads.

Bradley International Airport north of Hartford serves western Massachusetts as well, while smaller ones at Bridgeport, New Haven, and New London are used by regional commercial airlines. The port of New Haven is the second largest in New England, after Boston, and New London is a major base of the United States Coast Guard.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *State level.* Connecticut's state government, like that of most American states, is headed by a strong governor elected for a four-year term. He initiates legislation, prepares the state budget, appoints department heads, and can veto individual items of an appropriation bill. Members of the General Assembly are elected for two-year terms. The 36 senatorial districts are approximately equal in population. The House of Representatives was originally based on towns, with each town, regardless of size, having at least one representative. The 1965 constitution reapportioned the lower branch so that it, too, is based upon population.

The judicial system

Connecticut has the usual American judicial system of a supreme court; a superior court that hears cases of serious crimes and those civil cases involving large sums of money; a court of common pleas for lesser civil cases

and certain criminal ones; and a circuit court for minor criminal offenses. Connecticut was the first state to establish juvenile courts on a state-wide basis, and proceedings therein are entirely civil in nature. No child under the age of 16 can be convicted of a crime. Most judges are appointed by the General Assembly and may be re-appointed. Probate judges are elected.

Local government and politics. Below the state government are 169 territorial units. Legally they are creatures of the state; their rights and responsibilities are set out in state statutes. These local governments maintain roads and provide elementary and secondary education and police and fire protection. Larger municipalities may also provide water and sewage facilities and other services. The original form of government was based on the town meeting, at which the citizens elected selectmen to run the town between the annual meetings. As populations increased and problems of administration became more complex, other systems were substituted. The largest communities have opted for a city form with an elected mayor and council. Some smaller communities have elected mayors, some have town or city managers. A few of the largest towns have substituted a representative town meeting, but most retain the original form.

The social milieu. Connecticut generally is looked upon as a good state in which to live.

Education. From the earliest days, every town has been required to maintain public elementary schools and, as the town grew in size, secondary schools as well. The state supports public postsecondary education in five academic institutions and four technical colleges. In recent years the state has added eight community colleges. Connecticut has also been noted for the number and quality of its private schools and colleges. The most noted is Yale University, in New Haven, regarded as one of the great educational institutions in the world, but smaller institutions such as Wesleyan University, in Middletown, also enjoy nationwide recognition.

Health and welfare. Today the community and the state have become increasingly involved in health and medical care. Most people live within ten miles of hospital services, and doctors and other medical personnel are numerous. The most recent additions to the state's medical services are clinics for the treatment of alcoholism and drug addiction, some of the first established in cooperation with Yale University School of Medicine.

Urban redevelopment. Connecticut's progressiveness may be seen in the urban reconstruction going on in all of its larger cities. Redevelopment projects in Hartford and New Haven were begun in the 1950s to restore the blighted hearts of cities. Slums and run-down business sections were levelled and replaced by new housing, both for high-income and low-income people, and by new commercial structures.

State agencies

Governmental involvement. Connecticut's numerous special departments and agencies include the customary ones handling personnel, finance, public works, highways, agriculture, education, health, and development. Reflecting the rise of new problems, the legislature has in recent years added departments or commissions for community affairs, youth services, transportation, human rights, mental health, and aging. Local governmental units also face many of the same problems and have organized groups that work on them. Connecticut's conservatism is not evident here, and it has been among the first states to address itself to changing conditions.

CULTURAL LIFE AND INSTITUTIONS

Preservations. For the resident and nonresident alike, Connecticut provides a variety of landscapes: rocky headlands, beaches, forested hills, and, perhaps most attractive, the small towns around their tree-dotted village greens. On them and elsewhere in the towns the hundreds of houses dating from the 17th and 18th centuries are preserved by more than 100 local or national historical societies. Maintained by the state or private organizations are numerous sites important in Connecticut's past or associated with illustrious individuals. These include the Putnam Wolf Den in Pomfret, Mt. Riga Furnace in Salisbury, Ft. Griswold in Groton, Newgate Prison in East Granby, Mark Twain's home in Hartford, the Tapping Reeve House and First Law School in Litchfield, and the (William) Gillette Castle in East Haddam. Perhaps the best known is Mystic Museum in Mystic, where a small New England seaport has been recreated with all its ships and shops. The outdoorsman can tramp the many miles of trails and camp in one of the 29 state forests, covering more than 130,000 acres, or the 82 state parks, comprising almost 28,000 acres.

The arts. Recreation in another form is provided in the fine arts. Art exhibitions are held annually in many cities, which are well supplied with art galleries and museums. The best known are the Yale University Art Gallery, the Wadsworth Atheneum in Hartford, and the New Britain Museum of American Art. Symphony concerts and concerts by smaller groups are presented regularly in the larger communities. The New Haven Pops Concert, held in the Yale Bowl, is a weekly summer feature. Several educational institutions specialize in music and present concerts open to the public throughout the year. Repertory companies operate in or near the resort areas in the summer: the Ivoryton Playhouse in Essex, Westport Country Playhouse in Westport, and the Oakdale Musical Theater in Wallingford. Nationally known are the Shakespeare theatre in Stratford and the Goodspeed Opera House in East Haddam. The Yale School of Drama was, at its founding in 1925, the first such school at an institution of higher learning. The annual American Dance Festival at Connecticut College in New London, presenting companies in all mediums of dance, is almost unique in the United States. Southwestern Connecticut is also within easy reach of the vast artistic repertory of New York City.

The performing arts

Communications. The press of Connecticut includes 29 daily newspapers, two of which are university papers. The five largest cities have both morning and evening papers, while 41 other towns support weekly papers. Thirty-eight AM and 26 FM radio stations broadcast programs, and there are nine television stations. Most of the state's residents can also pick up television broadcasts originating in New York, Massachusetts, or Rhode Island, and Boston and New York newspapers are widely distributed.

Prospects. Living conditions can be expected to improve in the cities, which are vigorously attacking the problems of urban decay. The quality of the landscape is under intense study in the state, and laws controlling air and water pollution have been passed by the legislature.

BIBLIOGRAPHY. *Connecticut: A Guide to Its Roads, Lore and People* (1938), one of the "American Guide Series" describing all parts of the state; CHARLES M. ANDREWS, *The Colonial Period of American History: The Settlements*, vol. 2 (1934), the most authoritative account of early Connecticut history; JOHN W. BARBER, *Connecticut Historical Collections* (1836), a town-by-town description by an investigator who travelled and knew the state personally in the 1830s; CONNECTICUT DEVELOPMENT COMMISSION, *Connecticut Market Data* (annual), statistics on population and economy; *The Physical Geography of Connecticut* (1963), one of a valuable series including other economic, political, and social areas; CONNECTICUT, SECRETARY OF THE STATE, *Register and Manual* (annual), compendium of facts about Connecticut state and local governments; FLORENCE S.M. CROFUT, *Guide to the History and the Historic Sites of Connecticut*, 2 vol. (1937), the definitive work on this topic; JOHN W. DEFOREST, *History of the Indians of Connecticut from the Earliest Known Period to 1850* (1853), the basic reference on Connecticut's Indians; JOSEPH B. HOYT, *The Connecticut Story* (1962), a children's history, well illustrated and with geographical insights; LEAGUE OF WOMEN VOTERS, *Connecticut in Focus* (1970), essential description of Connecticut government today; ODELL SHEPARD, *Connecticut Past and Present* (1939), well-written essays about the appearance as well as the history of Connecticut; ALBERT E. VAN DUSEN, *Connecticut* (1961), a history of Connecticut that is weighted on political developments but is rather weak on geographical and economic aspects of the state; LAWRENCE F. WILLARD and A.V. SIZER, *Pictorial Connecticut* (1962), a collection of excellent photographs by an able photographer, with descriptive text.

(J.B.Ho.)

Connective Tissue, Human

The connective tissues are a heterogeneous group of tissues derived from the mesenchyme, a meshwork of stellate cells that develop in the middle layer of the early embryo. They have the general function of maintaining the structural integrity of organs and providing cohesion and internal support for the body as a whole. The connective tissues include several types of fibrous tissue that vary only in their density and cellularity, as well as more specialized variants ranging from adipose tissue through cartilage to bone. The cells that are responsible for the specific functions of an organ are referred to as its parenchyma, while the delicate fibrous meshwork that binds the cells together into functional units, the fibrous partitions or septa that enclose aggregations of functional units, and the dense fibrous capsule that encloses the whole organ, collectively make up its connective-tissue framework, or stroma. Blood vessels, both large and small, course through connective tissue, which is therefore closely associated with the nourishment of tissues and organs throughout the body. All nutrient materials and waste products exchanged between the organs and the blood must traverse perivascular spaces occupied by connective tissue. One of the important functions of the connective-tissue cells is to maintain conditions in the extracellular spaces that favour this exchange.

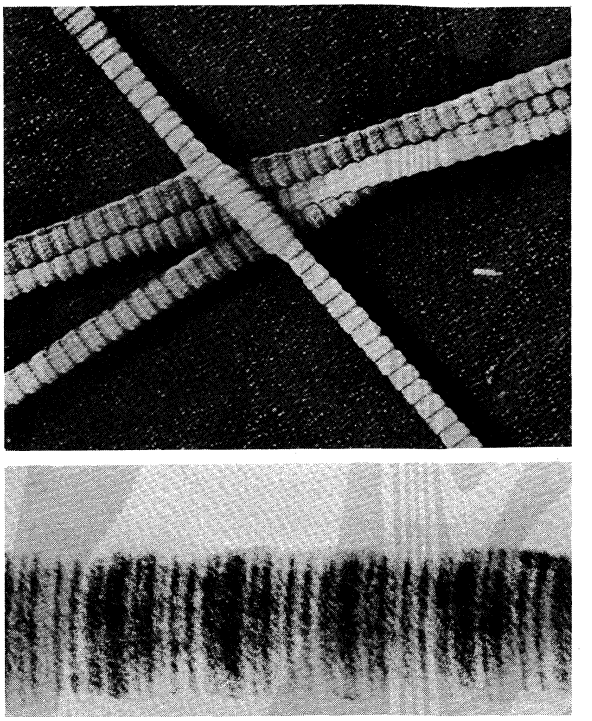
Some organs are suspended from the wall of a body cavity by thin sheets of connective tissue called mesenteries; others are embedded in adipose tissue, a form of connective tissue in which the cells are specialized for the synthesis and storage of energy-rich reserves of fat, or lipid. The entire body is supported from within by a skeleton composed of bone, a type of connective tissue endowed with great resistance to stress owing to its highly ordered, laminated structure and to its hardness, which results from deposition of mineral salts in its fibres and amorphous matrix. The individual bones of the skeleton are held firmly together by ligaments, and muscles are attached to bone by tendons, both of which are examples of dense connective tissue in which many fibre bundles are associated in parallel array to provide great tensile strength. At joints, the articular surfaces of the bones are covered with cartilage, a connective tissue with an abundant intercellular substance that gives it a firm consistency well adapted to permit smooth gliding movements between the apposed surfaces. The synovial membrane, which lines the margins of the joint cavity and lubricates and nourishes the joint surfaces, is also a form of connective tissue.

COMPONENTS OF CONNECTIVE TISSUE

All forms of connective tissue are composed of (1) cells, (2) extracellular fibres, and (3) an amorphous matrix, called ground substance. The proportions of these components vary from one part of the body to another depending on the local structural requirements. In some areas, the connective tissue is loosely organized and highly cellular, in others its fibrous components predominate, and in still others, the ground substance may be its most conspicuous feature. The anatomical classification of the various types of connective tissue is based largely upon the relative abundance and arrangement of these components.

Collagenous and reticular fibres

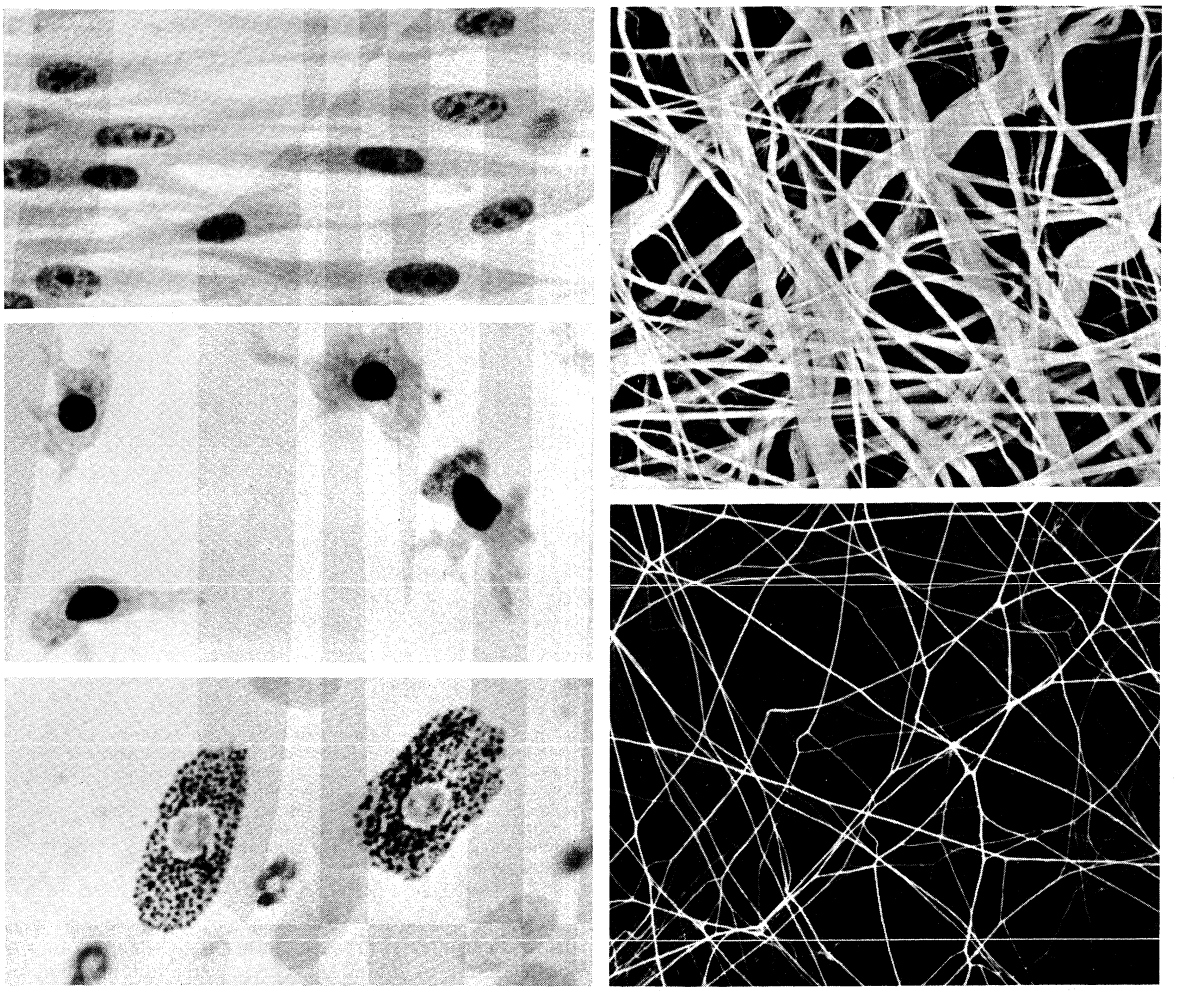
The fibrous components are of three kinds, collagenous, elastic, and reticular fibres. Most abundant are the fibres composed of the protein collagen. The fibrous components of loose areolar connective tissue when viewed with the light microscope appear as colourless strands of varying diameter running in all directions, and, if not under tension, these have a slightly undulant course (see photo). At high magnification, the larger strands are seen to be made up of bundles of smaller fibres. And the smallest fibres visible with the light microscope can be shown with the electron microscope to be composed of multiple fibrils up to 1000 Å in diameter. These unit fibrils are cross-striated with transverse bands repeating every 640 Å along their length.



(Top) Electron micrograph of four collagen fibrils shadowed with metal to increase their contrast and reveal their periodic 640-Å cross striation. (Bottom) In a more highly magnified electron micrograph of a collagen fibril stained with phosphotungstate, the cross striation can be resolved into several distinct bands which depend upon the arrangement of the tropocollagen molecules within the fibre.
By courtesy of (top) J. Gross, (bottom) B.R. Olsen

Extracellular fibres. Collagen is of commercial as well as medical interest because leather is the dense collagen of the dermis of animal skins preserved and toughened by the process called tanning. Fresh collagen dissolves in hot water, and the product is gelatin. Under appropriate conditions, collagen can be brought into solution without chemical change. The fundamental units in such solutions are slender tropocollagen molecules about 14 Å wide and 2800 Å long. Collagen appears to be secreted in this form by the connective-tissue cells called fibroblasts, and the tropocollagen molecules assemble extracellularly to form striated collagen fibrils. By an alteration of the physicochemical conditions, tropocollagen in solution can be induced to polymerize with the formation of cross-striated fibrils identical to native collagen, thus simulating in the test tube the process of assembly that is believed to take place during fibrogenesis in the living organism. Analysis of the structure of collagen by X-ray diffraction has shown that the tropocollagen molecule consists of three side-by-side polypeptide chains—linear combinations of a number of amino acids, which are subunits of proteins—each in the form of a left-handed helix. These three left-handed helices are further twisted around one another to form a major right-handed helix. Upon chemical analysis, the amino-acid composition of collagen is found to be unique in its extremely high proline content (22 percent) and in the fact that one-third of the amino acid residues are glycine. It is the only naturally occurring protein known to contain hydroxyproline and hydroxylysine. Two of the three polypeptide chains comprising the tropocollagen molecule are similar in amino-acid composition, while the third is distinctly different. In the tissues, the collagen fibrils are believed to be held together by a polysaccharide component that has not been fully characterized.

Reticular fibres are distinguished by their tendency to form fine-meshed networks around cells and cell groups, and by virtue of their property of staining black because of adsorption of metallic silver when they are treated with alkaline solutions of reducible silver salts.



Photomicrographs of connective tissue components.
(Top left) Long fusiform fibroblasts growing in tissue culture (magnified about 520 ×). (Top right) Randomly oriented collagenous fibres of varying size in a thin spread of loose areolar connective tissue (magnified about 370 ×). (Centre left) Four macrophages in cell culture (magnified about 825 ×). Their irregular outline is associated with amoeboid migration. (Bottom left) Metachromatic granules fill the cytoplasm of these mast cells, which have centrally located nuclei (magnified about 750 ×). (Bottom right) Selectively stained network of thin elastic fibres in a thin sheet of areolar connective tissue (magnified about 390 ×).

By courtesy of (top left) W. Bloom and D. Fawcett, *Textbook of Histology*, 9th ed.; W.B. Saunders Company, (top right, centre left, bottom left, bottom right) Don Fawcett

They were formerly believed to be composed of a distinct protein, reticulin, but electron microscopy has revealed that reticular fibres are small fascicles of typical collagen fibrils interwoven to form a network. It is now apparent that reticular fibres are simply a form of collagen and that their distinctive staining depends upon the mode of association of the fibrils and possibly upon subtle differences in their relation to the polysaccharide material that binds them together.

Elastic fibres

Elastic fibres are composed of the protein elastin and differ from collagenous fibres in dimensions, pattern, and chemical composition. They do not have uniform subunits comparable to the unit fibrils of collagen. They present a variable appearance in electron micrographs; sometimes they appear to have an amorphous core surrounded by minute fibrils, while in other sites they appear to consist exclusively of dense amorphous material. Whether there are in fact two components or whether these are differing forms of the same substance is not yet clear. At the light-microscope level, the fibres vary in diameter and often branch and reunite to form extensive networks in loose connective tissue. When present in high concentration, they impart a yellow colour to the tissue. In elastic ligaments, the fibres are very coarse and are arranged in parallel bundles. In the walls of arteries, elastin is present in the form of sheets or membranes perforated by openings of varying size. Elastic fibres are extremely resistant to hot water, to strong alkali, and even to digestion with the proteolytic enzyme trypsin. They can be digested, however, by a specific enzyme, elastase, present in the pancreas. Upon chemical analysis, elastin, like collagen, is found to be rich in glycine and proline, but it differs in its high content of valine and in the presence of an unusual amino acid, desmosine. As their name implies, elastic fibres are highly distensible and, when broken, recoil like rubber bands. Changes in this property and diminution in their numbers are thought to be, in part, responsible for the loss of elasticity of the skin and of the blood-vessel walls in old age.

Ground substance. The amorphous ground substance of connective tissue is a transparent material with the properties of a viscous solution or a highly hydrated thin gel. Its principal constituents are large carbohydrate molecules or complexes of protein and carbohydrate, often called mucopolysaccharides. One of these carbohydrates is hyaluronic acid, composed of glucuronic acid and an amino sugar, *N*-acetyl glucosamine. Other carbohydrates of the connective tissue are chondroitin-4-sulfate (chondroitin sulfate A) and chondroitin-6-sulfate (chondroitin sulfate C). The sugars of the sulfates are galactosamine and glucuronate. Multiple chains of chondroitin sulfate seem to be bound to protein. These substances in solution are viscous. All substances passing to and from cells must pass through the ground substance.

Variations in its composition and its viscosity may therefore have an important influence on the exchange of materials between tissue cells and the blood. Its physical consistency also constitutes a barrier to the spread of particulates introduced into the tissues. It is interesting, in this relation, that some bacteria produce an enzyme, hyaluronidase, which breaks up hyaluronic acid into subunits and alters the viscosity of ground substance. The ability of these bacteria to produce this enzyme is probably responsible for their invasiveness in the tissues.

Cells of connective tissue. The cells of connective tissue include three types that are relatively stationary, fibroblasts, macrophages, and adipose cells, and several types of motile, migrating cells—mast cells, monocytes, lymphocytes, plasma cells, and eosinophils. The ubiquitous fibroblasts are the principal cells of connective tissue, occurring as long spindle-shaped cells stretched along bundles of collagen fibrils. Their function is to secrete tropocollagen and constituents of the ground substance and to maintain these extracellular tissue components. When organs are injured, the fibroblasts of the stroma are stimulated to proliferate; they migrate into the defect and deposit an abundance of new collagen, which forms a fibrous scar. The macrophages or histiocytes are also important for tissue repair and for defense against bacterial invasion. Like fibroblasts, they are normally fusiform or stellate and are deployed along the collagen fibre bundles, but, if there is tissue damage or bacterial invasion, they withdraw the projections by which they are anchored and migrate by active amoeboid movements into the affected area. They have a great capacity for phagocytosis—the process by which cells engulf cellular debris, bacteria, or other foreign matter, and break them down by intracellular digestion. Thus they represent an important force of mobile scavenger cells.

Adipose cells are connective-tissue cells that are specialized for synthesis and storage of reserve nutrients. They receive glucose and fatty acids from the blood and convert them to lipid, which accumulates in the body of the cell as a large oil droplet. This distends the cell and imposes upon it a spherical form. The nucleus is displaced to the periphery, and other metabolically active constituents of the cell are confined to a thin rim of cytoplasm around the large, central droplet of lipid. Adipose cells may occur in small numbers anywhere in connective tissue, but they tend to develop preferentially along the course of small blood vessels. Where they accumulate in such large numbers that they become the predominant cellular element, they constitute the fat or adipose tissue of the body.

All the cells of connective tissue develop during embryonic life from the mesenchyme, a network of primitive stellate cells that have the potentiality for differentiating along several different lines depending upon local conditions. In addition to the specialized cell types of adult connective tissue described above, it is believed that small numbers of mesenchymal cells persist into postnatal life in the walls of small blood vessels and elsewhere and that these retain the capacity to differentiate into fibroblasts, adipose cells, or histiocytes as the need arises.

Mast cells

In addition to the relatively fixed cell types just described, there are free cells that reside in the interstices of loose connective tissue. These vary in their abundance and are free to migrate through the extracellular spaces. Among these wandering cells are the mast cells; these have a cell body filled with coarse granules that exhibit a characteristic metachromatic staining reaction. The function of these cells is still poorly understood, but they are known to produce and store in their granules two biologically active substances, histamine and heparin. Histamine affects vascular permeability, and heparin, when added to blood, delays or prevents its clotting. Mast cells respond to mechanical or chemical irritation by discharging varying numbers of their granules. Histamine released from them causes fluid to escape from neighbouring capillaries or venules; this results in local swelling, as seen in the welt that appears around an insect bite.

Eosinophils

Tissue eosinophils are a type of white blood cell or leukocyte. Some of these migrate through the walls of capillaries and take up residence in the connective tissues. They have polymorphous nuclei and, in the cell substance outside the nuclei, coarse granules that stain with eosin and other acid dyes. In electron micrographs, the granules contain conspicuous crystals. The granules have been isolated and shown to contain a variety of hydrolytic enzymes. Eosinophils are normally widespread in connective tissues of the body, but they are especially abundant in persons suffering from allergic diseases. The cells are believed to phagocytose and break down antigen-antibody complexes.

Plasma cells, which are present in limited numbers in loose connective tissues and in larger numbers in lymphoid tissue, are essential to the body's immunological defenses. They are specialized for synthesis and release of those specific immune globulins, called antibodies, that combine with and neutralize foreign proteins introduced into the body. Lymphocytes are among the normal cellular elements of the blood, but they may also leave the blood and migrate in the connective tissues. They are small round cells with a thin rim of cytoplasm and only limited synthetic activity. They appear to be able to recognize foreign protein and to respond to its presence by enlargement, proliferation, and transformation into plasma cells. They thus constitute an important reserve of relatively undifferentiated cells capable of sustaining an immunological response. They are the subject of intensive study because of their important role in defense against disease and because they participate in the rejection reaction that often frustrates the surgeon's efforts to transplant organs from one individual to another.

Monocyte and neutrophilic leukocyte

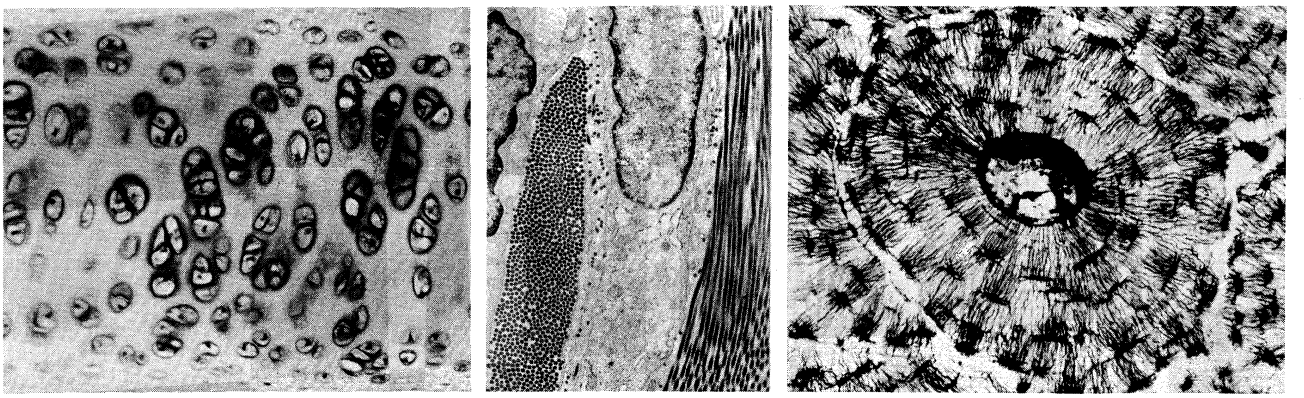
Another of the leukocytes that enter the connective tissues from the blood is the monocyte, a mononuclear cell larger than the lymphocyte and with different potentialities. These migratory cells can divide and, when appropriately stimulated, can transform into highly phagocytic macrophages. Thus at sites of bacterial invasion, when the resident population of fixed macrophages or histiocytes cannot cope with the situation, the monocytes of the connective tissues are reserves that can be mobilized at the focus of infection and can be continually reinforced by emigration of additional monocytes from the blood. This reaction of the blood and connective-tissue cells to injury is called inflammation and is usually accompanied by local heat, swelling, redness, and pain. Under these conditions, the neutrophilic leukocytes (white blood cells called neutrophilic because of their neutral staining characteristics with certain dyes), which are not normally present in connective tissue in significant numbers, may also migrate through the capillary walls in astronomical numbers and join the macrophages in the work of ingesting and destroying bacteria. Voraciously phagocytic, the neutrophils have a short life-span; having accomplished their mission, they die in great numbers. Pus, which may accumulate at sites of acute inflammation, is composed largely of dead and dying neutrophilic leukocytes.

TYPES OF CONNECTIVE TISSUE

The description presented thus far applies to the widely distributed loose areolar connective tissue, which is relatively unspecialized and can therefore be considered prototypic. In the more specialized forms of connective tissues, one component or another may predominate over all the others, depending upon the local structural or metabolic requirements.

Adipose tissue, for example, is a variant of loose areolar tissue in which large numbers of adipose cells make up the bulk of the tissue.

Dense fibrous connective tissue is composed of closely packed bundles of collagen and their associated fibroblasts, but there are relatively few elastic fibres and little ground substance. The term irregular dense fibrous tissue is applied to sites where the collagen bundles are randomly oriented and interwoven, as in the dermis of the skin and the capsules of joints. Regular dense fibrous tissue is the term used to describe tendons, ligaments, and



Three types of connective tissue.
(Left) Photomicrograph of hyaline cartilage from the trachea. The chondrocytes are enclosed in lacunae distributed in groups in an abundant matrix (magnified 250 ×). (Centre) Electron micrograph of a small area of connective tissue illustrating the intimate association of cells and fibres. In the centre is a portion of a fibroblast, and on either side are two collagen fibres. The collagen fibre on the left is cut transversely, showing round cross sections of the unit fibrils. The collagen fibre on the right has been cut nearly parallel to its long axis and shows extensive segments of the cross-striated fibrils (magnified about 6,625 ×). (Right) Photomicrograph of a ground section of bone showing a haversian system, or osteon, in cross section, and adjacent interstitial lamellae. The osteocytes occupy flat lacunae, which appear black in this preparation. Slender canaliculi connect neighbouring canaliculi (magnified about 125 ×).
By courtesy of (left) W. Bloom and D. Fawcett, *Textbook of Histology*, 9th ed.; W.B. Saunders Company, (centre, right) Don Fawcett

aponeuroses (fibrous sheets that form attachments for muscles), where the collagen fibres are precisely oriented in parallel bundles.

Synovial membrane lining joint capsules is composed of loose vascular connective tissue but has cells specialized for secretion of the viscous synovial fluid, which is rich in hyaluronic acid. This fluid serves as a lubricant and nutrient for the avascular joint surfaces. Similar tissue forms sheaths around tendons where they pass over bony prominences.

Cartilage

Cartilage is a form of connective tissue in which the ground substance is abundant and of a firmly gelated consistency that endows this tissue with unusual rigidity and resistance to compression. The cells of cartilage, called chondrocytes, are isolated in small lacunae within the matrix. Although cartilage is avascular, gaseous metabolites and nutrients can diffuse through the aqueous phase of the gel-like matrix to reach the cells. Cartilage is enclosed by the perichondrium, a dense fibrous layer lined by cells that have the capacity to secrete hyaline matrix. Cartilage grows by formation of additional matrix and incorporation of new cells from the inner chondrogenic layer of the perichondrium. In addition, the young chondrocytes retain the capacity to divide even after they become isolated in lacunae within the matrix. The daughter cells of these divisions secrete new matrix between them and move apart in separate lacunae. The capacity of cartilage for both appositional and interstitial growth makes it a favourable material for the skeleton of the rapidly growing embryo. The cartilaginous skeletal elements present in fetal life are subsequently replaced by bone.

Hyaline cartilage, the most widely distributed form, has a pearl-gray semitranslucent matrix containing randomly oriented collagen fibrils, but relatively little elastin. In elastic cartilage, on the other hand, the matrix has a pale yellow appearance owing to the abundance of elastic fibres embedded in its substance. This variant of cartilage is more flexible than hyaline cartilage and is found principally in the external ear and in the larynx and epiglottis. The third type, called fibrocartilage, has a large proportion of dense collagen bundles oriented parallel. Its cells occupy lacunae that are often arranged in rows between the coarse bundles of collagen. It is found in intervertebral disks, at sites of attachment of tendons to bone, and in the articular disks of certain joints.

Bone. Like other connective tissues, bone consists of cells, fibres, and ground substance, but, in addition, the extracellular components are impregnated with minute crystals of calcium phosphate in the form of the mineral hydroxyapatite. The mineralization of the matrix is responsible for the hardness of bone. It also provides a large reserve of calcium that can be drawn upon to meet unusual needs for this element elsewhere in the body. The structural organization of bone is admirably adapted to give maximal strength for its weight-bearing function with minimum weight. There are bones strong enough to support the weight of an elephant and others light enough to give internal support and leverage for the wings of birds.

The shaft of a typical long bone of the skeleton has a central medullary or marrow cavity surrounded by a cortex of compact bone. Toward either end, the compact bone is gradually reduced to a thin peripheral layer and the interior is occupied by cancellous or spongy bone consisting of a tridimensional lattice of branching and anastomosing trabeculae. Compact bone is penetrated by blood vessels that course longitudinally within its substance. The bone tissue is arranged in concentric layers or lamellae around these vascular channels. The concentric lamellar units so formed are called osteons or haversian systems. The intercellular substance of the bony lamellae is largely collagen whose parallel fibres are arranged helically within each concentric layer of the osteon. The pitch of the helix changes from one lamella to the next. The longitudinal channel in the centre of each osteon is referred to as a haversian canal. The angular areas between the cylindrical osteons of compact bone are filled in with interstitial lamellae that are parallel but have no particular orientation with respect to the vascular channels. Similarly in the trabeculae of spongy bone, which are avascular, the lamellae are not arranged in osteons but appear as a mosaic of angular units each composed of parallel lamellae. The laminated construction of bone and the internal reinforcement of its matrix with oriented bundles of collagen, and with crystals of mineral within the collagen fibrils, all contribute to its remarkable properties as a strong but relatively light-weight skeletal material.

Bone cells

The cells of bone are of four kinds, osteocytes, osteoblasts, osteoclasts, and osteoprogenitor cells. Osteocytes, the principal cells of bone, are lodged in flattened lacunae located within or between the successive lamellae. Radiating from these lacunae are many slender channels (canaliculi) that join them to neighbouring lacunae. Compared to cartilage matrix, the heavily mineralized matrix of bone is relatively impermeable. The sys-

tem of intercommunicating canaliculi between lacunae is therefore thought to be essential to permit diffusion of nutrients from blood vessels in the haversian canals to the osteocytes in lacunae some distance away. These cells are metabolically active and are believed to play a significant role in the maintenance of the surrounding matrix.

Bone is constantly being resorbed and renewed. Large multinucleate cells, the osteoclasts, often occupy shallow excavations in the surface of bone and are believed to be the active agents of bone resorption. The sizable resorption cavities that they produce are then filled in by osteoblasts, the formative cells of bone that line the cavities and secrete amorphous matrix and collagen precursors that are deposited layer upon layer to fill in the cylindrical cavity with a new osteon. In this process, some of the osteoblasts become incarcerated in the newly formed matrix and are buried in the continued deposition of matrix by osteoblasts that remain on the surface. Thus osteoblasts gradually diminish in number as they become isolated in lacunae and transform into osteocytes. The marrow cavity and the haversian canals are lined by a tenuous cellular layer called the endosteum, and the outside of a bone is invested by a dense connective-tissue layer called the periosteum. Cells of the endosteum and of the inner aspect of the periosteum constitute a population of relatively undifferentiated osteoprogenitor cells. When stimulated they have the capacity to develop into osteoblasts and initiate bone deposition, as in the healing of fractures. In other circumstances, they can coalesce and give rise to osteoclasts that are active in removing excess bone callus after fracture repair, or form resorption cavities in the continual internal remodelling of bone.

The common impression that bone is a hard, unyielding, inert material is erroneous. It is a living, actively metabolizing tissue, constantly being renewed. Its structure is responsive not only to changing patterns of external stress but also to the nutritional state and the internal endocrine environment of the body.

BIBLIOGRAPHY. Additional information on connective tissues may be found in W. BLOOM and D.W. FAWCETT, *Textbook of Histology*, 9th ed., ch. 10, pp. 131–164, and ch. 12, pp. 347–357 (1968); and in A.W. HAM, *Histology*, 6th ed., pt. 3, ch. 10, pp. 205–227 (1969).

(D.W.F.)

Connective Tissue Diseases

Connective tissue is the collective designation for the tissues that provide the supportive framework and protective covering of the body and of its internal organs. It includes bone and periosteum—*i.e.*, the membranous covering of the bones; cartilage, tendon, and ligament (cartilage is translucent elastic tissue that forms much of the temporary skeleton of the embryo and persists in such locations as in the nose, in the external ear, and in joints; tendons are the fibrous cords by which muscles are attached to bones; ligaments are bands of tissue that support organs or connect bones); fascia—sheets or layers that ensheath muscles and viscera and are continuous with other connective tissue structures; and the dermis, the inner or true skin. In addition, connective tissue is a major component of the arteries and veins. The connective tissues serve as the site for the transport of essential nutrients and the collection of metabolic wastes, as well as the arena for immunologic and inflammatory reactions that protect the body from invasion by micro-organisms (see CONNECTIVE TISSUE, HUMAN).

CLASSIFICATION AND GENERAL CHARACTERISTICS, AND THEORIES OF CAUSATION

Classification. Diseases of the connective tissue can be divided into (1) a group of relatively uncommon, genetically determined disorders that affect the primary structure of this tissue, and (2) a number of acquired maladies in which the connective tissues constitute the site of several more or less distinctive immunological and inflammatory reactions. The latter conditions formerly were called collagen diseases because the connective tissues, which were found to show similar changes, were once known as the collagen or collagen–vascular system. The term collagen is now restricted to a specific fibrous protein, and there is little or no evidence in the majority of these disorders of any fundamental abnormality in the structure or metabolism of collagen; consequently, the designation collagen disease should be discontinued. The connective tissue diseases are now generally taken to include, besides certain heritable diseases, the following acquired diseases: rheumatoid arthritis, systemic lupus erythematosus, progressive systemic sclerosis (scleroderma), polymyositis (dermatomyositis), necrotizing vasculitides, Sjögren's syndrome (sicca syndrome), rheumatic fever, amyloidosis, thrombotic thrombocytopenic purpura, and relapsing polychondritis.

General characteristics. The above acquired diseases, which will be touched upon in later sections, display certain common clinical features, including inflammation of the joints (polyarthralgia and arthritis), serous (fluid-exuding) membranes (pleurisy, pericarditis), and small blood vessels (vasculitis), and a high frequency of involvement of various internal organs that are particularly rich in connective tissue (*e.g.*, the lungs). The walls of inflamed blood vessels, portions of which may become necrotic (*i.e.*, may die), are often found to contain characteristic deposits of hyaline (translucent) material. This is called fibrinoid because staining with dyes (*e.g.*, eosin) reveals tinctorial properties similar to fibrin (a fibrous protein that forms the lattice of blood clots). It was the demonstration of such hyaline deposits in various organs that was responsible, in large part, for the original grouping of these disorders as collagen diseases. Fibrinoid is not uniform in composition, however, but varies from one condition to another, and the similarity of structural changes in the connective tissues does not necessarily imply a common pathogenesis (disease origin).

Theories of causation. A number of observations suggest that acquired connective tissue diseases are auto-immune diseases—*i.e.*, diseases that result from reactions against components of the body as if they were foreign substances. In general terms these observations are that: (1) there are abnormally high levels of immunoglobulins in the blood; the immunoglobulins, also called gamma globulins, consist wholly or chiefly of antibodies; (2) these antibodies include several directed against particular serum proteins and other components of the affected person's own tissues; (3) there are complexes (combinations) of these antibodies and their antigens at the sites of tissue damage; (4) at the sites of tissue damage there are also aggregations of the cells (plasma cells and lymphocytes) that are responsible for the production of antibodies; (5) there is a favourable response to treatment with substances, such as corticosteroid hormones, known or believed to inhibit the production of antibodies; (6) connective tissue diseases are associated with other disorders known or suspected to be the result of an aberrant immune reaction.

The nature of the bodily components that serve to excite the production of antibodies in connective tissue diseases is the subject of continuing investigation.

An account of the effects of the interactions between antibodies and antigens may be found in the article ALLERGY AND ANAPHYLACTIC SHOCK. The effects include the damaging of cells to which antibodies have become attached so that the cells release substances that act upon the blood vessels and bring about such symptoms as those encountered in asthma or hay fever (allergic rhinitis) or in certain instances of drug allergy. The interactions may result in the destruction of red or white blood cells or platelets or may inactivate circulating hormones or enzymes. The antibody-antigen complexes may be deposited in the walls of blood vessels and there combine with a substance in the blood called complement, with a variety of injurious effects, including those seen in serum sickness, rheumatoid arthritis, and the kidney damage seen in systemic lupus erythematosus (see below). Last, the interaction may result in a complex reaction known as cellular immunity, which is thought to play an important role in certain auto-immune disorders that in-

Antibody-antigen interactions

volve solid organs, as well as in such processes as transplant rejection and cancer immunity.

HERITABLE DISORDERS OF CONNECTIVE TISSUE

Heritable disorders of connective tissue are a heterogeneous group of generalized, single-gene-determined disorders that affect one or another of the primary elements of the connective tissues (collagen, elastin, or ground substance [mucopolysaccharide]). Many cause skeletal and joint abnormalities that may interfere seriously with normal growth and development. All of these conditions are uncommon or rare when compared to the acquired connective tissue diseases.

Marfan's syndrome. Also called arachnodactyly (spider fingers), Marfan's syndrome is the least uncommon of the heritable disorders of connective tissue, having an estimated prevalence of about 15 cases per 1,000,000. The main skeletal characteristic is excessive length of the extremities. Weakness of joint capsules, ligaments, tendons, and fasciae is responsible for such manifestations as hyperextensible joints (double-jointedness), recurrent dislocations, spinal deformities, flat feet, hernias, and dislocation of the lens of the eye (ectopia lentis). The latter is present in at least three-quarters of the cases and usually occurs in both eyes. Cardiovascular abnormalities, which result from degenerative changes in the media (middle coat) of the great vessels, include insufficiency of the aortic valve and aneurysm (weakening of the wall and consequent bulging) of the ascending segment of the aorta.

Marfan's syndrome is inherited as an autosomal dominant trait; in other words, the gene involved is not a sex gene (all the genes that are not sex genes are autosomal). No more than 15 percent of cases occur as an isolated instance in a family and may be attributable to a new mutation. Death is usually due to heart failure or an aneurysm of the aorta. A normal life span is possible.

Although the basic abnormality of some unidentified component of connective tissue cannot be remedied, wound-healing occurs normally, and surgical correction of some of the defects is practicable.

Homocystinuria. Homocystinuria, so called because of the presence of homocystine in the urine, may closely resemble Marfan's syndrome. Distinctive from the latter, however, is the frequent occurrence of progressive mental deterioration, a fine skin with a tendency to flushing, osteoporosis (thinning of the bones), which may result in fractures, and thrombosis (blood clotting) of the coronary blood vessels and the peripheral blood vessels of medium size. The joints tend to have restricted mobility rather than hyperextensibility. Death from vascular occlusion is common during childhood, but persons with the disorder have survived into their 50s.

Homocystinuria is inherited as an autosomal recessive trait (it is not manifested unless inherited from both parents). Affected persons have a deficiency of an enzyme (cystathionine synthetase) required for the conversion of the amino acid cystathionine to cysteine. Treatment of affected infants with a diet low in methionine and high in cystine shows some promise, as does the regular administration of large amounts of pyridoxine, a component of the vitamin B complex.

Characteristics of Ehlers–Danlos syndrome

Ehlers–Danlos syndrome. The Ehlers–Danlos symptom complex is manifested particularly by an abnormal skin elasticity and fragility and by loose-jointedness. Some persons with this syndrome perform as contortionists. The skin, peculiarly stretchable even in early childhood, gradually loses its elasticity. Minor injury can cause lacerations that tend to bleed severely and to extend. In addition to the fragility of blood vessels in the skin, large vessels may develop aneurysms, varicosities, intervascular fistulas, or free rupture. Varicosities are enlargements of vessels over considerable lengths; they usually occur in superficial veins. Intervascular fistulas are openings between the walls of adjacent vessels. Scoliosis (lateral curvature of the spine), recurrent dislocations of joints, and hernias of the abdominal wall or the diaphragm (the muscular partition between the chest and the abdomen) are seen as in the Marfan syndrome, and there may be blue sclerae (the "whites" of the eye), such as occur in osteogenesis imperfecta (see the subsection below).

The underlying defect has not been defined. It is likely that it consists of the abnormal organization of collagen bundles, the individual fibres of which are normal. Some investigators have also detected an excess of elastin fibres. (Collagen and elastin are two of the fibrous proteins in connective tissue.) It is now clear that there are several distinct varieties of the Ehlers–Danlos syndrome. The disease is most commonly inherited as an autosomal dominant trait. Death from rupture of a major blood vessel may occur in childhood, but most affected persons live at least to middle age. Surgical repair of cutaneous, vascular, or other lesions is difficult because of the fragility of the tissues, which retain sutures poorly.

Osteogenesis imperfecta. Osteogenesis imperfecta is a general disorder of connective tissue that involves bone, sclera, inner ear, ligamentous structures, and skin. Several syndromes have been described, but they probably represent different degrees of expression of the same heritable disorder. The principal variants are osteogenesis imperfecta congenita, osteogenesis imperfecta tarda, and the van der Hoeve syndrome. In osteogenesis imperfecta congenita, stillbirth is common, or fractures are evident at birth; severe crippling occurs as a result of numerous fractures, and survival to adulthood is uncommon. In osteogenesis imperfecta tarda, the infant is normal at birth, but, depending on the severity of disease, few or many fractures occur over the following years, usually as a result of trivial mishaps. The frequency of fractures tends to diminish after puberty. The van der Hoeve syndrome consists of osteogenesis imperfecta tarda, with bluish sclerae and deafness. The peculiar colour of the sclera is due to its abnormal thinness, which permits the pigmentation of the choroid (the middle coat of the eyeball) to show. Hearing loss may be due to deformities of the bones of the inner ear as well as pressure on the auditory nerve because of deformity of its canal in the skull. Hyperextensibility of joints—double-jointedness—and abnormally thin skin are also characteristic.

Van der Hoeve syndrome

The fundamental defect in this disorder, which is usually inherited as an autosomal dominant trait, appears to involve the collagen fibres. Treatment is limited to surgical fixation of fractures.

Alkaptonuria. Alkaptonuria is a rare, inherited (recessive) disorder in which absence of the enzyme homogentisic acid oxidase results in an abnormal accumulation of homogentisic acid, which is a normal intermediate in the metabolism of the amino acid tyrosine. A portion of the homogentisic acid is excreted in the urine, to which, upon alkalinization and oxidation, it imparts a black colour. The remainder is deposited in cartilage and, to a lesser degree, in the skin and sclerae. The resultant darkening, or blackening, of these tissues by this pigment is termed ochronosis.

The pigment is bound particularly to collagen fibres in the deeper layers of joint cartilage and intervertebral disks (the fibrous pads between adjacent bones of the spine) and causes these tissues to lose their normal resiliency and become brittle. The erosion of the abnormal cartilages leads to a progressive degenerative disease of the joints, which usually becomes manifest by the fourth decade of life. Usually, the intervertebral disks become thinned and calcified first, and later the knees, shoulders, and hips are affected.

Alkaptonuria is compatible with normal life expectancy. Whether the arteriosclerosis (hardening of the arteries) that has been described in some cases is a manifestation of the disease or coincidental is uncertain. No effective treatment is available for the underlying metabolic disorder.

Pseudoxanthoma elasticum. Pseudoxanthoma elasticum, also known as the Grönblad–Strandberg syndrome, has its principal effects on the skin, the eyes, and the blood vessels. The word pseudoxanthoma refers to the yellowish papules (pimple-like protuberances) that occur most commonly in the fold of the skin of the neck, armpits, and groin. The colour results from degenerative

changes in the elastic fibres in the deep layers of the skin. Calcium deposition may occur in the skin, and premature arteriosclerosis is common. The characteristic eye lesion is that of angioid streaks (streaks resembling blood vessels) of the retina, which are found in at least 80 percent of the cases. Through the ophthalmoscope (an instrument for viewing the interior of the eye), the angioid streaks can be seen as brown or brownish-black bands that resemble the normal retinal vessels but are generally much wider. These streaks are produced by breaks in the elastin-rich internal membrane of the choroid, which permits the choroidal pigment to be seen. Deterioration of vision may occur because of bleeding or degenerative changes. Bleeding in the stomach is also fairly common. The basic lesion appears to be a premature deterioration of elastic fibres of connective tissue. Pseudoxanthoma elasticum is inherited as an autosomal recessive disorder. No specific treatment is available.

Angioid streaks in the eye

Cutis laxa. Cutis laxa is a rare disorder in which the skin hangs loosely in folds. In this disorder, unlike the Ehlers–Danlos syndrome, with which it is often confused, the skin is neither abnormally elastic nor unduly fragile, and the joints are not affected. There are several forms of the disorder, which are separable into inherited and acquired varieties. In the inherited variety there may be a characteristically hooked nose, with nostrils opening outward, and a long upper lip. Included among the complications are diverticula (abnormal pouches or pockets in the walls) of the gastrointestinal tract and bladder, hernias of the diaphragm or the abdominal wall, and rectal and vaginal prolapse (falling from a normal position). Present evidence indicates that cutis laxa is a disease of elastin, the collagen of the skin being normal.

Mucopolysaccharidoses. The mucopolysaccharidoses include six or more separate varieties of a disorder that, in varying degrees, affects the skeleton, brain, eyes, heart, and liver. Biochemically, the varieties have in common the abnormal production, storage, and excessive excretion of one or more mucopolysaccharides. (The mucopolysaccharides, which are complex, high molecular weight carbohydrates, form the chief constituent of the ground substance that fills the space between the connective tissue cells and fibres.)

Hurler's syndrome

Hurler's syndrome, or mucopolysaccharidosis 1, is the least rare and most rapidly fatal. Few children afflicted with it reach the age of ten. Abnormalities begin to appear when the infant is a few months old. The lower vertebrae develop abnormally, resulting in hunchback; the neck is usually short, and the head becomes large and misshapen. This is in part due to intrinsic maldevelopment of the skull, and in part results from hydrocephalus (the accumulation of abnormal amounts of fluid in the brain). Cerebral function deteriorates gradually. Various deformities of the extremities develop, accentuated by stiffness of the joints. The corneas become clouded. The liver and spleen enlarge because of deposition of mucopolysaccharides. Death most often results from heart failure, which is attributable to infiltration of heart muscle and coronary vessels with mucopolysaccharides.

The facial deformities and dwarfed, deformed bodies that occur characteristically in Hurler's syndrome and in mucopolysaccharidosis 2 (Hunter's syndrome) are referred to as gargoylism. Individuals with a mucopolysaccharidosis other than Hurler's syndrome commonly live to adulthood, but a normal life-span is unusual.

These diseases may be identified biochemically according to the type of mucopolysaccharide excreted excessively in the urine.

The mode of inheritance is autosomal recessive in all the types except mucopolysaccharidosis 2, which is sex-linked recessive (only males show the disease). Symptoms can be alleviated, but no cure is available.

Myositis ossificans progressiva. In myositis ossificans progressiva bone develops in tendons, fasciae, and striated (striped or voluntary) muscle. Skeletal growth is normal, although certain abnormalities occur in the majority of cases, particularly shortening of the thumbs or the great toes, or both. Symptoms usually begin in childhood and progress irregularly until the third decade. Lesions may begin abruptly with local tenderness, swelling, and fever, or develop very gradually, with increasing stiffness and firmness as the only symptoms. Involvement of the chewing muscles may make eating impossible. Death most often results from an infection.

Why certain connective tissues become sites for the aberrant formation of bone is completely unknown.

ACQUIRED CONNECTIVE TISSUE DISEASES

Rheumatoid arthritis. Rheumatoid arthritis is a chronic disease in which a predominant nonsuppurative inflammation (an inflammation that does not discharge pus) of the peripheral joints is frequently combined with a variety of manifestations not directly associated with the joints. The most prominent of these latter ones are inflammation of tendons and their sheaths and granulomatous nodules that occur in the connective tissue beneath the skin and beneath the periosteum. This, the most common of the inflammatory arthritides, occurs in individuals of all ages, including young children, and with particular frequency in young and middle-aged women; it is approximately three times as common in women as in men.

Incidence of rheumatoid arthritis

According to epidemiological surveys conducted in many different parts of the world, rheumatoid arthritis affects at least 1–2 percent of the population. It is estimated that there are, in the United States, 3,000,000–5,000,000 individuals with the disease, approximately 30 percent of whom are less than 55 years of age. A high rate of serious interference with joint function, combined with the tendency for the disease to occur among the young and middle-aged, makes rheumatoid arthritis the leading cause of occupational disability among the rheumatic diseases.

In the majority of cases the onset of rheumatoid arthritis is insidious, with generalized weakness, feeling of ill health, and aching and stiffness, often poorly localized to joints. This is followed by the gradual appearance of frank joint inflammation in the form of pain, swelling, redness, warmth, and tenderness. The symptoms usually originate in the small joints of the hands and feet but in most cases spread to other parts, especially the wrists, elbows, hips, knees, and subtalar joints (the subtalar, or talocalcanean, joint is in the ankle); in severe cases, virtually all of the peripheral junctures may be affected, as well as a number of other joints, including the jaw, the intervertebral joints, and even such fine structures as the cricoarytenoid joints of the larynx.

The many manifestations of rheumatoid arthritis not associated with the joints include evidence of blood-vessel inflammation in the form of tiny areas of necrosis in the fingertips; chronic leg ulcers and lesions in the peripheral nerves; inflammation of the pericardium, the sac enclosing the heart, and of the sclerae; inflammation and nodule formation in the lungs and pleura (tissue covering the lungs); anemia; enlargement of the lymph nodes; and Sjögren's, or sicca, syndrome (see below). Enlargement of the spleen occurs in approximately 5 percent of affected persons.

The disease process within the joints begins as an inflammation of the synovium (joint-lining tissue). In most cases there is an increase, often considerable, in the amount of synovial (joint) fluid.

The course of the disease varies greatly from person to person and is characterized by a striking tendency toward spontaneous remission (temporary recovery) and exacerbation. With continuing inflammation of the joints, there is destruction of the joint cartilage. The degree of articular (joint) disability present in rheumatoid arthritis depends in large measure upon the amount of damage done to this cartilage. If the injury is severe, large areas of bone may be denuded of cartilage, so that adhesions form between the articular surfaces. Subsequent transformation of these adhesions into mature fibrous or bony connective tissue leads to firm union between the bony surfaces (ankylosis), which interferes with motion of the joint and may render it totally immobile. In other instances, the loss of cartilage and bone, coupled with

weakening of tendons, ligaments, and other supporting structures, results in instability and partial dislocation of the joint. In a small minority of cases, the disease pursues a rapidly progressive course marked by relentless joint destruction and evidence of diffuse vasculitis (inflammation of blood vessels).

Causation of rheumatoid arthritis

There is now convincing evidence that immunologic reactions play an important role in the causation of the processes just listed. Suspicion of this originated with the observation that the serum of approximately 80–90 percent of persons with rheumatoid arthritis contains an immunoglobulin (called rheumatoid factor) that behaves as an antibody and reacts with another class of immunoglobulin. This immunoglobulin is produced by cells (plasma cells) that are present in sites of tissue injury. At this time the nature of the agent (antigen) responsible for the initiation of the rheumatoid inflammatory reaction is not clear. There is evidence to suggest that this agent may be one or more viruses or viral antigens that persist in the joint tissues.

Systemic lupus erythematosus. Systemic lupus erythematosus is a chronic inflammatory disease that affects, either singly or in combination, the skin, joints, kidneys, nervous system, serous membranes lining body cavities, and often other organs as well. The disease is characterized by a striking tendency toward remissions and exacerbations and by the presence of a multitude of immunologic abnormalities, including antibodies that react with various components of cell nuclei, as well as antibodies directed against circulating proteins, blood cells, and solid organs. The disease may develop at any period of life but appears with highest frequency during the second to fourth decades. The majority of affected persons are women.

Systemic lupus erythematosus exists in many forms, from the very mild to severe and rapidly fatal, and is far more common than once recognized. The annual incidence of the disease has been estimated to be at least three to six cases per 100,000 population.

This disorder, like many others, was first recognized by a characteristic rash, and the term lupus erythematosus (literally "red wolf") is based on the early confusion of this rash with certain disorders that commonly ended in ragged ulcerations of the skin and the erosion of underlying parts in a manner similar to the destruction wrought by a hungry animal.

Systemic and discoid lupus erythematosus must be distinguished. The latter is a localized disorder of the skin and is usually confined to a facial rash, which may be similar in appearance to that which occurs in the systemic disease. Only a small minority of persons with discoid lupus erythematosus (probably less than 5 percent) develop evidence of involvement of organs and structures other than the skin.

Characteristics of systemic lupus erythematosus

The identification of systemic lupus erythematosus is based primarily on certain clinical findings, the most specific and frequent of which include the following: (1) facial erythema (reddening), which often takes the form of a butterfly-shaped rash over the bridge of the nose and the cheeks; this rash occurs in only a minority of adults with the disease; (2) discoid lupus—an erythematous raised patchy eruption that heals with scarring and atrophy of the skin and may be found anywhere on the body; (3) Raynaud's phenomenon (see below); (4) photosensitivity, manifested by unusual skin reaction after exposure to sunlight; (5) nondeforming arthritis; (6) inflammation of the kidneys (glomerulonephritis); (7) pleurisy or pericarditis or both (inflammation of the chest lining or the membranous sac enclosing the heart); and (8) evidence of involvement of the central nervous system, in the form of psychosis or convulsions. In addition to this evidence of specific organ involvement, the affected persons also have constitutional symptoms—including fever, weakness, fatigability, and weight loss—that are often the first manifestation of illness.

Some degree of anemia (effects of deficiency of red blood cells or hemoglobin) is found in the majority of persons with the disease; in a number of cases, this is the result of an increased rate of red cell destruction attributable to antibodies that coat the cell and damage its membrane. Low white blood cell counts (leukopenia) and platelet counts (thrombocytopenia) are also characteristic; these, too, can often be traced to the presence of specific auto-antibodies. Abnormal bleeding may result from thrombocytopenia or from an antibody that combines with and inactivates certain plasma proteins (clotting factors) involved in blood coagulation. False positive serological tests for syphilis are found in up to 20 percent of patients; in a significant number of cases this, or other isolated serologic abnormalities, may be the first evidence of systemic lupus erythematosus and may antedate other developments by a period as long as several years. Rheumatoid factor (see the subsection above on rheumatoid arthritis) occurs in about 25 percent of cases. Most important from the standpoint of diagnosis are the antibodies that combine with various components of cell nuclei. One or more of these antinuclear antibodies are present in virtually 100 percent of persons with active disease. The first of these antibodies to be recognized was the lupus erythematosus cell factor, discovery of which permitted detection of many previously obscure and unrecognized forms of systemic lupus erythematosus and greatly expanded knowledge of the clinical spectrum of this disorder.

The compound thus formed is avidly ingested by certain phagocytic (particle-engulfing) white blood cells, neutrophilic leukocytes, and these cells, distended by the compound of antibody and of lymphocyte-nucleus component, are the characteristic lupus erythematosus cells. Other antibodies fairly characteristic of lupus erythematosus include two that react specifically with deoxyribonucleic acid (DNA) and double-stranded ribonucleic acid (RNA) of cell nuclei. The presence of these two antibodies is associated with clinically active disease, and in particular with inflammation of the kidney, and with skin and brain lesions.

Recent evidence strongly suggests the possibility that one or more viruses may be the ultimate source of the antigenic stimulation responsible for the development of these auto-antibodies.

The treatment of systemic lupus erythematosus is designed to reduce or control inflammation and to limit the damage done to vital organs. Salicylates (aspirin) are used to relieve pain, particularly when joints are involved, and to reduce fever. In most cases, however, it is necessary to employ corticosteroid hormone to reduce inflammation in acute crises of the disease. Certain antimalarial drugs, such as chloroquine and hydroxychloroquine, have been found to exert an anti-inflammatory effect on skin and joint lesions and are widely used for the treatment of milder forms of the disease.

The course of systemic lupus erythematosus is highly variable. Acute episodes occur, but more commonly the disease gives rise to a subacute or chronic illness that smoulders on for many months or years, subject to spontaneous remissions and exacerbations. There may be long intervals (up to 20 years or more) in which the affected person is entirely free of symptoms, with little or no evidence of the disease aside from serologic abnormalities, which tend to persist indefinitely. In most cases the prospects of survival are determined by the degree of kidney involvement and its responsiveness to treatment with corticosteroids and other measures. Persons with severe and persistent thrombocytopenia (deficiency of blood platelets) and hemolytic anemia or with central-nervous-system disease fare poorly compared to those whose illness mainly involves the joints.

Drug-induced lupus

There are a number of drugs that have been found to be responsible for the induction of a lupus-like disease. Typically, this state is manifested by the appearance of fever, joint pain, pleurisy, deficiency of white blood cells, and the development of various antinuclear antibodies (antibodies that interact with cell nuclei). The clinical features thus resemble closely those of natural lupus, with the exception that evidence of kidney involvement is notably rare and the fact that the symptoms generally disappear upon discontinuation of the offending drug. Drugs that have had this effect include hydralazine (used

in the treatment of arterial hypertension), procainamide (used for the control of irregular heart rhythms), various anticonvulsants, isoniazid (an antituberculous agent), the antibiotic penicillin, and penicillamine. In one study, these or other drugs were found to account for 3 to 12 percent of all the cases identified as lupus. Only a small proportion of persons who receive these drugs present evidence of a lupus-like disease, although, in the instance of procainamide, it appears that a much larger number develop antinuclear antibodies without (other) signs of illness.

Progressive systemic sclerosis. Progressive systemic sclerosis, or scleroderma, is a generalized disorder of connective tissue of uncertain causation characterized by inflammatory, fibrotic (increase of fibrous tissue), and degenerative changes in the skin, joints, muscles, and certain internal organs. The term scleroderma refers to the thickening and tightening of skin, by which the disease was first recognized. The disease affects women approximately twice as often as men. The initial symptoms, which usually appear in the third to fifth decade of life, include painless swelling or thickening of the skin of the hands and fingers, pain and stiffness of the joints (polyarthralgia)—often mistaken for rheumatoid arthritis—and paroxysmal blanching and cyanosis (becoming blue) of the fingers induced by exposure to cold (Raynaud's phenomenon). The skin changes may be restricted to the fingers (sclerodactyly) and face but often spread. The disease may remain confined to the skin for many months or years, but in most cases there is insidious involvement of the esophagus (gullet), intestinal tract, heart, and lungs. A number of affected persons develop changes in the vessels of the kidneys or lungs accompanied by severe systemic arterial and pulmonary hypertension (high blood pressure), respectively. The diagnosis is based on the typical clinical features, particularly the findings in the skin, and X-ray evidence indicative of a disturbance in esophageal or intestinal motility, or both, and otherwise unexplained pulmonary fibrosis. (The motility involved is the waves of contractions, or peristaltic movements, by which the esophageal and intestinal contents are moved along.) Confirmation is often sought by means of examination of specimens of skin, which discloses an increase in skin thickness because of deposition of compact collagen in the dermis, or true skin.

Outlook for persons with scleroderma

Although there may be spontaneous improvement in the condition of the skin, those persons with more diffuse scleroderma tend to lose the ability to straighten their fingers. In general, the outlook is poor in cases marked by rapid progression of dermal changes, intestinal malabsorption, or heart failure, and is particularly serious in those patients who have kidney involvement and the grave type of high blood pressure known as malignant hypertension. In many cases, however, the disease progresses extremely slowly. There is no drug that is generally accepted as effecting a cure. The use of corticosteroids has proved disappointing, although these agents may help in those persons in whom there is inflammation of the muscles (myositis).

Polymyositis. Polymyositis is characterized by inflammation and degeneration of skeletal muscle, especially the muscles of the shoulder and pelvic girdles. In dermatomyositis, a closely related disorder, polymyositis is accompanied by an inflammation of the skin. Polymyositis occurs in all ages and is slightly more common in women. The muscle disease is manifested primarily by weakness and later by atrophy and contractures. Involvement of the muscles of the esophagus and the larynx causes difficulty in swallowing and in uttering sounds. Heart muscle may be affected. The skin changes, which occur in approximately half the persons affected and are highly variable, include a characteristic puffiness around the eyes with lavender discoloration of the upper eyelids and sharply demarcated reddened patches overlying the small joints of the fingers, which heal with atrophy and whitening of the skin. Calcium deposits under the skin, around the joints, and in the muscles are a common sequel to the inflammation in skin and muscle; when extensive, the condition is called calcinosis universalis. Other less common features of the disease include Raynaud's phenomenon, polyarthritis, and interstitial fibrosis (formation of fibrous tissue between the air sacs) of the lung. In at least 15 percent of affected adults, especially those with reddening of the skin, cancers are present.

The diagnosis is supported by an increase in the serum levels of certain enzymes that are released into the bloodstream when there is active destruction of muscle fibres and is confirmed by microscopic examination of affected muscle. There is evidence to suggest that the injury of muscle is produced by a toxic substance produced by sensitized lymphocytes. Certain structures that appear to be viruses have been observed by electron microscopy in muscle cells from several cases of chronic polymyositis. It is not yet clear, however, whether such virus infection is primarily responsible for polymyositis or represents a secondary invasion of injured tissue.

Course of dermatomyositis

The course of the disease is highly variable. Children with dermatomyositis usually react more favourably to corticosteroids than do adults. It has recently been reported that persons with corticosteroid-resistant polymyositis may respond to the drug methotrexate. Significantly fewer adults than children with the disease survive.

Necrotizing vasculitides. The symptom complexes included in this category are characterized by inflammation of segments of blood vessels, chiefly small and medium-sized arteries. Clinical manifestations depend upon the site and severity of the arterial involvement and are therefore highly variable.

No single cause or disease mechanism has been identified. In some cases the lesions are similar to those encountered in human serum sickness and in animals given large amounts of foreign protein, in which conditions there is convincing evidence to link the disease to the deposition of immune (antigen-antibody) complexes in the walls of small blood vessels. An antigen (Australia antigen) associated with viral hepatitis (liver inflammation) has recently been found in the serum of several persons with polyarteritis nodosa, raising the possibility that some cases of polyarteritis may result from the deposition in blood vessels of immune complexes of viral antigen and antibody.

In polyarteritis nodosa, formerly known as periarteritis nodosa, inflammation and necrosis of small and medium-calibre arteries leads to local dilation and the formation of small aneurysms. The kidneys are the most frequently involved organs, and the disease is often first manifested by hypertension or other evidence of nephritis (kidney inflammation). Unexplained fever is common, as well as various combinations of abdominal pain, gastrointestinal bleeding, heart failure, disease of the peripheral nerves, asthma, or pneumonia. The diagnosis is established by the finding of characteristic inflammation of the blood vessels on examination of tissue specimens and sometimes by the X-ray demonstration of aneurysms of the renal arteries. Corticosteroids are the principal form of treatment, although this may aggravate the hypertension. Survival for more than one year after the onset of multiple-system manifestations is uncommon.

Hypersensitivity angiitis

Hypersensitivity angiitis tends to involve smaller blood vessels than those affected in polyarteritis nodosa. Frequently, the affected person seems to have experienced hypersensitivity to various drugs, particularly penicillin, sulfonamides, and iodides. Treatment is with corticosteroids, and the outlook tends to be somewhat better than in polyarteritis nodosa.

Wegener's granulomatosis is a disorder marked by the combination of granulomatous lesions of the upper air passages and lower respiratory tract; destructive inflammation of blood vessels, both arteries and veins, especially in the lungs; and localized kidney disease. The disease affects adults of either sex. Corticosteroid therapy is often ineffective, and, in the past, most affected persons succumbed of kidney failure or lung infection within a year. Recently, complete, long-lasting recovery following treatment with various immunosuppressive agents has been reported.

Takayasu's arteritis, with variants called pulseless disease, branchial arteritis, and giant-cell arteritis of the

aorta, involves principally the thoracic (chest portion of the) aorta and the adjacent segments of its large branches. Symptoms and signs, including obliteration of the pulses in the arms, are related to narrowing and obstruction of these vessels. Most reported cases have occurred in young Oriental women. The diagnosis and extent of vascular involvement can be established by means of angiography (X-ray observation of the blood vessels). Corticosteroids administered early during the course of illness have a beneficial effect, accompanied on occasion by return of pulses. Anticoagulants may prevent thrombosis (formation of blood clots).

Giant-cell or temporal arteritis occurs chiefly in older people and is manifested by severe temporal or occipital headache (headache in the temples or at the back of the head), mental aberrations, visual difficulties, unexplained fever, anemia, aching pains and weakness in the muscles of shoulder and pelvic girdles (polymyalgia rheumatica), and—in a minority of cases—tenderness and nodularity of the temporal artery. This vessel is the site of an inflammation that is characterized by the presence of numerous giant cells. Treatment with small doses of corticosteroids usually leads to dramatic relief of symptoms.

Sjögren's syndrome. Sjögren's syndrome, or sicca syndrome, is a chronic inflammatory disorder characterized by severe dryness of the eyes and mouth that results from a diminution in secretion of tears and saliva. Dryness may also involve the nose, pharynx, larynx, and tracheobronchial tree. Approximately half the persons affected also have rheumatoid arthritis, or, less commonly, some other connective tissue disease, such as scleroderma, polymyositis, or systemic lupus erythematosus. The great majority of the persons affected are women.

Features of Sjögren's syndrome

Infiltration (gradual assemblage) of lymphocytes and plasma cells leads to enlargement of the parotid or other salivary glands in half the patients. There may also be enlargement of the spleen, diminution in white blood cell numbers, Raynaud's phenomenon, vasculitis (inflammation of vessels) with chronic leg ulcers, a disease of the peripheral or trigeminal nerves, nonthrombocytopenic purpura (purpura is the presence in the skin of a number of red spots from the escape of blood into the tissues from small blood vessels; if it is nonthrombocytopenic, it is not due to deficiency of blood platelets), chronic (Hashimoto's) thyroiditis (inflammation of the thyroid), enlargement of the liver, and inflammation of the pancreas. A number of persons with sicca syndrome of long duration have developed neoplasms of the type called reticulum cell sarcoma, or primary macroglobulinemia (the presence in the blood of globulins of high molecular weight). The frequent occurrence of elevated levels of immunoglobulin in the serum and of various antibodies, including antisalivary duct antibody, antinuclear antibody, and antiglobulin antibody (rheumatoid factor) has strengthened suspicion that Sjögren's syndrome is an autoimmune disorder. Treatment is directed toward relief of symptoms. The ocular dryness may be relieved by the use of artificial tears. Corticosteroids or immunosuppressive drugs have been employed with some success for the more serious manifestations.

Rheumatic fever. Rheumatic fever is an inflammatory disease that represents a delayed sequel to infection with the group A hemolytic streptococcus and predominantly affects children between the ages of five and 15 years. Although its name is based upon involvement of the joints, rheumatic fever owes its importance to its tendency to damage the heart. The attack rate of rheumatic fever ranges as high as 3 percent in cases in which streptococcal infection is associated with sore throat and pharyngeal exudate (oozing from the throat surfaces). The occurrence of the disease does not confer immunity to second attacks, and persons who have had it are more susceptible to recurrences than the general population is to an initial attack.

Rheumatic fever may be gradual and unnoticed in onset, or it may develop rapidly. Typically, clinical evidence of the disease appears after a symptom-free latent interval of a few days to several weeks after the inciting streptococcal infection. The major indications of its presence in children include inflammation of the heart (especially the valves, manifested by heart murmurs), arthritis, chorea (a nervous disorder involving unceasing, involuntary movements), subcutaneous nodules, and skin rashes, the most characteristic of which is erythema marginatum (reddening of the skin in disk-shaped areas with elevated edges). Fever is common, but not invariably present. Symptoms in adults are usually confined to the heart, the joints, or both. Antibiotics, especially penicillin, are employed during the attack to eradicate the streptococci, whereas aspirin and corticosteroids are used to treat the acute symptoms. Both decrease discomfort but neither shortens the course of illness. Occasionally, death occurs from an overwhelming inflammation of the heart, but, unless there is damage to heart valves, recovery usually is complete. Scarring and deformity of the valves may lead to their narrowing or failure to close properly, and this may eventually lead to the development of heart failure when the heart muscle is unable to keep up with the extra work load imposed by the abnormal functioning of the valves. The prophylactic use of antibiotics (chiefly penicillin) has led to a dramatic reduction in the frequency of streptococcal infections and resultant recurrences of rheumatic fever.

Complications of rheumatic fever

Several antibodies against the streptococcus develop in response to infection. These are directed against various constituents of the micro-organism or its products. The mechanisms whereby streptococcal infection initiates the process of rheumatic fever appear to be immunological in nature and to be based in large measure on antigenic cross-reactive relationships between a protein constituent of the streptococcal cell walls and human heart tissue and between other fractions of the streptococcus and a component of joint cartilage. Immunoglobulins produced in response to these bacterial antigens may act as autoantibodies and be responsible for the inflammation of the heart and joints.

Amyloidosis. Amyloidosis is a disorder characterized by the accumulation of the substance amyloid in the connective tissue. Amyloid's name was originally assigned it in the mistaken belief that this material is akin to starch. Amyloid consists of a filamentous protein that is derived from immunoglobulins. The deposition of this substance may be widespread, with involvement of major organs leading to serious clinical consequences, or it may be very limited in extent with little effect on health. Amyloidosis has been separated into a primary form, unrelated to any other disease, and a secondary form, which is associated with chronic infections and inflammatory disorders. In addition, it appears that amyloid is related to the phenomenon of aging in that deposits are found with increasing frequency in the heart and brain of individuals past the age of 70. Included in the category of primary amyloidosis are a number of heritable forms, which have been reported from many different parts of the world. As seen by electron microscopy, all the types of amyloid examined consist chiefly of fine fibrils (minute fibres).

Manifestations and diagnosis of amyloidosis

Because any organ may be involved, the clinical manifestations of amyloidosis are varied. Accumulation of amyloid in and about peripheral nerves and the skin, tongue, joints, and heart is particularly common in the primary forms, whereas in secondary amyloidosis the liver, spleen, and kidney are the sites chiefly involved. The diagnosis is established by means of examination of tissue specimens. Correction of the underlying infection or inflammation may prove helpful in the amelioration of secondary amyloidosis, but the primary types remain untreatable.

Osteoarthritis. Osteoarthritis, also called degenerative joint disease, is a ubiquitous noninflammatory disease of joints; the weight-bearing joints are particularly affected, including the knees and the hips. The disease is characterized by the progressive deterioration of joint cartilage and by the reactive formation of dense bone and of bony projections at the margins of the joint. Although osteoarthritis remains the most popular name for this condition, the term is inaccurate because its suffix, "-itis," implies the existence of a basically inflammatory disorder.

Osteoarthritic changes have been noted in skeletal re-

mains of Neanderthal man (40,000 BC) and in a wide variety of animal species, both large and small, including fossil dinosaur skeletons. Some erosion of joint cartilage is virtually universal in the elderly and, indeed, appears to be an inherent part of the aging process. Recent surveys in the United States and Great Britain are the basis of estimates that 40–50 percent of the adult population have X-ray visible changes of osteoarthritis in the hands or feet and that approximately 5,000,000–10,000,000 Americans have symptoms as a result of these changes; thus, osteoarthritis is by far the most common form of joint disease.

In addition to the aging process, there are a variety of predisposing conditions as well as local joint factors that influence the site and severity of degeneration of joint cartilage. These include excessive wear and tear related to occupation, injury, developmental structural abnormalities, and obesity; certain metabolic disorders that affect articular cartilage directly; repeated bleeding in a joint, such as occurs in hemophilia; and impaired proprioception (reception of sensory stimuli from within the body) of the joints, such as occurs in joint disorders resulting from nervous-system diseases (Charcot joints).

Heberden's nodes

Finally, it appears that for some forms of osteoarthritis —*e.g.*, so-called Heberden's nodes—there exists a genetically determined predisposition. Heberden's nodes are bony protuberances (osteophytes), which are found at the margins and on the dorsal surface of the terminal joints of the fingers. The nodes develop more frequently in women, tend to occur in families, and are often associated with degenerative change in other joints.

Osteoarthritis is considered more at length in JOINT DISEASES AND INJURIES.

Thrombotic thrombocytopenic purpura. Thrombotic thrombocytopenic purpura is a rare disorder that has been included with the connective tissue diseases chiefly because of certain clinical similarities to systemic lupus erythematosus. The main features of this disorder, which usually appears suddenly and intensely in young women, include thrombocytopenic purpura (presence in the skin of red spots from the escape of blood into the tissues as a result of scarcity of blood platelets), hemolytic anemia (anemia resulting from destruction of red cells), changing neurological manifestations, fever, and kidney failure. The principal lesion is that of widespread blockage of small blood vessels—arterioles, venules, and capillaries—by material consisting principally of fibrin, the principal constituent of blood clots. The heart, kidneys, and brain are particularly affected. Corticosteroids remain the principal agent in treatment. Few persons survive longer than a month after onset of the disease.

Relapsing polychondritis. Relapsing polychondritis is an uncommon inflammatory disease that primarily affects cartilages. It begins usually in the fourth or fifth decade and is marked by recurrent periods of inflammation of various cartilages, lasting several weeks to months. The external ear and nose are affected most frequently and are eventually disfigured ("cauliflower ear") in a high percentage of cases. Involvement of joint cartilages produces pain and swelling of the joints, and the destruction of these cartilages results in a degenerative joint disease that may be disabling. Most serious is involvement of the trachea (windpipe), which may lead to respiratory obstruction or recurrent pneumonia. Hearing may be impaired when the inner ear is involved. About half of the affected persons have inflammation of the eyes. The aorta and the ring of the aortic valve may be involved. The acute manifestations of the disease can usually be suppressed with corticosteroid therapy, but the changes in the cartilages are permanent.

BIBLIOGRAPHY. Detailed accounts of the connective tissue diseases and of the rheumatic disorders in general may be found in JOSEPH L. HOLLANDER (ed.), *Arthritis and Allied Conditions*, 8th ed. (1972); W.S.C. COPEMAN (ed.), *Textbook of the Rheumatic Diseases*, 4th ed. (1969); and JAMES A. BOYLE and W. WATSON BUCHANAN, *Clinical Rheumatology* (1971). VICTOR A. MCKUSICK, *Heritable Disorders of Connective Tissue*, 3rd ed. (1966), is the most complete and authoritative text dealing with the genetically transmitted connective tissue diseases. IAN R. MACKAY and F. MACFARLANE BURNET, *Autoimmune Diseases* (1963), was one of the first to clearly formulate and champion the role of autoimmunity in the connective tissue diseases and a number of other maladies of obscure etiology. MAX SAMTER (ed.), *Immunological Diseases*, 2nd ed., 2 vol. (1971), describes the mechanisms whereby aberrant immunity leads to tissue damage. Briefer, clinically oriented accounts may be found in MICHAEL MASON and HARRY L.F. CURREY (eds.), *An Introduction to Clinical Rheumatology* (1970), and the *Primer on the Rheumatic Diseases*, 7th ed. (1972). LEON SOKOLOFF, *The Biology of Degenerative Joint Disease* (1969), considers the comparative pathology of this ubiquitous arthropathy and reviews recent important experimental studies. EDMUND L. DUBOIS (ed.), *Lupus Erythematosus* (1966), provides an exhaustive summary of the condition that has been influential in the development of the modern concept of connective tissue disease.

(G.P.R./T.G.B.)

Conodonts

The conodonts, toothlike microfossils consisting of calcium phosphate, are among the most frequently occurring fossils in marine sedimentary rocks of Paleozoic age. The organisms to which the conodonts belonged are unknown and probably have been extinct since the middle or late Mesozoic Era (65,000,000 to 225,000,000 years ago).

Initial discovery

Conodonts were first found and described in 1856 in Lower Ordovician greensand in the East Baltic area. They were at first thought to be the teeth of primitive fish, and several authors since then have endorsed the fish hypothesis. Others, however, have pointed out the similarity between certain conodonts and the jaws of the worm group polychaetes (see ANNELIDA). The polychaete, or annelid, hypothesis became particularly attractive after the discovery (1934) that several kinds of conodonts must have functioned together in individual animals as a kind of mechanism resembling the jaw apparatus of polychaetes.

The conodonts also have been regarded as internal, skeletal structures set within soft tissue. At least one worker has assumed that they supported the food-gathering organ (consisting of a lophophore or braccia appendage) of a free-swimming lophophorate animal, surrounding the mouth and extending into the mouth cavity. They have also been compared with the radular teeth (muscular food-tearing mechanism) of gastropods, with parts of crustaceans, and with the copulatory structures (spicules) of nematodes. It has even been claimed that they do not belong to the animal kingdom at all but are algal structures.

Conodonts occur in all kinds of sedimentary rocks that have their origin in the sea, and have been found at practically all stratigraphic levels of their historical range. Because they permit precise and accurate correlation of rock sequences over great distances, they are of considerable practical importance, particularly in defining stratigraphic boundaries for petroleum exploration.

This article treats the morphology and physical properties of conodonts, describes their classification, geological occurrence, and stratigraphic use, and presents the several theories of origin of conodonts. For further information on the relationship between conodonts and other fossils, see FOSSIL RECORD. See also STRATIGRAPHIC BOUNDARIES for a discussion of the types of problems in which correlation by conodonts is useful.

THE NATURE OF CONODONTS

Among the well-known hard parts of fossil or living animals, those that bear the greatest resemblance to conodonts are those elements of the polychaete jaw apparatus known as scolecodonts (worm mandibles) and the teeth of fish. Scolecodonts and conodonts are distinguished from each other because the former consist of acid-resistant organic material, whereas the conodonts, phosphatic, are soluble in hydrochloric acid. In rocks that have not undergone high-temperature alteration, the conodonts are light yellow to yellowish brown and at least partially translucent, whereas scolecodonts are dark brown to black and are opaque.

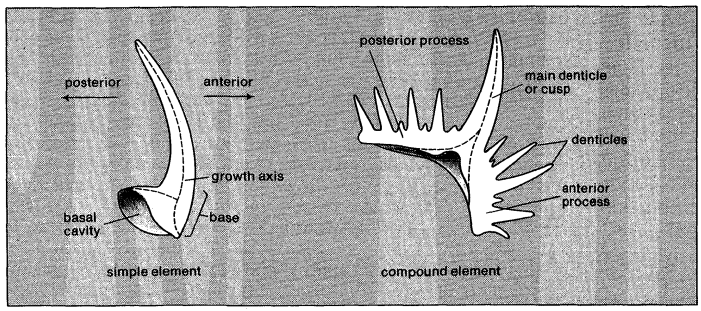


Figure 1: Simple and compound elements and conodont nomenclature.

There are also differences in outer shape between conodonts and scolecodonts. The denticles (small spinelike teeth) that rise from the base of conodonts may be long and slender and occur either in branches or in several parallel rows, whereas denticles of the scolecodonts are relatively short and occur in a single row. Platform conodonts may have a tuberculate or pitted surface, whereas scolecodonts have a smooth one.

Fish teeth and conodonts are similar in composition. The outer resemblance between certain stout and simply constructed conodonts and teeth may be great. Fish teeth, however, do not have the thin and strongly recurved denticles and the long, narrow denticulate processes found in many conodonts. The most important distinguishing characters lie, however, in the inner structure. Only moderate magnification is sufficient to show that well-preserved conodonts consist of thin concentric lamellae, without a trace of the bony tissue and blood-vessel canals that occur in teeth.

Arthropods and certain brachiopods may yield fragments that are mistaken for conodonts, but the former are nonphosphatic or, if phosphatized, lack the finely layered structure of conodonts. Phosphatic brachiopod shells may carry denticulate structures resembling conodonts. With respect to certain primitive, conical fossils with a phosphatic composition, the question of their place within the classification of animals is still open.

Chemical and structural distinctiveness

Chemical composition and internal structure. The chemical composition of conodont phosphate contains no hydroxyl groups and approaches that of the phosphate mineral apatite. Minor amounts of organic compounds are also present; amino acids from Devonian conodonts were identified in 1966. Leucine, phenylalanine, and lysine have been found to be particularly important constituents of the conodonts. Fossil bone and teeth differ significantly in this respect. Because the environments of the material studied were rather special, it is possible that different results may be obtained in the future.

Though it has been customary to make a structural distinction between lamellar and fibrous conodonts, the distinction may be a false one. One argument favouring a structural distinction is the fact that in fracturing, lamellar conodonts break straight across the denticles, whereas in fibrous forms the fractured denticles may give lengthwise splinters, like wood. Another argument is that lamellar conodonts have opaque parts consisting of white matter, and in well-preserved specimens the entire fibrous conodont is more or less translucent. Certain well-defined genera contain only translucent elements, which break in the fibrous way. Working against the validity of a structural distinction is the fact that fibrous conodonts actually have a concentric lamellar structure similar to that of the main lamellar group, and typically lamellar conodonts have a lengthwise fibrous structure within the lamellae.

The lamellae, usually a few microns thick, may be separated by narrow interlamellar spaces and consist of columnar phosphate crystallites with dimensions considerably less than one micron. Judging from microscopic observation, the crystallites are probably arranged at right angles to the lamellar surfaces.

Whereas the base of well-preserved specimens is translucent, the denticles may contain white matter that does not let light through. The lamellar structure probably continues through the white matter, although this cannot easily be ascertained. Electron-microscopic studies indicate that the white matter owes its aspect to closely set empty spaces, the diameters of which may be about one micron. The hollow spaces are known to cut discordantly through the fabric of crystallites, and it has been suggested that this might indicate that they were formed by resorption.

Morphology and nomenclature. Distinctive forms of conodonts are called elements. An element normally consists of a base that carries one or several denticles (Figure 1). Conventionally, the denticles are oriented upward. On the lower side of the base many conodonts, including the most primitive, have a conical basal cavity. The oldest part of the element is at the tip of the basal cavity. This part is usually somewhat curved. The convex side is said to be anterior (forward), the concave side posterior (backward). This terminology is conventional and arbitrary, because the orientation in the animal is not known. As a rule the denticles also curve toward the posterior. The lamellae about this oldest part of the element are concentric. They are interrupted by the basal cavity, the interior of which shows the free margins of successive lamellae. In certain platform conodonts the basal cavity is replaced by a flat or slightly convex area, the attachment surface, or escutcheon, on which the lamellae appear as concentric striations. Even these elements originate with a basal cavity that remains as a small pit at the centre of the escutcheon.

The basal cavity or the escutcheon serves as an attachment surface of the basal filling. In simple and compound conodonts, this may be a hollow cone (basal funnel) inserted in the basal cavity. Platform conodonts may have a padlike basal filling. The basal filling is lamellar and phosphatic, but with somewhat less phosphate and more organic matter than the conodont. The crystallites are smaller than in the conodont. In most specimens the basal filling has been lost.

Conodont elements

The three principal kinds of conodont elements are simple, compound, and platform. The distinction depends on the development of the base and the number and arrangement of the denticles.

The simple conodonts consist of a main denticle (cusp), base, and basal cavity. In certain genera they are patterned, for instance, with lengthwise ridges and furrows.

In compound elements, the base carries more than one denticle. The main denticle, or cusp, originates at the tip of the basal cavity. The base develops branches (processes or bars), and denticles are aligned on the upper, or outer, side of the processes. This side is called oral, which suggests orientation in a mouth. The term is, however, due to historical usage and gives no indication of place or function. The basal cavity (escutcheon) continues along the lower (aboral) side of the processes.

Compound elements with one, two, three, or four processes, bilaterally symmetrical or asymmetrical, may belong to a single species that normally contained all these kinds of elements, as well as transitions between them (symmetry transitions).

Platform conodonts are compound forms in which the oral surface is widened to the sides and carries some kind of ornamentation, usually crests, rows of denticles, or

tubercles, as well as a finely reticulate pattern. In some of the most important groups of platform conodonts, there is a main row of denticles with the oldest part near the centre. In the anterior part the denticles may be fused so as to form a high blade. The platform usually occupies the posterior part of the unit, though the superfamilies Prioniodontacea and Bryantodontacea have platforms anteriorly as well as posteriorly, and the Gondolellacea have the platform practically restricted to the anterior process. On the platform, the main denticle row is low and may be fused into a ridge, or carina.

The platform may develop in either of two ways. The sides of the processes may thicken into ledges and develop platforms by further sideward thickening, as occurs in the polygnathids. On the aboral side the basal cavity is transformed into an escutcheon with central pit. The other possibility, represented by the idiognathodontids, is that the whole base expands, including the basal cavity. The oral surface of the expanded posterior part of the basal cavity then provides the space for platform ornamentation.

Conodont assemblages. Conodonts are collected as separate elements. Until 1934 it was the general belief, expressed in classification, that a conodont-carrying animal had only one kind of element. The discovery of natural assemblages on shale surfaces of Pennsylvanian (Upper Carboniferous) age, made independently in North America and Europe, showed that elements that were until then regarded as different genera could belong to a single individual. The composition of the most frequently occurring Pennsylvanian assemblages has been found to be fairly constant: a pair of platform elements, right and left; a pair of bladelike compound elements classified as *Ozarkodina;* a pair of pick-shaped compound elements (*Neoprioniodus* or *Synprioniodina*); and four or five pairs of long, barlike compound elements (*Hindeodella*). Assemblages have been found in different shales of Pennsylvanian and Mississippian age, recently also in Upper Devonian Kellwasser Limestone in West Germany. The elements of the assemblages may be arranged in a more or less orderly fashion. Amorphous organic material may be concentrated around the assemblages. Otherwise, no recognizable organic structure other than the conodonts occurs.

Because the elements occur in constant proportions, it should be possible to reconstruct assemblages from series of samples of isolated elements. This method has been successfully applied to samples from Ordovician-age rocks. Probable assemblages have also been outlined for the Silurian. Most of the reconstructed assemblages are comparable to the classical assemblages found in the Pennsylvanian.

In order to avoid confusion with the taxonomic nomenclature, conodonts found in assemblage are referred to by adding an ending to the name of the genus to which they would have been referred in the old taxonomy (thus, polygnathiform, ozarkodiniform, or neoprioniodiform elements, etc.).

THE CLASSIFICATION OF CONODONTS

Problems of classification

A taxonomy founded on isolated elements persisted after the discovery of assemblages, though researchers became increasingly aware that it was artificial. Efforts to stabilize the traditional taxonomy as a form taxonomy (based on similarity between species and genera only in form) were frustrated in 1958 by the International Congress of Zoology. Since then, a taxonomy based on isolated elements, in reality though not by right a form taxonomy, has been further elaborated, while at the same time a taxonomy based on natural assemblages (in which several conodont elements occur in the same species or genera) gradually has taken its place. The decision of the International Congress of Zoology means that the same genus and species names must be used in both classifications until the form-taxonomic approach has been completely abandoned.

Taxonomic units above the level of the genus cannot be more securely founded than the genera themselves. A classification in the *Treatise on Invertebrate Paleontology* recognizes seven families (Distacodontidae, Belodontidae, Coleodontidae, Prioniodinidae, Prioniodontidae, Polygnathidae, and Idiognathodontidae), all of which are defined as utilitarian. These families are too heterogeneous to be useful. Most conodont publications therefore leave the question of suprageneric classification aside.

In 1944 a classification that appears to indicate some basic relationships was proposed. In this scheme, the order Conodontophorida, of uncertain affinity, comprises the two suborders Neurodontiformes and Conodontiformes. Neurodontiformes, fibrous conodonts, contains the families Chirognathidae, Coleodontidae, and Trucherognathidae. Conodontiformes, or lamellar conodonts, is divided into the families Distacodidae (correctly: Distacodontidae), Prioniodidae (correctly: Prioniodontidae), Prioniodinidae, Polygnathidae, and Gnathodontidae (younger synonym of Idiognathodontidae). The validity of this classification is unaffected by the circumstance that the structural difference between fibrous and lamellar conodonts was misunderstood. The definitions of the families are different from those in the utilitarian *Treatise* classification. Both classifications falter because they ignore the assemblages, however.

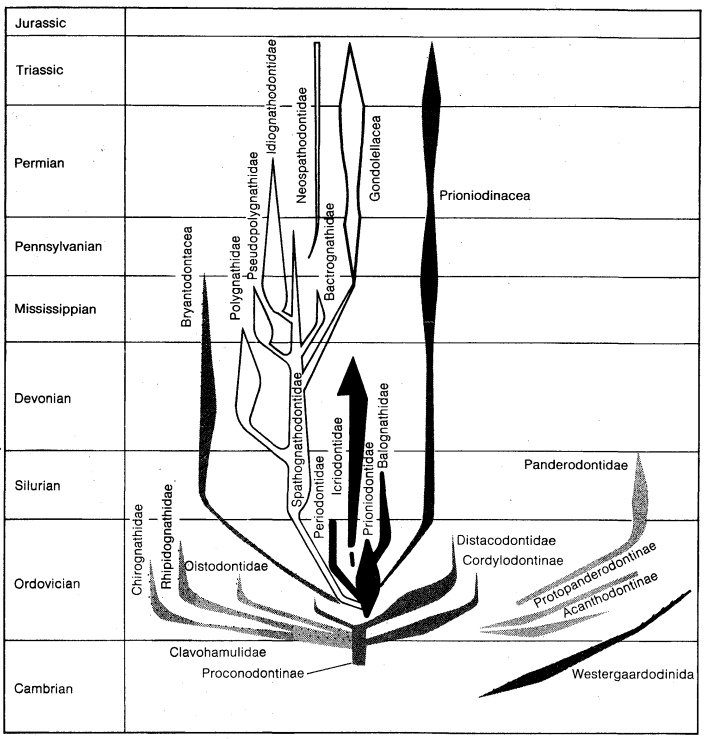


Figure 2: Phylogeny of main superfamilies, families, and subfamilies of conodonts.

Biological classification

The available criteria for a classification are outer shape; micromorphology; internal structure, including the distribution and structure of white matter; and the composition of the natural assemblage or conodont apparatus. The following classification is based on these criteria.

Order Westergaardodinida. U- or W-shaped, or conodont-like structures, consisting mainly of organic material, with phosphate as a minor component. (Middle?–) Upper Cambrian to Middle Ordovician.

Order Conodontophorida. Conodonts in the strict sense. Upper Cambrian to Upper Triassic (to Upper Cretaceous?).

1. Superfamily Distacodontacea
   Family Proconodontidae. Subfamilies: Proconodontinae and Cordylodontinae.
   Family Clavohamulidae
   Family Distacodontidae
2. Superfamily Chirognathacea (= "Suborder Neurodontiformes")
   Family Oistodontidae
   Family Chirognathidae

Family Rhipidognathidae
3. Superfamily Panderodontacea
Family Acanthodontidae. Subfamilies: Acanthodontinae and Protopanderodontinae
Family Panderodontidae
4. Superfamily Prioniodinacea
5. Superfamily Prioniodontacea
Family Periodontidae
Family Prioniodontidae
Family Balognathidae
Family Icriodontidae
6. Superfamily Bryantodontacea
7. Superfamily Gondolellacea
8. Superfamily Polygnathacea
Family Spathognathodontidae
Family Polygnathidae
Family Pseudopolygnathidae
Family Bactrognathidae
Family Idiognathodontidae
Family Neospathodontidae

Superfamilies 4–8 are closely related; the Prioniodontacea probably evolved from distacodontids in the Early Ordovician.

THE GEOLOGICAL OCCURRENCE OF CONODONTS

Geologic ranges

The oldest fossils claimed to be conodonts occur in the Middle Cambrian Kalby Clay of Bornholm, Denmark. They are horn-shaped, hollow objects, evidently consisting of phosphate. Because no further diagnostic characters are known, it is uncertain whether these fossils are conodonts. The Upper Cambrian contains a group of more or less conodont-like fossils here brought to the order Westergaardodinida. They have a lamellar structure and consist of organic material and phosphate. The colour is black or dark brown. Only simple forms and U- or W-shaped structures are known. Westergaardodinida are not known from beds younger than Middle Ordovician. Workers on these fossils have regarded them as primitive conodonts, but this supposition has been questioned.

Conodonts in the strictest sense appear in the uppermost part of the Upper Cambrian (Figure 2). They are simple forms belonging to the Proconodontidae. In the Ordovician, conodonts become very frequent and diversified. Simple distacodontids are dominant forms in many Lower Ordovician faunas. The most long-lived major group of simple forms is that of the panderodontids that are common throughout the Silurian.

Compound forms belonging to the Prioniodontacea are characteristic elements of most Ordovician faunas. The Devonian, like the Ordovician, contains exceptionally rich conodont faunas. These are often found to be dominated by the Polygnathacea. The continuous conodont record ends in the Upper Triassic, and reports on Jurassic conodonts appear to be doubtful.

Conodonts have been discovered in Upper Cretaceous beds at the Mungo River in Nigeria. The species are the same as in the Upper Triassic. From the taxonomic and evolutionary point of view, it is therefore most likely that the fauna is reworked from Upper Triassic beds. No such beds are known in the region, however, and it may be difficult to reconcile the existence of Triassic marine fossils in this area with widely accepted ideas on continental drift (*q.v.*), according to which the South Atlantic Ocean did not exist until the Cretaceous.

Conodonts are among the most useful index, or guide, fossils of the geological systems in which they occur. They are easy to recover in quantities sufficient for stratigraphic purposes. The standard method is to dissolve samples from marine limestone beds in diluted acetic or monochloroacetic acid. Acetic acid works more slowly but more gently than the latter. The conodonts are screened from the undigested residue. A good yield may be about 100 conodonts in a one-kilogram (2.2-pound) sample, although certain samples from the Ordovician and Devonian may contain over 10,000 conodonts in one kilogram of rock. Determinable conodonts also may be found in sandstone and on shale surfaces.

The rapid evolution of certain conodonts, most notably the platform elements, has provided a great number of easily recognizable index species. The short life of such species gives precision to age determinations by conodonts. The near-ubiquity of conodonts in marine deposits is a further reason for their usefulness as stratigraphic indicators.

Guide fossils

Rapid dispersal kept pace with rapid evolution, allowing conodont species to spread over the whole world. In the Devonian and Mississippian, practically identical conodont successions have been encountered in Europe, North America, and Australia. Similar conditions are known from the other systems, though, for instance, in the Ordovician certain conodont faunas are geographically restricted. Thus, some early prioniodontaceans of the Lower Ordovician are rare outside a North European-Atlantic province, and the genus *Phragmodus*, a Middle–Upper Ordovician periodont, seems to be restricted to areas outside this province (*e.g.*, central North America). *Cavusgnathus*, a Lower Carboniferous idiognathodont, is missing locally in western Europe. Nevertheless, conodonts are exceptionally useful for intercontinental as well as local correlation.

The lowermost Ordovician is characterized by proconodontids; *e.g.*, the genus *Cordylodus*. In the Lower Ordovician of Europe, index fossils are found among the distacodontids and prioniodontids. The Lower Ordovician of other geographic provinces appears to be dominated by panderodontaceans belonging to the subfamilies Acanthodontinae and Protopanderodontinae. In the Middle and Upper Ordovician, the balognathids yield widespread index fossils (*Amorphognathus* and *Polyplacognathus*). Chirognathaceans (*Chirognathus, Leptochirognathus, Rhipidognathus*) are important and locally very common in areas of central and western North America.

In the Silurian, several zonal indices belong to the spathognathodonts. In the Lower–Middle Devonian, species of the genus *Icriodus* are particularly important for local and intercontinental correlation. Upper Devonian conodont faunas are mostly dominated by polygnathids. Species of *Palmatolepis* are among the best index fossils known to stratigraphy. *Polygnathus* and *Ancyrodella* are also stratigraphically important. The Lower Mississippian is characterized by the polygnathid genus *Siphonodella*. The Mississippian to Pennsylvanian sequence is marked by a succession of idiognathodontids belonging to the genera *Gnathodus, Streptognathodus, Idiognathodus, Idiognathoides*, and *Cavusgnathus*.

The gondolellaceans and neospathodontids appear in the middle to upper parts of the Pennsylvanian. The superfamily Gondolellacea contains a number of important guide species for stratigraphic levels in the Permian and Triassic.

A particular stratigraphic problem depends on the chemical resistivity and hardness of the conodonts. Conodonts may remain as a residue even after the embedding sediment has been eroded or weathered away. The exposed conodonts may become parts of a new sediment. The reworked fauna indicates the age of the first, destroyed sediment, not that of the younger deposit in which it is included. When recognized, such ghost faunas are highly useful for paleogeographic reconstruction, because they may be all that is left of certain phases of marine sedimentation.

Rock types and environments

Conodonts are found only in marine sediments. In deltaic and coal swamp deposits they are restricted to marine bands, formed during temporary sea transgressions. They occur in all kinds of sedimentary rocks formed in the sea, together with fossils of all major groups of marine animals. They are most frequently encountered in thin, very slowly formed deposits. The greatest conodont faunas have been found in shales, cherts, and limestones with evidence of slow sedimentation but also in certain thin bands of sandstone. In most cases the type of sediment on the sea floor did not influence the conodonts very strongly. Only a few cases are known in which the sedimentary substrate and certain conodont species are consistently related. A particularly instructive instance from the Upper Ordovician of the Cincinnati region has been described. There two subspecies of *Rhipidognathus symmetricus* lived, one of which evident-

ly preferred extremely shallow water, whereas the other dwelt in more tranquil water farther from the shore. Other cases are known in which shallow-water deposits do not contain conodonts.

THE QUESTION OF ZOOLOGICAL AFFINITY

When conodonts were described for the first time, they were regarded as fish teeth, and this idea never has been universally abandoned. It is founded principally on superficial morphological resemblance to teeth and on chemical composition. The latter is not decisive, for there are phosphatic skeletal parts in other animal groups (*e.g.*, inarticulate brachiopods). The resemblance to teeth is fallacious. In vertebrate teeth the mineral lamellae are added from the outside inward, whereas conodonts grew outward. Vertebrate teeth are external and if broken cannot be regenerated. Conodonts, on the other hand, were internal structures that could be repaired if broken. Furthermore, the seizing, cutting, and grinding function of teeth could not be fulfilled without breakage by the more slender forms of conodonts.

The inner structure of conodonts is not diagnostic, because it consists of simple lamellae. A simple, lamellar structure, consisting of phosphate, occurs in certain primitive vertebrates. It is known as aspidin and seems to be the only vertebrate tissue with any resemblance to conodonts. Its outer form is random, however, and bears no evident relation to a constant, vital function. The fundamental plan of the main conodont assemblages, on the other hand, remained the same from the Early Ordovician to the Late Triassic. The structure of conodont white matter has not been identified in vertebrate tissues. Unlike the vertebrates, conodonts evidently never invaded the freshwater environment.

The polychaete hypothesis

Whereas no apparatus recalling the conodont assemblage is known in the vertebrates, there is a considerable resemblance between this assemblage and the jaw apparatus of the polychaetes. Of the hypotheses assuming close affinity between the conodonts and any still-existing group, the polychaete hypothesis is the most plausible. Polychaete jaws, since their first known appearance in the Ordovician, have consisted of organic material. This does not absolutely refute the affinity. Polychaete jaws are exposed in the pharynx, however, where they perform seizing and masticating functions. Conodonts seem to have been embedded in living tissue and, for the greatest part, were unsuited for a principal function as jaws. Polychaete jaws have no part resembling the ornamented platform of the conodonts.

Most asymmetrical conodonts occur as right and left elements. The conodont animals were as a rule bilaterally symmetrical. This symmetry, and the absence of a protective shell or external skeleton, indicates that they were free moving and possibly wormlike. Their comparatively great independence of sedimentary substrate, as well as the global spread of a great many species, suggests a swimming, pelagic mode of life. The porous white matter could serve to reduce weight in a swimming animal, but its disadvantage would be reduction of the strength of the elements. The elements of most conodont genera stayed separate; fusion during growth has been observed in only one instance. Thus, mobility between the elements probably was essential for the functioning of the apparatus.

The lophophore hypothesis

The conodonts grew embedded in the tissue by which they were secreted. Broken parts were crudely regenerated by lamellae laid down about the stump. The constant, fairly elaborate plan of the main types of conodonts suggests a fixed relationship to the function of surrounding tissue. It is unlikely that the outer morphology of this tissue was greatly different from that of the conodonts. This suggests a papillate or tentaculate surface. Much of conodont morphology can be understood as a means of surface enlargement, which would be of help to any organ serving food-gathering or metabolic function. Conodonts tend to add complexity during growth. For a food-gathering mechanism this is necessary. The required volume of food generally increases as the cube of the length of the animal. According to this interpretation, the conodonts supported a lophophore-like organ that strained small microplankton from the seawater and, if possible, also was able to manipulate larger, soft-bodied microplanktons. The biologically successful group of animals to which the conodonts belonged is probably distinct from all other known groups.

BIBLIOGRAPHY. C.H. PANDER, *Monographie der fossilen Fische des silurischen Systems der russisch-baltischen Gouvernements* (1856), the first monograph published on the morphology, taxonomy, and nature of conodonts; F.H.T. RHODES, "The Zoological Affinities of the Conodonts," *Biol. Rev.*, 29: 419–452 (1954), a review of known facts about conodonts; W.H. HASS, "Conodonts," in R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology*, pt. *W*, pp. 3–69 (1962), on the morphology of conodonts—a review of known genera, and literature references; M. LINDSTROM, *Conodonts* (1964), a textbook on conodonts; S.M. BERGSTROM and W.C. SWEET, "Conodonts from the Lexington Limestone (Middle Ordovician) of Kentucky and Its Lateral Equivalents in Ohio and Indiana," *Bull. Am. Paleont.*, 50:271–441 (1966), a description and classification of Ordovician conodonts according to the assemblage principle; H. PIETZNER *et al.*, "Zur chemischen Zusammensetzung und Mikromorphologie der Conodonten," *Palaeontographica*, 128A:115–152 (1968), an extensive, well-illustrated study of morphology, structure, and composition of conodonts, with a discussion of results (summary in English).

(M.L.)

Conrad II, Emperor

German king and Western emperor Conrad II, in a short reign, from 1024 to 1039, proved that the German monarchy had become a viable institution. Since the survival of the monarchy was no longer primarily dependent on a compact between sovereign and territorial nobles, it was henceforth invulnerable to prolonged rebellion on their part. Conrad also proved that at that time—within certain limits and despite the elective character of the monarchy—a self-made man could rise to kingship and establish a new Salian dynasty, which would last for almost a century.

By courtesy of the Real Biblioteca de San Lorenzo de El Escorial, Spain



Conrad II, detail from a miniature from the manuscript *Codex Aureus Echternach*, *c.* 1045; in the Real Biblioteca de San Lorenzo de El Escorial, Spain.

Conrad was born around 990, the son of Count Henry of Speyer, who had been passed over in his inheritances in favour of a younger brother. Henry was descended, through the marriage of his great-grandfather Conrad the Red to a daughter of Emperor Otto, from the Saxon house. Left poor, Conrad was brought up by the Bishop of Worms and did not receive much of a formal education; but, conscious of the deprivations suffered by him and his father, he matured early. Prudent and firm, he often displayed great chivalry as well as a strong sense of justice, and he was determined to gain the status that fortune had denied him. In 1016 he married Gisela, the widowed duchess of Swabia and a descendant of Charlemagne. Conrad, however, was distantly related to Gisela. When strict canonists took exception to the marriage, Emperor Henry II, who was jealous of the growth of Conrad's personal influence, used their findings as an excuse for forcing Conrad into temporary exile. The two men later became reconciled, and, by the time Henry II

died, in 1024, Conrad presented himself to the electoral assembly of the princes at Kamba on the Rhine as a candidate for the succession. After prolonged debates, the majority voted for him, and he was crowned king in Mainz on September 8, 1024.

Crowned king

Intelligent and genial, Conrad was also fortunate. Soon after his election, even the minority opposition was persuaded to pay their homage. Early in the following year, the sudden death of Bolesław Chrobry of Poland, a tributary to the German monarchy who had styled himself an independent king, spared Conrad the necessity of military interference. In Germany, a rebellion fomented by nobles and relatives of Conrad was joined by many lay princes of Lombardy; and, although the Italian bishops paid homage at a court in Constance in June 1025, the lay princes sought to elect William of Aquitaine as antiking. But, when the King of France refused his support, the rebellion collapsed. Early in 1026, Conrad was able to go to Milan, where Archbishop Ariberto crowned him king of Italy. After brief fighting, Conrad overcame the opposition of some towns and nobles and managed to reach Rome, where he was crowned emperor by Pope John XIX on Easter 1027. When a renewed rebellion in Germany forced him to return, he subdued the rebels and imposed severe penalties on them, not sparing members of his own family.

Conrad not only showed strength and incorruptible justice in maintaining his power but also displayed enterprise in legislation. He formally confirmed the popular legal traditions of Saxony and issued a new set of feudal constitutions for Lombardy. On Easter Sunday 1028, at an imperial court in Aachen, he had his son Henry elected and anointed king. In 1036 Henry was married to Kunigunde, the daughter of King Canute of England. Eventually, he became inseparable from his father and acted as his chief counsellor. Thus, the succession was virtually assured and the future of the new house looked bright.

In the meantime, Conrad had been compelled, after all, to campaign against Poland in 1028. After severe fighting, Mieszko—Bolesław's son and heir—was forced to make peace and surrender lands that Conrad's predecessor had lost. Even so, Conrad had to continue to campaign in the east, and in 1035 he subdued the heathen Liutitians.

Political triumphs

Although occupied intermittently in the east, Conrad was able to gain political triumphs in the west. Earlier, the childless King Rudolf of Burgundy had offered the succession to his crown to Emperor Henry II, who, however, died before Rudolf. Thus, when Rudolf died in 1032, he left his kingdom to Conrad over the opposition of the Burgundian princes, who two years later, on August 1, 1034, paid homage to Conrad at Zürich.

Although Conrad's relations with his son remained close, King Henry at times showed independent initiative. He once concluded a separate peace with King Stephen of Hungary and on another occasion gave his oath to Duke Adalbero of Carinthia never to side against him. Thus when Conrad fell out with Adalbero in 1035, Henry's oath severely strained relations between father and son. Conrad managed to overcome his son's partisanship only by humbling himself before him. In the end, Conrad's determination prevailed, and Adalbero was duly punished.

In 1036, Conrad appeared for a second time in Italy, where he proceeded with equal vigour against his old ally, Archbishop Aribert of Milan. Italy was rent by dissensions between the great princes, who, together with their vassals—the *capitanei*—had suppressed both knights and the burghers of the cities, the *valvassores*. Conrad upheld the rights of the *valvassores*, and, when Aribert, claiming to be the peer of the emperor, rejected Conrad's legislative interference, Conrad had him arrested. Aribert managed to escape, however, and succeeded in raising a rebellion in Milan. Through luck and skillful diplomacy, Conrad succeeded in isolating Aribert from his Lombard supporters as well as from his friends in Lorraine. Conrad was thus able to proceed in 1038 to southern Italy, where he installed friendly princes in Salerno and Anversa and appointed the German Richer as abbot of Montecassino.

On his return to Germany the same year along the Adriatic coast, his army succumbed to a midsummer epidemic in which both his daughter-in-law and his stepson died. Conrad himself reached Germany safely and held several important courts in Solothurn (where his son Henry was invested with the kingdom of Burgundy), in Strassburg, and in Goslar. He fell ill during the following year (1039) and died on June 4 in Utrecht. He was buried in the newly built cathedral of Speyer.

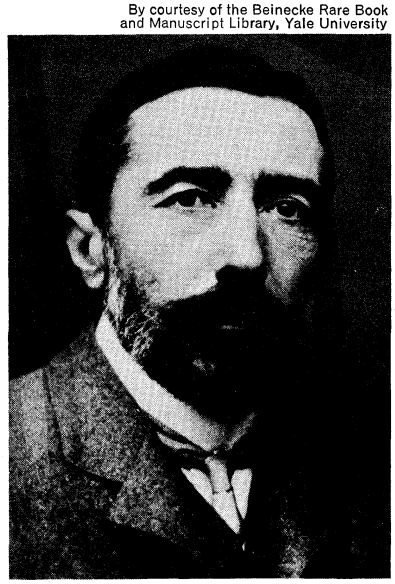
BIBLIOGRAPHY. There is no biography of Conrad II. The only useful treatment in English is G. BARRACLOUGH, *The Origins of Modern Germany*, 2nd rev. ed. ch. 3–4 (1947); in German, K. HAMPE, *Deutsche Kaisergeschichte in der Zeit der Salier und Staufer*, 11th ed., ch. 1 (1963). For important revision due to modern research, see T. SCHIEFFER, "Heinrich II und Konrad II," *Deutsches Archiv*, 8:384–437 (1951). An authoritative summary by M.L. BULST-THIELE may be found in B. GEBHARDT, *Handbuch der deutschen Geschichte*, 9th ed., vol. 1, pp. 299–307 (1970).

(P.Mu.)

Conrad, Joseph

By birth a Pole of Russian nationality, British by adoption, Joseph Conrad was during his lifetime admired for the richness of his prose and his renderings of dangerous life at sea and in exotic places. But his initial reputation, based on the public's view of him as a masterful teller of colourful adventures of the sea, masked his fascination with the individual when faced with nature's invariable unconcern, man's frequent malevolence, and his inner battles with good and evil. To Conrad, the sea, of which he held not the romantic but the sailor's workmanlike view, meant the tragedy of loneliness. A writer of complex skill and striking insight, but above all of an intensely personal vision, he has, since his death, been increasingly regarded as one of the greatest English novelists and short-story writers.

By courtesy of the Beinecke Rare Book and Manuscript Library, Yale University



Conrad, 1904.

Józef Teodor Konrad Korzeniowski was born December 3, 1857, at Berdichev, in what is now the Soviet Ukraine. On both sides he came of Polish landed gentry. His father, Apollo Nałęcz Korzeniowski, a poet and an ardent patriot, was one of the organizers of the clandestine City Committee of Warsaw that, as the National Central Committee, in 1863 directed the Polish insurrection against Russian rule. He himself took no part in the insurrection led by the Committee, for he had been arrested by the Russians in October 1861 and seven months later sent into exile at Vologda in northern Russia, where his family followed him.

Youth

The climate of Vologda was "murderous"; the four-

year-old Józef almost died on the way there. His mother was suffering from advanced tuberculosis, and though the family was allowed to move to a milder climate, she died in April 1865. Apollo, shattered by his wife's death and by the failure of insurrection, degenerated according to Conrad, "into mysticism touched with despair." Broken in health and resources, and no longer capable of original work, Apollo turned to translation, particularly of Shakespeare and Hugo. In *A Personal Record* Conrad relates that his first introduction to the English language was at the age of eight, when his father was translating *The Two Gentlemen of Verona.* In those solitary years with his father, he read Sir Walter Scott, James Fenimore Cooper, Captain Marryat, Dickens, and Thackeray in Polish and French. Apollo was also ill with tuberculosis, however, and died in Cracow in 1869.

Responsibility for the boy was assumed by his maternal uncle, Tadeusz Bobrowski, a lawyer, who for the rest of his life was indefatigable in plying his nephew with advice, admonition, financial help, and love. He sent Conrad to school at Cracow, but the boy was bored by school, and by 1872 he yearned to go to sea. A year later, his uncle sent him to Switzerland, accompanied by a tutor, whose function, Conrad believed later, "was to talk me out of my romantic folly." They argued endlessly and eventually the tutor left, calling Conrad "an incorrigible, hopeless Don Quixote." In October 1874, Conrad left Cracow for Marseilles.

French merchant service. Bobrowski made him an allowance of 2,000 francs a year and put him in touch with a merchant named Delestang, in whose ships Conrad sailed while he was in the French merchant service. His first voyage, on the "Mont-Blanc" to Martinique, was as a passenger; on her next voyage he sailed as an apprentice. In July 1876, he again sailed to the West Indies, as a steward on the "Saint-Antoine." What happened to Conrad during this voyage of 18 months is obscure, except that he seems to have taken part in some unlawful enterprise, probably gun-running, and to have sailed along the coast of Venezuela, memories of which were later to find a place in *Nostromo.* The first mate of the vessel, a Corsican named Dominic Cervoni, was the model for the hero of that novel and was to play a picturesque role in Conrad's life and work.

Voyages to West Indies

Between voyages, Conrad seems to have frequented the company of the Delestangs and made himself at home in artistic circles in Marseilles; but his sojourn in the city is shrouded in mystery. It is known, however, that his uncle visited him in February 1878, summoned by a cable announcing that Conrad was wounded and in need of money. Eighteen months later, Bobrowski wrote a friend of Conrad's father that he had found Conrad saddled with debt and recovering from a self-inflicted wound, although he told everyone else that Conrad had been wounded in a duel. What actually happened can only be surmised. In the chapter "The Tremolino," in *The Mirror of the Sea,* Conrad tells how, with three others—an Englishman, an American, and a Provençal—he bought a 60-ton vessel, the "Tremolino," and, with Dominic Cervoni as master of the ship and Cervoni's nephew César, engaged in smuggling arms into Spain on behalf of the pretender Don Carlos. César Cervoni, however, betrayed Conrad and his associates, and to escape capture they drove the "Tremolino" on the rocks.

The gun-running episode also occurs in the late novel *The Arrow of Gold,* in which Conrad (the M. George of the novel) fights a duel with the American, Blunt, for love of the beautiful young Doña Rita and is wounded in the chest. The events related in "The Tremolino" and *The Arrow of Gold* were accepted as autobiographical by Conrad's wife and by his first biographer. Conrad himself allowed his family and friends to believe that the scar on his left breast was the result of a duel. Yet, his uncle Bobrowski would scarcely have written that Conrad had attempted suicide unless he had good reason for thinking so.

Whatever had happened, certainly Conrad was heavily in debt. Moreover, as a sailor in the French Merchant Navy he was liable to conscription when he came of age. In April 1878, he signed on as a deckhand on the "Mavis," a British freighter bound for Constantinople with a cargo of coal. On the return journey, the ship, with a cargo of linseed, docked at Lowestoft, England, on the 18th of June. It was Conrad's first English landfall. Knowing no one in England and ignorant of the language, he spent three weeks in London and then joined the coaster plying between Lowestoft and Newcastle, on which he made six voyages. In the following October, he shipped as an ordinary seaman on a wool clipper, on the London-Sydney run.

Arrival in England

Career in the British merchant navy. Conrad was to serve 16 years in the British merchant navy. In June 1880, he passed his examination as second mate, sailing as a ship's officer for the first time two months later on another wool clipper on the Australia run. He was back in London the following April, and five months later joined the "Palestine," a bark of 425 tons. This move proved to be an important event in his life; it took him to the Far East for the first time, and it was also a continuously troubled voyage, which provided him with literary material that he would use later. Caught in a gale, the "Palestine" was 15 days sailing from the Thames to the Tyne in the north of England and, on leaving, was rammed by a steamer and delayed another three weeks. Then gales forced her to put into Falmouth, where there was trouble with the crew; 30 seamen either deserted or were discharged. Months later, as the "Palestine" neared Java Head, her cargo of coal caught fire and the crew had to take to the lifeboats, so that Conrad's initial landing in the East, on an island off Sumatra, took place only after a 13-and-a-half hour voyage in an open boat. In 1898 Conrad published his account of his experiences on the "Palestine," with only slight alterations, as the short story "Youth," a remarkable tale of a young officer's first command.

Voyages to the East

He returned to London by passenger steamer and soon after spent a month with his uncle at Marienbad (now Mariánské Lázné). In September 1883, he shipped as mate on the "Riversdale," leaving her at Madras to join the "Narcissus" at Bombay. This voyage gave him material for his novel *The Nigger of the "Narcissus,"* the story of an egocentric Negro sailor's deterioration and death aboard ship. The winter of 1884 he was in London studying for his first mate's certificate. The next April he signed aboard a sailing ship bound for Singapore. From there, he wrote his earliest known letters in English, to a Polish émigré friend in Cardiff to whom he revealed his dissatisfaction with the sailor's life and discussed plans for going into whaling or commerce. Soon after his return to London, two notable events occurred: he became a British subject on August 19, 1886, and, three months later, obtained his master mariner's certificate.

He did not wait for a ship of his own before going to sea again. In February 1887, he sailed as first mate on the "Highland Forest," bound for Samarang, Java. Her captain was John McWhirr, whom he later immortalized under the same name as the heroic, unimaginative captain of the steamer "Nan Shan" in *Typhoon.* During the voyage, Conrad was disabled by a spar and went into the hospital in Singapore. Discharged, he joined the "Vidar," a locally owned steamship trading among the islands of the southeast Asian archipelago. During the five or six voyages he made in four and a half months, Conrad was discovering and exploring the world he was to re-create in his first novels, *Almayer's Folly, An Outcast of the Islands, Lord Jim,* and several short stories. He met Almayer himself and got to know him "pretty well"; he also met others who appear in his works under their own names—Tom Lingard, for instance, who was a well-known trader in the region.

After leaving the "Vidar," Conrad unexpectedly obtained his first command, of the "Otago," sailing from Bangkok, an experience out of which he was to make his stories "The Shadow-Line" and "Falk." He took over the "Otago" in unpropitious circumstances. The captain Conrad replaced had died at sea, and by the time the ship reached Singapore, a voyage of 800 miles that took three weeks because of lack of wind, the whole ship's

First command

company, except Conrad and the cook, was down with fever. Conrad then discovered that his predecessor had sold almost all the ship's supply of quinine. At Singapore, he took on a new crew, went to Sydney, then sailed to Mauritius and then to Melbourne. He returned to England as a passenger on a steamship.

Experiences in Africa. Back in London in the summer of 1889, he took rooms near the Thames and, while waiting for a command, began to write *Almayer's Folly*. The task was interrupted by the strangest and probably the most important of his adventures. As a child in Poland, he had stuck his finger on the centre of the map of Africa and said, "When I grow up I shall go there." In 1889 the Congo Free State was four years old as a political entity and already notorious as a sphere of imperialistic exploitation. Conrad's childish dream took positive shape in the ambition to command a Congo River steamboat. Using what influence he could, he went to Brussels and secured an appointment. What he saw, did, and felt in the Congo are largely recorded in "Heart of Darkness," his most famous, finest, and most enigmatic story, the title of which signifies not only the heart of Africa, the dark continent, but also the heart of evil—everything that is corrupt, nihilistic, malign—and perhaps the heart of man. The story is central to Conrad's work and vision, and it is difficult not to think of his Congo experiences as traumatic. He may have exaggerated when he said that "before the Congo I was a mere animal," but in the real sense the dying Kurtz's cry, "The horror! The horror!" was Conrad's. He suffered psychological, spiritual, even metaphysical shock in the Congo, and his physical health was also damaged; for the rest of his life, he was racked by recurrent fever and gout.

Conrad was in the Congo for four months, returning to England in January 1891. He was in a hospital for some weeks and then visited a hydropathic establishment near Geneva, where he worked on *Almayer's Folly*. Back in London in the summer in a depressed state and unable to get a ship, he worked as manager of a waterside warehouse, which increased his depression. In November, he joined the "Torrens," a 1,300-ton clipper and one of the most famous passenger ships of the day, as first mate and made two voyages on her to Adelaide, Australia, and back, on the second of which he met a passenger who was to become one of his closest friends, the novelist John Galsworthy.

Illness

Back in London in July 1893, he found a letter from Bobrowski urging him to visit him. Within a month he was in the Ukraine, ill and being nursed by his uncle. On his return to London, he resumed *Almayer's Folly* and looked around for a ship. Offered the post of first mate on a ship scheduled to transport emigrants to Canada, he joined her at Rouen but, because no emigrants arrived, he returned to London in January 1894. Two weeks later, he heard that Tadeusz Bobrowski, the "wisest, firmest, most indulgent of guardians," as he was to call him, was dead.

Termination of sea life. Though he did not know it, Conrad's sea life was over. In the spring of 1894 he sent *Almayer's Folly* to the London publisher Fisher Unwin, whose reader, the critic Edward Garnett, soon to be Conrad's close friend, urged him to begin a second novel; so that, before *Almayer's Folly*, dedicated "To the memory of T.B.," was published in April 1895, he was already writing *An Outcast of the Islands*. It was as the author of *Almayer's Folly* that Conrad adopted the name by which he is known: he had learned from long experience that the name Korzeniowski was impossible on British lips.

Almayer's Folly achieved a *succès d'estime* only. The first edition consisted of 2,000 copies, and it was seven years before a third impression was called for. *An Outcast of the Islands*, published March 1896, had a comparable reception, though the anonymous reviewer of the *Saturday Review* called it "perhaps the finest piece of fiction that has been published this year, as *Almayer's Folly* was one of the finest that was published in 1895." The reviewer was H.G. Wells.

Three weeks after *An Outcast of the Islands* was published, Conrad, aged 38, married the 22-year-old Jessie George, by whom he had two sons, Borys (1898) and John Alexander (1906). Inevitably, the nature and scope of his life changed, confined to the southeast corner of England with occasional forays abroad and reduced to the ardours and miseries of authorship, which were exacerbated by poor health, near poverty, and difficulties of temperament. Besides John Galsworthy, Edward Garnett, and H.G. Wells, from his earliest days as a writer he could count among his friends the writers Henry James, Stephen Crane, R.B. Cunninghame Graham, W.H. Hudson, Ford Madox Hueffer (later Ford Madox Ford), and Arnold Bennett, but his critical reputation was out of all proportion to his readership and even after the turn of the century he was still contemplating going back to sea.

It was not, indeed, until 1910, after he had written what are now considered his finest novels—*Lord Jim*, *Nostromo*, *The Secret Agent*, and *Under Western Eyes*, the last three novels of political intrigue and romance—that his financial situation became relatively secure. He was awarded a Civil List pension of £100. The American collector John Quinn began to buy his manuscripts—for what now seem ludicrously low prices, as they did to Conrad himself when Quinn put them on the market in New York in 1923. In 1910 also the *New York Herald* asked him for a novel for serialization. He thereupon went back to *Chance*, which he had put aside unfinished five years earlier. Serialization began in January 1912, and when the novel was published in London two years later its American success was repeated. *Victory*, published in 1915, was no less successful, but it was not until his best work had been done that he became a popular or even fashionable author.

Attitude toward World War I

He was in Cracow, then in Austria, when war broke out in 1914, but was enabled to return to England by the good offices of the American ambassador to Vienna. "This war," he said, "attends my uneasy pillow like a nightmare." He was too old and unwell for active service and found writing increasingly difficult, though, in order to write articles on the war at sea, he did make some short trips on minesweepers and even a flight in a naval airplane. His relief, when the fighting ended, was offset by his pessimism about the future: he found an "awful sense of unreality" in President Wilson's idealistic postwar plans, was horrified when the Bolsheviks were invited to send delegates to the peace conference, and inflamed in his patriotism when the Red Army attacked the new republic of Poland.

In 1919 he settled at Oswalds, Bishopsbourne, near Canterbury. In 1923 he visited the United States as the guest of his publisher and enjoyed a reception in New York appropriate to his status as one of the most famous of living novelists. He crossed on ocean liners, but the inveterate sailor regarded them with contempt as "nothing but locomotives." By now he was a sick man. In March 1924, he sat for the splendid portrait bust of him by Sir Jacob Epstein, who records in his autobiography that Conrad "was crippled by rheumatism, crotchety, nervous, and ill." A month or so later he refused the offer of a knighthood from the prime minister, Ramsay MacDonald. He died on August 3, 1924, after a heart attack and was buried at St. Thomas' Roman Catholic Church, Canterbury. His second name, Teodor, is spelled incorrectly on the tombstone.

Character and reputation. Devoted to England though he was, Conrad was very much an exotic in the English scene. "He had the most perfect manners," the writer Virginia Woolf recalled on his death, "the brightest eyes, and spoke English with a strong foreign accent." His perfect manners, the evidence suggests, were the counterpart of a temperament of extreme sensibility, one irascible and quick to take offense. He was the most formal of men, which made it difficult for him to come to terms in the ordinary traffic of life with men like Shaw and Wells, for whom formality meant nothing. Perhaps the most vivid impression that remains of him, and certainly the noblest, is from a man who on the surface would seem to have nothing in common with him except aristocratic

lineage. Bertrand Russell met him in 1913, saw him infrequently, but was so impressed with him that eight years after their initial meeting, he sought permission to name his first son Conrad. Russell sums up the novelist's vision in these words: "he thought of civilised and morally tolerable life as a dangerous walk on a thin crust of barely cooled lava which at any moment might break and let the unwary sink into fiery depths," and concludes his brief memoir, "His intense and passionate nobility shines in my mind like a star seen from the bottom of a well."

Assessment

Conrad's influence on later novelists, William Faulkner in the United States, André Malraux in France, and Graham Greene in England, has been profound, not only because of his masterly technical innovations but also because of the vision of man expressed through them. He is the novelist of man in extreme situations. "Those who read me," he wrote in his preface to *A Personal Record*, "know my conviction that the world, the temporal world, rests on a few very simple ideas; so simple that they must be as old as the hills. It rests, notably, among others, on the idea of Fidelity." For Conrad; fidelity is the barrier man erects against nothingness, against corruption, against the evil that is all about him, insidious, waiting to engulf him, and that in some sense is within him unacknowledged. But what happens when fidelity is submerged, the barrier broken down, and the evil without acknowledged by the evil within? At his greatest, that is Conrad's theme.

MAJOR WORKS

NOVELS: *Almayer's Folly* (1895); *An Outcast of the Islands* (1896); *The Nigger of the "Narcissus"* (1897); *Lord Jim* (1900); with F.M. Hueffer (Ford), *The Inheritors* (1901) and *Romance* (1903); *Nostromo* (1904); *The Secret Agent* (1907); *Under Western Eyes* (1911); *Chance* (1913); *Victory* (1915); *The Shadow-Line: A Confession* (1917); *The Arrow of Gold* (1919); *The Rescue* (1920); *The Rover* (1923); with F.M. Ford, *The Nature of a Crime* (1924).

SHORT STORIES: *Tales of Unrest* (1898); *Youth: A Narrative, and Two Other Stories*, "Heart of Darkness" and "The End of the Tether" (1902); *Typhoon and Other Stories* (1903); *A Set of Six* (1908); *Twixt Land and Sea* (1912); *Within the Tides* (1915); *Tales of Hearsay* (1925).

OTHER WORKS: *The Mirror of the Sea, Memories and Impressions* (1906); *Some Reminiscences* (1912; U.S. title, *A Personal Record*, 1912); *Notes on My Books* (1921); *Last Essays* (1926).

BIBLIOGRAPHY. KENNETH A. LOHF and EUGENE P. SHEEHY, *Joseph Conrad at Mid-Century: Editions and Studies, 1895–1955* (1957); THEODORE G. EHRSAM, *A Bibliography of Joseph Conrad* (1969).

Works: JOSEPH CONRAD, *Collected Works*, 22 vol. (1923–28, reprinted 1946–); *Complete Works*, 26 vol. (1938).

Biography: JOCELYN BAINES; *Joseph Conrad: A Critical Biography* (1960), the standard life; G.J. AUBRY, *Joseph Conrad: Life and Letters*, 2 vol. (1927); J. ALLEN, *The Thunder and the Sunshine: A Biography of Joseph Conrad* (1958), *The Sea Years of Joseph Conrad* (1965); JESSIE CONRAD, *Joseph Conrad as I Knew Him* (1926), *Joseph Conrad and His Circle* (1926); F.M. FORD (HUEFFER), *Joseph Conrad: A Personal Remembrance* (1924); E. GARNETT, *Letters from Joseph Conrad, 1895–1924* (1928); G. MORF, *The Polish Heritage of Joseph Conrad* (1930); Z. NAJDER, *Conrad's Polish Background: Letters to and from Polish Friends* (Eng. trans. 1964); BERTRAND RUSSELL, *Portraits from Memory* (1956); NORMAN SHERRY, *Conrad's Eastern World* (1966).

Criticism: J.D. GORDAN, *Joseph Conrad: The Making of a Novelist* (1940); A.J. GUERARD, *Conrad the Novelist* (1958); DOUGLAS HEWITT, *Conrad: A Reassessment* (1952); F.R. KARL, *A Reader's Guide to Joseph Conrad* (1960); F.R. LEAVIS, *The Great Tradition* (1948); T.J. MOSER, *Joseph Conrad: Achievement and Decline* (1957).

(W.E.A.)

Consanguinity

Consanguinity is the sharing of some common ancestors. The word is derived from the Latin *consanguineus*, "of common blood," which implied that Roman individuals were of the same father and thus shared in the right to his inheritance. Kin are of two basic kinds: consanguineous (sharing common ancestors) and affinal (the kinsmen of one's spouse). In some societies other pairs of individuals also treat each other as relatives—for example, the wives of a pair of brothers, relatives by adoption, and persons such as godparents who have special kinlike relationships (fictive kin). Consanguineous kinship is a universal type; it includes those with common ancestors and it excludes individuals who lack ancestors in common.

In the modern sense, consanguinity is a genetic concept. From the biological point of view the term itself is unsuitable (as are the terms mixed blood and good blood) because the genetic contributions of ancestors do not flow to their descendants as blood, but through the genes contained in the chromosomes in the cell nuclei. The chromosomes are made up of nucleic acids (DNA, or deoxyribonucleic acid) and proteins. The DNA is the part of the chromosome that carries the genes and is coded in specific ways to produce and control protein synthesis; parts of each parent's coded message are transmitted to the offspring. Besides genetic determinants carried in the DNA, there are other biological influences of parents on offspring, such as the environment in the mother's womb; indeed, there is even cultural inheritance through learning, which in turn influences nutritional and other habits and hence affects growth and development. In genetics, consanguinity affects the probabilities of specific genotypes—the combinations of genetic characteristics. Consanguinity results in inheritance, from common ancestors of both parents, of their transmissible capacities to synthesize and control nucleic acids and proteins, the essential substances of all organisms.

Degrees of consanguinity

Consanguineous relatives are said to be of various degrees, according to the likelihood of their sharing genetic potentialities from the common ancestors. Thus, pairs of brothers and sisters (siblings) have all the same ancestors, whereas pairs of first cousins (cousins germane) who are not otherwise related share only one-half of their ancestors. A child inherits only about one-half of the coded information from each parent; hence a pair of brothers or sisters have about half of their chromosomal constitution in common (the doubt about the exact fraction is due to the chance element in transmission of material during meiosis, the cell division that produces sperm and egg).

Genetically the degree of consanguinity of siblings is thus the same as that between a parent and child, and both may be called consanguineous in the first degree. An aunt or uncle shares with his niece or nephew about half the chance of common inheritance of a pair of siblings, and aunts and uncles may be called consanguineous kin of the "second degree." First cousins may be called consanguineous of the "third degree."

Legal systems of designating degrees of consanguineous kin exist in Roman law, common law, and canon law. None of these systems depends on the genetic facts, however, and they have no place in a classification of the biological relationships between individuals.

A great-grandfather and great-grandson are genetically related to the same degree as a pair of first cousins. The grandparent is a linear kinsman, however, whereas the cousin is a collateral one. In genetics the degree of consanguinity is the sole factor of significance, but in social relations in various societies important considerations of collateral versus lineal types of relationship as well as age, birth order, and other factors may determine social behaviour. In fact, consanguineous kin of various degrees and even nonconsanguineous kin may be called by the same name and treated similarly by custom or law (the term uncle, for instance, may be applied to a granduncle or to the husband of an aunt).

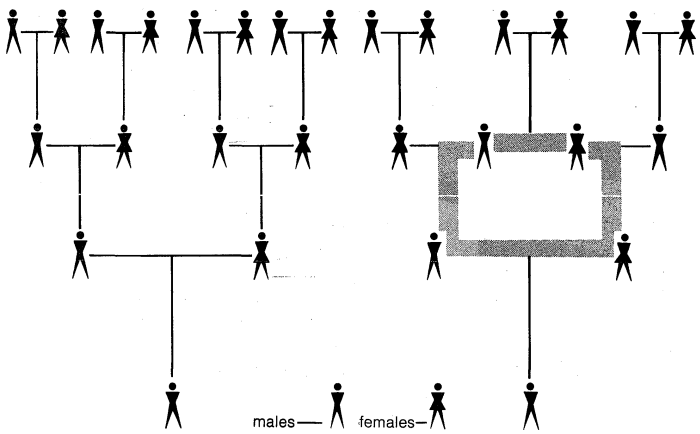
One of the two main applications of data on consanguinity is with respect to the probability that two individuals of a known degree of consanguinity with another individual will share the traits of that person. This probability depends upon the mode of inheritance and on the degree of penetrance or expressivity of genetic factors. The mode of inheritance may, for example, be dominant or recessive. A pair of genes in the same relative positions in a set of two chromosomes in the cell nucleus (these genes are called alleles) are for two alternative traits, such as greenness and yellowness in peas; both alleles

may be for one of the traits, or they may differ, so that one allele is for one trait and one for the other. The trait that appears—is "expressed"—in an individual, when the pair of alleles differ, is the dominant trait and the inheritance is dominant; if a trait does not appear unless both alleles are for that trait, it is recessive. Another mode of inheritance is sex-linked inheritance. Genes for hemophilia, for example, are present in both males and females, but the disease tends to affect only males. The degree of penetrance is the frequency with which any trait or effect is shown in a group or population that has the gene corresponding to that trait. Expressivity is the degree to which traits are shown in an individual.

Inbreeding

Another significant application of data on consanguinity is the measurement of inbreeding by the degree of consanguinity between two parents. The coefficient of inbreeding (*F*) is used. This is the probability that two alleles will be identical and derived from the same forebear. The application of this principle is most easily demonstrated by example. If brother and sister marry, the offspring would have one chance in four of inheriting a pair of identical alleles from the grandparent. With each further degree of consanguinity the likelihood is halved, so that the likelihood in the child of a mating between aunt and nephew would be one in eight, and in a child of first cousins, one in 16.

In construction of pedigrees, if horizontal lines are used to connect symbols of siblings and mates and vertical lines to connect parents with their offspring, all inbreeding is represented by one or more loops (see the Figure);



(Left) Absence of inbreeding: horizontal lines connect mates, vertical lines connect parents with their child. (Right) An inbreeding loop: the horizontal bar at the top of the loop connects brother and sister; all other lines and bars as at left.

and each loop involves consanguinity. The coefficient of inbreeding for the individual is the sum of that calculated for all the loops that include his parents. The inbreeding coefficient of a population is the average *F* of its members. The highest values of *F* are found in small populations whose members marry one another over many generations. Such groups are called isolates. Thus, the Samaritans, who remained a small but distinctive group since before the time of Christ, are considerably inbred. In the United States some religious groups also live in agricultural colonies as isolates (for instance, the Amish and Hutterites).

Effects of consanguinity. In genetics an allele that is carried at the same position in both of a pair of chromosomes is called homozygous. An allele may be rare in the general population, but, if the parent possesses it, it is transmitted from parent to child with the same probability as a common one. Therefore the chance of receiving a rare allele in the chromosomes derived from both mother and father—that is, the chance of being homozygous for that allele—is greatest in the offspring of consanguineous mating. In theory, since repeated mutations are rare, homozygosity of even common alleles may be ascribed to distant consanguinity.

Mendel's classical experiments with peas and much subsequent work show that when an allele is present in double dose (homozygous), the effects may be very different from those when it is inherited only from one parent (heterozygous). In medical genetics there are many proteins, especially enzymes, that are produced in adequate amounts if either chromosome carries the appropriate allele. Absence of the gene in both of a pair of alleles produces a deficiency in the protein it determines. The rare diseases and anomalies of this kind are relatively less rare in the offspring of consanguineous unions. In fact, in 1902, within two years after the rediscovery of Mendel's laws, the high frequency of consanguinity in the parents of individuals with three inborn errors of metabolism was used as evidence of recessive Mendelian inheritance in man. One of the defects noted was albinism, a condition in which the skin is pink and the hair white, the eyes lack pigment, and the subjects are uncomfortable in bright sunlight and usually squint. Such adverse genetic effects in the offspring of consanguineous unions is appreciable only in rare hereditary diseases; the rarer the occurrence, the more frequently the parents are found to be consanguineous. A large proportion of offspring of consanguineous mating of the first degree die or have serious defects by six months of age. In the offspring of first-cousin marriages and in other instances of similar or lesser consanguinity, few deleterious effects have been observed. Furthermore, because rarity is a factor in this effect, the overall influence of inbreeding in the general population is very limited. The modest socioeconomic differences between the kinds of people who mate with consanguineous kin and those who do not may account for some of the increase in disease and mortality that has been ascribed to consanguinity.

From the social, as opposed to the individual, point of view, close inbreeding does not increase the number of deaths from recessive disease; it merely precipitates these so-called genetic deaths in earlier generations. In heterozygous form, with no adverse influence on the individual who carries them, such alleles contain the possibility of future deaths from recessive disease; and death for infant offspring of consanguineous parents reduces that possibility for the next generation. The principle of deliberate inbreeding is used with domestic animals to eliminate such covert recessive alleles from the stock. Nevertheless, there seem to be problems of health from very highly inbred "pure" lines, and heterozygosity in some alleles seems to be advantageous. Many species, including man, have been established by episodes of isolation and inbreeding interspersed with outbreeding; they apparently thrive in this way.

Legal and social considerations. All human societies have some incest taboos. These are rules and laws that prohibit marriage or sexual relations, or both, between certain kinds of kin. The kinds of kin always include some consanguineous classes, and one theory of the establishment of incest laws is folk knowledge of undesirable inbreeding effects in offspring of such unions. Incest taboos are not uniform restrictions to a particular grade, however, and often extend to nonconsanguineous relations. Thus, in traditional Chinese society a man may marry his mother's brother's daughter, for instance, but he may not marry any person with his own surname. Other theories of the origin of incest, therefore, include analysis of its effects on stability of the family as an economic and educational unit and ascribe the definition of incest in various societies to social and psychological motives.

BIBLIOGRAPHY. The genetic aspects are well covered in C. STERN, *Principles of Human Genetics*, 2nd ed., ch. 19 (1960); the classic social study is L.H. MORGAN, "Systems of Consanguinity and Affinity of the Human Family," *Smithson. Contr. Knowl.*, vol. 17, no. 218 (1870). Perhaps the most extensive study of inbreeding is W.J. SCHULL and J.V. NEEL, *The Effects of Inbreeding on Japanese Children* (1965); some evidence of the deleterious effects of close inbreeding is contained in M.S. ADAMS and J.V. NEEL, "Children of Incest," *Pediatrics*, 40:55–62 (1967).

(G.W.L.)

Conservation Laws and Symmetry

Conservation laws in physics state that certain measurable quantities, called physical properties, do not change in the course of time within an isolated physical system. Common examples are the conservation laws of electrical charge, mass–energy, and linear and angular momentum. The existence of conservation laws is directly related to the symmetry of the laws of nature; *i.e.*, to their invariance or changelessness under various symmetry operations such as rotations, translations, and reflections of the spatial and temporal coordinates.

GENERAL CONSIDERATIONS

The most familiar law of conservation concerns matter. A clearly stated philosophical recognition that certain aspects of matter might remain unaltered through every conceivable kind of transformation was first made, according to some interpretations of his works, by the Greek philosopher Anaxagoras (*c.* 450 BC). The principle was reiterated from time to time by such natural philosophers as Sir Francis Bacon (1620), but the first scientific statement of it was made toward the end of the 18th century—on the basis of precise observations of chemical reactions—by Antoine Lavoisier, the French chemist. This law of conservation of matter may be stated as follows: the sum total of matter in the universe cannot be changed; *i.e.*, matter can neither be created nor destroyed. The next concept of conservation concerned energy, which, by the middle of the 19th century, had become mathematically definable. The law of conservation of energy merely substitutes the word energy for the word matter: the sum total of energy in the universe remains unchanged no matter what events take place; *i.e.*, energy cannot be created or destroyed. Eventually, after Einstein posited the equivalence of matter and energy, these two conservation laws had to be restated into the law of conservation of matter and energy: the sum total of matter and energy in an isolated system remains the same.

Familiar symmetry

The concept of symmetry is familiar from such figures as a circle or square in elementary geometry because they can be made to coincide with themselves under the operation of rotation. Other kinds of symmetry are revealed in the behaviour of subatomic particles.

The underlying basis for the existence of all conservation laws in macroscopic physics—*i.e.*, the physics of bulk matter—can be shown to be the geometric symmetry of space, which can be extended to include time as another dimension. In microscopic physics—*i.e.*, physical phenomena at or below the atomic level of magnitude—the same symmetry-based conservation laws apply, but here they play an even more important role; they account for the transport or propagation of such quantities as energy and electric charge. Finally, in the theory of subatomic particles, which must be treated according to quantum mechanical symmetry operations (a mathematical description of the behaviour of subatomic systems undergoing rotations, reflections, and translations, not only in real space–time but in certain abstract spaces), the application of symmetry operations leads to a classification of the particles and their interactions with each other and to a number of new conserved quantities that have no macroscopic counterparts.

This whole subject, clearly, involves modern physics, and in a brief discussion such as this it is impossible to provide the necessary mathematical background that supports concepts and physical details. Thus, in this article, the nature of conservation laws and symmetry can be treated only generally. (For more detailed and technical treatment, see ATOMIC STRUCTURE; PARTICLES, SUBATOMIC; NUCLEUS, ATOMIC; GEOMETRY, EUCLIDEAN; GEOMETRY, PROJECTIVE; ALGEBRAIC STRUCTURES; MECHANICS, CLASSICAL; MECHANICS, QUANTUM; RELATIVITY.)

Geometrical symmetry. Of Greek origin, the word symmetry at first had the meaning "balanced proportions, or beauty of form arising from such proportions," but in later usage, it took on a more precise geometrical meaning. For example, an object is said to have bilateral symmetry if there is a median plane such that the left half of the object is the exact mirror image of the right half. If any two-dimensional curve is rotated about an axis lying in its plane, the resulting three-dimensional figure is said to possess cylindrical symmetry or rotational symmetry about the axis. Other examples of this concept are rotational symmetry about a point (spherical symmetry) and the symmetries possessed by the regular polygons (equilateral triangle, square) in a plane, or the regular solids (polyhedrons) in space.

Geometrical symmetries are characterized by a set of mathematical operations that can shift a figure but leave it undistinguishable from itself in its former position, called the group of symmetry transformations. To illustrate, any regular polygon, a closed plane figure bounded by straight lines, may be considered, with its corners labelled counterclockwise A, B, C, etc. If the polygon has n sides, then in the symmetry operation, rotating the polygon counterclockwise about its centre will bring the polygon into congruence (coincidence) with itself a certain number of times. For example, if the polygon is a square, there are four sides ($n = 4$), and therefore there will be congruences when the figure is rotated ¼, ½, ¾, and 1 times 360°; *i.e.*, 0°, 90°, 180°, and 270°. In general terms, if m is any integer that is equal to 0, or 1, or 2 through $(n - 1)$, then congruence will take place by rotating the polygon (m/n) times 360°. Each rotation is thus a member, or element, of a symmetry group.

Polygon rotations

Another set of congruences of the regular polygon is obtained by considering reflection, as in a mirror. In such a case, the counterclockwise labelling A, B, C, etc., would appear to be clockwise. Thus, reflection in one of the planes of bilateral symmetry is an operation that changes the ordering of the labels A, B, C, etc., from counterclockwise to clockwise.

If, instead of the polygon, one considers a circle, any rotation about its centre will bring the circle into itself if one imagines its points to be ordered counterclockwise.

It is possible also to consider translational symmetry, by which a straight line is carried into congruence with itself (*i.e.*, slid along its length), and dilatational symmetry, which results in a consideration of similarities rather than mere congruences, such as a square within a square.

Symmetry in art and nature. Many examples of all these symmetries can be found both in art and in nature. One- and two-dimensional translation symmetry are frequently found in friezes and wallpaper patterns if these are imagined to be continued indefinitely. Columns and rosette decorations possess rotational symmetry, as do some entire buildings and monuments. Biological forms frequently possess bilateral or rotational symmetry and, in the case of such free-floating microscopic animals as the radiolarians, even have the shapes of the regular polyhedrons. Some molecules, as well as the external forms of most minerals, exhibit the symmetries of crystals; *i.e.*, combinations of rotations and reflections. These external crystalline forms arise from the symmetrical arrangements of the atoms within the smallest or unit crystal cell, which, in turn, is repeated three-dimensionally throughout the volume of the crystal in a symmetrical way to form a periodic lattice (see CRYSTALLOGRAPHY).

The arrangements of the atoms in the unit crystal cell, and its repetition in the lattice, determine its electrical, optical, acoustical, and other physical properties. In the case of magnetic materials, a given magnetic state can be reversed by reversing all currents, the effect being the equivalent of the exchange of north and south poles of the magnetic crystal. The magnetic crystal can be restored to its original state by a combination of this operation, called time-reversal operation, T, and one of the geometrical operations described above. There are 32 types of crystals, or point groups, that exhibit symmetry in the absence of a time-reversal operation and 58 more that require an additional time-reversal operation, giving a total of 90 magnetic-symmetry groups.

SYMMETRY AND THE LAWS OF PHYSICS

From the earliest days of natural philosophy (Pythagoras in the 6th century BC), symmetry has furnished insight

into the laws of physics and the nature of the cosmos. The two outstanding theoretical achievements of the 20th century, relativity and quantum theory, involve notions of symmetry in a fundamental way. Einstein's special theory of relativity can be economically stated as the invariance (that is, unchanging form) of the laws of physics under a continuous group of symmetry transformations known as the Poincaré (or inhomogeneous Lorentz) group. It is notable that in quantum theory every symmetry yields a conservation law; *i.e.*, for every symmetry operation that leaves the laws invariant, there exists a measurable physical property (*e.g.*, charge) of an isolated or undisturbed system that does not change during the development of that system in the course of time.

Derivation of conservation laws from symmetry

Some of these conservation laws can also be deduced in classical physics, in which they are closely related to the classical laws of physics. For example, the physics of Galileo and Newton states that the total linear momentum of a system (the sum of the linear momenta, mass times velocity, of each mass point) is constant, providing that the system is not acted upon by an external force. This conservation law is related to the translational symmetry of geometrical space, as can be seen intuitively by considering a two-dimensional analogy: on a huge, flat, smooth, horizontal table, a flat object is sliding without friction; until the edge is reached, nothing will act to change the momentum. If the table is curved, tipped, or rough, however, or if it contains holes, the momentum will not be conserved; for then the system will be acted on by an external force, gravity or friction. Similarly, a three-dimensional space in which the law of conservation of momentum holds contains by analogy no roughness, curvature, edges, or other local features to disturb the motion. That is, each part of empty space is assumed to be like every other part. Similarly, every direction in empty space is assumed to be equivalent to every other direction, a complete rotational symmetry, the consequence of which is the conservation of the quantity called angular momentum, a physical property of a body having magnitude and direction that is related to its mass and angular velocity. These symmetries assumed for the geometrical space are summarized by saying that the space is homogeneous (translational symmetry) and isotropic; *i.e.*, has properties with the same values when measured along axes in all directions (rotational symmetry).

In relativity, and in classical theory as well, time can be imagined as a fourth dimension, distinct from the three dimensions of ordinary space. If the laws of physics are assumed to be the same at different times, they are said to be invariant under translations in time. The consequence of this assumption can be shown to be the conservation of energy. Finally, if it is assumed that the laws of physics are unchanged and that the velocity of light is constant for observers moving with any constant relative velocity, the result is the special theory of relativity in which, for example, moving objects appear contracted in length and their masses appear to increase, and moving clocks appear, to an observer outside the moving frame of reference, to run more slowly (see RELATIVITY).

Conservation of mass and energy. Perhaps the best known result of the special theory is Einstein's relation, which states that the total energy (E) of an isolated system is equal to its mass (m) times the square of the velocity (c) of light: $E = mc^2$. The total energy includes the kinetic energy of the particle so that the mass represents the increased mass caused by the motion; *i.e.*, relativistic effect. Two separate conservation laws of classical physics, that of mass and that of energy, are thus united into a single conservation law—that of mass–energy. Particles held together by mutually attractive forces—*e.g.*, neutrons and protons bound together in the atomic nucleus—have a total mass smaller than that of their separate masses, whereas particles repelling each other have a larger mass. Indeed, the stability of the nucleus is based upon the fact that energy must be supplied to separate it into its parts. In the case of certain nuclei, such as the naturally occurring uranium-235 nuclei, which are heavier than the two parts (other nuclei) into which it can split, the nucleus is unstable and undergoes spontaneous fission with the release of a considerable amount of energy.

The subatomic particles. It will be helpful to review the forces, or interactions, that prevail between subatomic particles, which have symmetries. The most familiar of the subatomic, or elementary, particles (see PARTICLES, SUBATOMIC) are the proton and neutron, constituents of the nucleus of the atom and called nucleons, and the electron. A simple lettering system is used in discussing these: an atom of mass number A (equal to the sum of neutrons and protons) has an atomic number Z equal to the number of protons in its nucleus, each proton having a single positive electrical charge; the number of electrically neutral neutrons is, then, $A - Z$. In a normal atom, the nucleus is surrounded by electrons, Z in number, each having a single negative electrical charge. The nuclear particles (nucleons) are much heavier than the electron, the proton being 1,836.1 times as heavy and the neutron being slightly heavier than the proton. Many other subatomic particles are known and some are listed in Table 1 with their masses given in energy units as millions of electron volts (MeV) according to the relation $E = mc^2$.

Atomic mass and atomic number

Because every elementary particle has wavelike properties (see MECHANICS, QUANTUM), its size is not well defined but depends upon its mass, its interaction properties, and its state of motion. A measure of the size of one of the massless particles listed in Table 1 is furnished by its wavelength (represented by the Greek lambda, λ), the value of which can be obtained by combining Planck's relation that a quantum of energy (E) is equal to a constant (h) times the frequency (represented by the Greek nu, ν) of the wave that it represents, or $E = h\nu$, with the simple wave equation that the product of a wave's frequency and wavelength is constant, or $\lambda\nu = c$ (E being its energy, h Planck's constant of action, ν the wave frequency, and c the speed of light). This equation yields a wavelength (λ) equal to the product of Planck's constant and the velocity of light divided by the energy of the particle, namely, $\lambda = hc/E$. The same measure applied to the particles with mass of the lepton group (*e.g.*, electron and muon), gives, upon substitution of the mass–energy equivalent (mc^2) for the total energy E in the above equation, a quantity known as the Compton wavelength, $\lambda_C = h/mc$. Although the Compton wavelength is a good measure of the particle size when its velocity is large, at lower velocities the Heisenberg indeterminacy principle (which states that uncertainty in the momentum of a particle multiplied by its uncertainty in position is greater than or equal to $h/4\pi$) prevents specifying its position anywhere within a radius of one de Broglie wavelength; *i.e.*, the wavelength associated with a particle, h/mv, in which v is the speed of the particle and m is its mass. Thus in the hydrogen atom, the "size" of the electron is roughly the size of the atom itself (about 10^{-8} centimetre in diameter).

Table 1: Some Elementary Particles

particle	symbol	spin (units of $h/2\pi$)	rest mass (mc^2 in MeV)
Graviton	g	2	0
Photon	γ	1	0
Leptons			
Neutrino (electron)	ν_e	$\frac{1}{2}$	0
Neutrino (muon)	ν_μ	$\frac{1}{2}$	0
Electron	e^-	$\frac{1}{2}$	0.5111
Muon	μ^-	$\frac{1}{2}$	105.7
Hadrons			
Mesons (B = 0)			
Pion (S = 0)	π^+, π^-	0	139.6
	π^0	0	135.0
Kaons (S = 1)	K^+	0	493.8
	K^0	0	497.8
Kaons (S = −1)	$\bar{K}^0$	0	497.8
	K^-	0	493.8
Eta (S = 0)	η	0	548.8
Baryons (B = 1)			
Proton (S = 0)	p	$\frac{1}{2}$	938.3
Neutron (S = 0)	n	$\frac{1}{2}$	939.6
Lambda (S = −1)	Λ	$\frac{1}{2}$	1,115.6
3 −3 resonance (S = 0)	Δ	3/2	1,236.0

Strong interaction

In the case of the particles that respond to nuclear forces—that is, strongly interacting particles (called hadrons), including baryons and mesons (see Table 1)—their size is taken as the distance over which they can interact with other particles, about 10^{-13} centimetre. This distance, which can be measured experimentally, is called the Fermi unit. These particles have the ability to emit and reabsorb numbers of other hadrons and may be visualized by thinking of each hadron as having surrounding "clouds" of other hadrons. For example, the proton "size" is nearly the Compton wavelength of the pion, which is equal to h/mc, Planck's constant divided by the velocity of light times the pion's mass, which forms the most extended cloud.

Quantum statistics. In quantum mechanics a system of particles is described by a mathematical expression known as a wave function. When two or more particles of the system are of identical type, that wave function must remain unchanged or, at most, change sign when the coordinates of the two particles are exchanged in the function. Wave functions, which change sign through exchange, are said to have odd exchange symmetry or to be antisymmetric; if they do not change sign they have even exchange symmetry and are said to be symmetric. If two identical particles have an antisymmetric wave function they are said to obey Fermi–Dirac statistics; if their wave function is symmetric, they are said to obey Bose–Einstein statistics.

Spin properties. Another important property of an elementary particle is its intrinsic angular momentum, or the spin it has about its own axis, measured in units of Planck's constant divided by 2π, or $h/2\pi$, that can be represented by a vector (an arrow symbolizing magnitude and direction). This spin is always one of the values 0, 1/2, 1, 3/2 . . . (*i.e.*, either integer or half-integer) allowed by quantum mechanics. Identical particles with half-integer spins (leptons and baryons) have what are called Fermi–Dirac statistics and obey the Pauli exclusion principle; *i.e.*, a given quantum state in a system can be occupied by at most one of a set of identical particles. In consequence, these particles form systems having layered or shell structure (neutron and proton shells in atomic nuclei, electron shells in atoms). Particles with integer spins have a different type of statistics, called Bose–Einstein statistics, which tends to concentrate identical particles in the same state (*e.g.*, photons of light in a laser beam. The same distinction as to statistics applies to identical agglomerates of elementary particles. Total spin of an atomic nucleus depends on the spins of the constituent particles; when the total spin is integral (*e.g.*, that of the helium isotope with mass 4, helium-4) the nucleus obeys Bose–Einstein statistics, whereas when the total spin is half-integral (*e.g.*, the isotope of helium with mass 3, helium-3) it obeys Fermi–Dirac statistics.

Shell structures

Magnetic properties. An important property of electrical charges possessing angular momentum is magnetic moment; *i.e.*, magnet-like behaviour. Even electrically neutral spinning hadrons (see Table 1), like the neutron, possess magnetic moments—another indication that they have internal structure. This structure may be due to electrically (but not magnetically) compensating clouds of mesons or, possibly, to the circulation of other more elementary constituents.

Forces within the nucleus. In order to clarify the notions of range and strength of interaction mentioned above, it is necessary to deal with the forces found within the nucleus. They are the electrostatic force between charged particles (like charges repel, unlike attract), also common to macroscopic physics, and two nuclear forces (called strong and weak) found only in subatomic physics. While it is possible to begin with the electrostatic interaction of two point charges (e_1 and e_2), separated by a distance r_{12}—that is, the charges can be said to attract or repel each other with a force equal to the product of the electrostatic charges divided by the square of their distance of separation, or e_1e_2/r_{12}^2—it is more convenient to consider their potential energy (or what is called their Coulomb interaction energy), which is simply the product of the charges divided by the distance to the first power, namely, e^2/r_{12}. If the charges are in motion relative to each other, the force is much more complicated (see MAGNETISM) but is always proportional to the charges, which, in that sense, measure the strength of interaction. In the case of the strong nuclear forces, the laws analogous to those of electricity—that is, of moving charges or electrodynamics—are unknown, although they are thought to be at least as complicated as those of electricity. In the following section, interactions are stated mathematically.

Mathematical formulations of interactions. An approximation of the nuclear force between two stationary nucleons, however, can be represented as the product of the strengths of interaction (g_1 and g_2) divided by the distance of separation (r_{12}) between them, times the exponential value of minus the ratio of the separation to the range (λ) of the force; *i.e.*, the approximate distance of separation at which the nuclear force becomes relatively small: $g_1g_2/r_{12} \exp(-r_{12}/\lambda)$. This form for the potential energy, known as the Yukawa interaction, was proposed in analogy to the quantum theory of moving charges, or electrodynamics, which considers the electromagnetic interaction as being caused by the exchange of photons (a quantum of energy, or the particle form of electromagnetic radiation) between electric charges. In Yukawa's theory, these field quanta, known as mesons, analogous to the photons in quantum electrodynamics, have a mass m intermediate between that of electron and proton and the range of the force that they carry is λ, equal to their Compton wavelength divided by 2π—that is, equal to Planck's constant divided by the product of 2π times meson mass times the velocity of light, or $\lambda = h/2\pi mc$.

Yukawa potential

In this last equation, if the mass of Yukawa's meson is set in million electron volts equal to 140 divided by the velocity of light squared ($m = 140/c^2$), the observed mass of the pion (from $mc^2 = E$), a Compton wavelength for the meson equal approximately to 10^{-13} centimetre is obtained, which agrees with the observed range of nuclear forces. Also, if the strengths of interaction are set equal to a coupling constant (g)—that is, $g_1 = g_2 = g$, in which the coupling constant measures the strength of the meson interaction with the nucleon—experiment gives a dimensionless measure of the interaction between nucleons a value ($g^2/2hc$) of approximately 15.

In the electromagnetic theory, from which the meson analogy was drawn, the electric charges measure the strength of coupling of the photon to its source, the charged particle. For two electrons, equating $e_1 = e_2 = -e$, the analogous dimensionless measure of the strength of electrical forces ($e^2/2hc \approx 1/137$) of approximately 1/137 is obtained.

Comparison of forces. In comparing the above strengths of interactions, the nuclear forces are seen to be hundreds of times stronger than electrical forces. Because the mass of the photon is zero, the range of force (λ) is infinite, and examination of the Yukawa equation for nuclear interaction shows that it tends to the (Coulomb) interaction equal to the coupling constant squared divided by the separation (g^2/r_{12}) when the range tends to infinity. (A similar quantum field theory of gravitation can be formulated, the massless field quantum analogous to the photon being called the graviton. The range of force is again infinite and the source of the quanta is the particle's mass. No dimensionless measure of the strength exists, but in the physics of elementary particles, the gravitational forces between particles are about 10^{40} times weaker than the electrical forces and thus negligible.)

In addition to the pion, more than 20 heavier mesons have been identified experimentally, and it is probable that there are many others not yet observed. Each of these mesons can serve as the quantum of a field having shorter range than the pion field, so that the account given in the previous paragraph describes only the longer range part of the nuclear force. In high-energy scattering experiments, which probe the inner nucleon structure, a complicated set of forces is revealed. The mathematical description of this "shape" of the nucleon, revealed by using a beam of high-energy electrons as an electromagnetic probe, is called the nucleon form factor.

Nucleon shape

Weak interaction. Another type of nuclear force, the weak interaction, has a range so short that the corresponding field quantum would have to be at least several times as massive as the proton. This conjectured particle, called a weak boson (or W-meson), has been searched for using high-energy accelerators but so far has eluded observation. As in the case of gravitation, no dimensionless measure of the strength of weak interactions is possible, but it is roughly 10^{12} to 10^{14} times weaker than the strong nuclear forces.

To summarize, three types of force are important in the atomic and subatomic domain: electromagnetic forces, and strong and weak nuclear forces. Their ranges and strengths differ greatly. Particles that participate in strong interactions are called hadrons (mesons and baryons); the others are leptons. The photon is exceptional, being neither a hadron nor a lepton. Both baryons and leptons have weak interactions. All particles that have either electric charge or electric-charge structure (such as the neutron) have electromagnetic interactions. As will be seen below, each type of interaction has its characteristic symmetry properties, and, corresponding to the hierarchy of interactions, there is a remarkable hierarchy of symmetries such that the stronger the interaction the more symmetrical it is.

Relationship of quantum theory to conservation laws. In the elementary particles, quantum theory and the special theory of relativity play pre-eminent roles. Conservation laws and their related symmetries become most important. Some reasons for this may be briefly stated:

1. The equations of motion from classical mechanics are known; thus while the recognition of the symmetries they imply is aesthetically satisfying and often provides a powerful analytical method, the laws in themselves are already complete and nothing essentially new is added. In those parts of quantum theory in which the laws are only partially known, however, an attempt is often made to find those predictions that follow only from accepted (or assumed) symmetry principles and conservation laws.
2. Quantum theory is usually applied to the simpler systems, such as crystals, molecules, atoms, nuclei, and elementary particles, in which intrinsic symmetries are more readily observable. Also, there are more symmetries in quantum theory than in classical theory.
3. Classical physics can be considered as a special application of quantum physics when the effect of the constant of action h (Planck's constant) is negligible; that is, in the limiting case, when the effect of Planck's constant tends to zero, the separation between states of definite energy tends to zero, so that a classical state is an ensemble of many quantum mechanical states. For example, an elementary magnet in a uniform magnetic field is restricted to one of a finite set of discrete orientations in quantum theory, whereas in classical theory it may have any orientation whatsoever. When the number of possible orientations is small, powerful restrictions can be placed on the internal complexity of the system because of symmetry principles, whereas in classical theory this is not possible.

Dipole moment of neutron

As an example of this type of reasoning, there is the proof that the neutron cannot have an electric dipole moment (separation of charges to produce effective positive and negative regions), providing time-reversal is a valid symmetry. The neutron, being electrically neutral but possessing an internal structure, may be regarded as an equal mixture of positive and negative charge. If the spin is represented by a vector (an arrow pointing upward along the axis when the sense of the spin is positive and vice versa), a positive electric dipole moment would mean a net separation of the charges, with more of the positive charge near the point of the arrow. The time-reversal transformation, however, would reverse the spin and hence would exchange the point and tail of the arrow but would not change the distribution of the charge. Thus the sign of the electric dipole moment, referred to the sense of the spin, would change. If the time-reversal symmetry is a valid one for the neutron, however, it would mean that an intrinsic static property like the electric dipole moment (static because the electric dipole moment does not depend on rotation) cannot change sign and must therefore be zero.

Dynamical symmetries. Two types of symmetries can be seen to lead to conserved quantities in elementary particle physics. The first type is referred to as dynamical symmetry, associated with properties of the space–time continuum assumed by the special theory of relativity, sometimes called geometrical symmetry. This leads to the absolute conservation (constancy in time for an isolated system) of the vector linear momentum (*i.e.*, the product of mass and its velocity in a certain direction), the tensor of total angular momentum (*i.e.*, the analogous quantity for rotation), and the total energy. The behaviour of the physical laws under the operations of space and time inversion can also be studied as dynamical symmetries.

The term operator as used in theoretical physics represents a symbolic means for changing a mathematical expression that describes a physical entity or system. In quantum mechanics such operators are generally theoretical representations of measurements on the described system. Operations (or measurements) that leave the system unchanged have the special significance of providing numbers that describe the quantum mechanical state called quantum numbers (except for the trivial identity operator that leaves all states unchanged). Tests or measurements of symmetry are, of course, also represented by operators.

Space inversion. The space inversion operator, called the parity operator P, has the meaning of reflection in the origin of the space coordinates of a particle or system; that is, the three space dimensions x, y, and z become, respectively, $-x$, $-y$, and $-z$. A right-handed coordinate system is one in which turning the $+x$-axis into the $+y$-axis, the system advances in the direction of increasing z, just as a right-hand screw would advance when rotated. Under parity change, a right-handed coordinate system changes to a left-handed coordinate system and vice versa. As another example, when a particle is moving along a path and is also spinning about its direction of motion in the same direction that would advance a right-hand screw, it is said to have positive helicity; under the parity operation P it would acquire, instead, negative helicity. If the laws of physics were invariant under parity, the laws and all their predictions could be expressed mathematically in such a way that it would be impossible to tell whether a right-handed or left-handed coordinate system had been employed.

Helicity

Time inversion. The time inversion, or time reversal operator T, means the reversal of motion. Thus under the time reversal operation T, a particle of positive helicity remains of positive helicity, applying the right-hand rule above, because both the sense of spin and the direction of motion are reversed by the operation. Time-reversal invariance (or symmetry) holds if, whenever a motion is allowed by the laws of physics, the reversed motion is also an allowed one. Thus if a motion picture were taken of the motion of a body and viewed when run backward, it would be possible to say that the observed motion is an allowed motion.

The laws of classical physics are, indeed, invariant under the space-inversion and time-reversal operations P and T, but in the classical case this condition does not lead to any conservation laws. For an example of this invariancy under the parity operator, the components of a vector representing electric field change direction (as x, y, and z directions cited earlier), but the components of the magnetic field vector do not; the currents that are the source of the magnetic field and those that are acted upon by magnetic fields, however, also reverse their sense of direction, with the result that there is no observable effect of changing from a right-handed to a left-handed description of magnetic phenomena.

The invariance of physical phenomena under time reversal may seem surprising at first. If, for example, an egg is cracked into a bowl and beaten, a motion picture of these actions shown in reverse would strain one's credulity. It is necessary to go further and to imagine the situation if all the final microscopic motions of the egg

episode were reversed. Theoretically, the egg would reform itself; however, this is an operation impossible to achieve in practice, and this is the reason that time appears to be unidirectional.

Violation of conservation law. Until 1956 the laws of quantum physics were generally assumed to be invariant under parity and time reversal. In that year, however, in order to explain an apparent lack of conservation of the parity quantum number in the decay of *K*-mesons (kaons) (see Table 1) into two or three π-mesons (pions), two Chinese physicists in the United States, Tsung-Dao Lee and Chen Ning Yang, pointed out that the conservation of parity had not been experimentally tested in the weak nuclear interactions previously observed (*e.g.*, nuclear beta minus decay, symbolized β^-, the emission of negative electrons from a nucleus) and proposed several new experiments. When these experiments were carried out during 1957, the nonconservation of parity was decisively verified. In 1964 when the decay of kaons into pions was investigated by others, another new and surprising result was obtained and was interpreted as being due to a violation of the conservation law associated with time reversal in a new superweak nuclear interaction.

Charge conjugation. The inversion symmetries of space and time are associated in relativistic quantum field theory with another symmetry, that of charge conjugation (*C*). In theories that conserve parity and time reversal, such as quantum electrodynamics or the strong nuclear interaction, invariance under charge conjugation, which changes the sign of the electric charge, implies that every charged elementary particle has an oppositely charged partner, its antiparticle. The antiparticle of an electrically neutral particle may be identical to the particle, as in the case of the neutral pion, or it may be distinct, as in the case of the antineutron. All of the expected antiparticles have been observed, beginning in 1933 with the discovery of the antielectron, or positron. Weak nuclear interactions, however, do not conserve charge conjugation, a phenomenon that poses a paradox for the theoretical prediction of the existence of antiparticles.

Antiparticles

Time reversal/charge conjugation/parity. The difficulty may be resolved. Physicists have shown that invariance under the combined symmetry time reversal/charge conjugation/parity (*TCP*) is sufficient to demand the existence of antiparticles. The meaning of a combined symmetry may be illustrated in the case of the weak (but not superweak) nuclear interaction that conserves *CP* but not *C* and *P* separately. In the above discussion of parity, reference was made to the notions of positive and negative helicity. Neutrinos, which enter only into the weak interactions, are found to have negative helicity, whereas antineutrinos have positive helicity. Thus invariance under charge conjugation alone, which would imply the existence of neutrinos and antineutrinos of the same helicity, is violated. The same conclusion would follow if *P*-invariance were to hold because parity changes the sign of the helicity without changing neutrino to antineutrino. Under the combined inversion *CP*, however, a neutrino of negative helicity becomes an antineutrino of positive helicity, in accord with observation. Similarly, the existence of antiparticles is predicted even if charge conjugation, parity, and time reversal are violated separately, providing *TCP* is still a valid symmetry.

Internal symmetries. The dynamical symmetries considered so far are related to the properties of space and time, and the continuous symmetry transformations such as translation in space, time, and velocity (or, alternatively, Lorentz transformation) all lead to such absolutely conserved quantities as energy and momentum. In the case of the discrete transformations, charge conjugation, parity, and time reversal, however, it has been seen that certain interactions—the weak nuclear interactions—would not conserve these symmetries separately but would do so in certain combinations. A number of other symmetries in elementary particle physics, some absolute and some approximate, are extremely useful for classifying the particles and lead to selection rules; *i.e.*, rules that tell which reactions among the particles are allowed and which are forbidden. The information in the following sections requires a broader knowledge of high energy physics that can be found under: ELECTRON; FIELDS, THEORY OF (PHYSICS); MATTER AND ANTIMATTER; NUCLEUS, ATOMIC; and PARTICLES, SUBATOMIC.

Absolutely conserved charges. There are four absolutely conserved charges: electric charge Q, baryon number B, and two lepton numbers l_e and l_μ, associated respectively with electrons and muons. In the sense that all known electric charges are integral multiples of the fundamental charge $e = 1.60 \times 10^{-19}$ coulomb, the electric charge Q may be regarded as a quantum number that, like B, l_e and l_μ, takes on only integer values. In other respects, however, the electric charge is distinguished from the other absolutely conserved quantities, being a property of both leptons and hadrons and being the source of the electromagnetic field. In any elementary particle reaction, if a and b represent reactants and c, d, etc., the products, an equation can be written: $a + b \rightarrow c + d + \ldots$; in this equation, the sum of the charges on the left side of the arrow must equal the sum of the charges of the particles on the right. (Here, and below, the arrow indicates decay or transformation of particles on its left into particles on its right.)

Baryon number

Baryon number B is a property only of the hadrons; those in which B equals zero are called mesons, those in which B equals one are called baryons, whereas antibaryons have B equal to minus one. Protons and neutrons are baryons, and hence the mass number A of the atomic nucleus is identical with its baryon number. It is the absolute additive conservation of baryon number and of energy that makes the free proton absolutely stable because it is lightest of the baryons. The neutron, being slightly heavier than the proton, can undergo the weak nuclear interaction in which the neutron (n) decays into a proton (p), electron (e^-), and antineutrino ($\bar{\nu}$) according to the equation: $n \rightarrow p + e^- + \bar{\nu}_e$. This process is called beta decay. If one did not take into account the energy and momentum carried by the massless neutral particle $\bar{\nu}_e$ (the antineutrino of the electron), the other three quantities would appear to be not conserved. It was on this basis that the neutrino's existence was predicted by Wolfgang Pauli, a German physicist, in 1931. It is sometimes (energetically) possible for protons bound in nuclei to undergo the following decay reaction in which e^+ is the positron (antiparticle of the electron) and ν_e is the neutrino (the massless particle) associated with the production of the positron, $p \rightarrow n + e^+ + \nu_e$.

Additive conservation of the lepton (electrons, neutrinos, and muons; see below) number l_e is also illustrated by the decay processes of the neutron and proton mentioned in the previous paragraph. The leptons (ν_e and e^-) are assigned the lepton number $l_e = +1$, and the antileptons ($\bar{\nu}_e$ and e^+) are assigned the lepton number $l_e = -1$. On both sides of each arrow above (in time: before and after the reaction takes place), the total electron lepton number l_e (the sum of the lepton numbers for each particle) must be the same. In addition to the electron, there is another charged lepton, the muon, which is much heavier than the electron and which occurs typically as a product in the decay of pions or kaons (see Table 1). Free charged pions, for example, undergo the decay processes $\pi^+ \rightarrow \mu^+ + \nu_\mu$, and $\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$. Again, the particles μ^- and ν_μ are assigned (conventionally) the lepton number $l_\mu = +1$, and their antiparticles μ^+ and $\bar{\nu}_\mu$ are assigned $l_\mu = -1$. That the neutrinos associated with electron and muon are different, implying the existence of two different lepton numbers l_e and l_μ, was established in the early 1960s. One consequence is that whereas the muon decays (in about 10^{-6} second) by the process $\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$, processes such as $\mu^- \rightarrow e^- + e^- + e^+$ are forbidden and do not occur.

Strangeness and hypercharge. Other internal symmetry properties such as strangeness, treated below, belong to hadrons alone (see Table 1). From 1947 on, a number of hadrons, both mesons and baryons, have been observed with the following property: although produced copiously, in strong interactions such as proton–proton collisions, they are unable to decay by strong or electro-

magnetic interactions even when that is allowed by all the conservation laws considered so far in this article. As a consequence, these particles are relatively stable, with lifetimes of 10^{-10} to 10^{-8} second, characteristic of the weak nuclear interaction. The apparent contradiction between their behaviour in production and decay is resolved by assigning a new quantum number, the strangeness number S, and requiring that it be conserved by strong and electromagnetic interactions. In production, then, two or more particles of nonzero S can be produced, providing their total strangeness adds up to zero. The conservation of strangeness does not hold for the weak interactions by which the particles decay, but it can change by one unit. The strange particles have antiparticles of opposite strangeness. Examples of reactions permitted by the strangeness selection rule are a pion and proton producing a kaon: $\pi^- + p \rightarrow + K^+ + \Lambda$ (associated production), and an antiproton and proton producing lambda plus antilambda particles: $\bar{p} + p \rightarrow \bar{\Lambda} + \Lambda$; as the π^- (pion), and $\bar{p}$ (antiproton) have strangeness numbers $S = 0$, Λ (lambda) has $S = -1$, and K^+ (kaon) and the antilambda $\bar{\Lambda}$ have $S = +1$ (see Table 1).

A quantity closely related to the internal symmetry property strangeness, which is useful in the symmetry classification of hadrons, is the hypercharge Y, defined as equal to the sum of the baryon and strangeness numbers: $Y = B + S$. Because the baryon number is absolutely conserved, the change of strangeness in any reaction is always the same as the change of its hypercharge.

Isospin symmetry. A typical characteristic of the hadrons is that they occur in families of the same spin, intrinsic parity, baryon number, and of closely similar mass. Such groupings of particles, differing in electric charge but not in hypercharge—such as proton and neutron, or such as the three pions—are called charge multiplets, or isospin multiplets. It is believed that the small mass differences that can exist between differently charged members of a charge multiplet would vanish if only the strong nuclear interactions were present and that, except for small electromagnetic and weak interaction corrections, their strong interactions would be identical. In atomic nuclei, for example, the forces between a pair of protons, a pair of neutrons, or a neutron and a proton appear to be the same except for corrections that can be reasonably ascribed to their differing electromagnetic properties. This concept is called the charge independence of nuclear forces.

In this situation, the proton and neutron are regarded as different charge states (of charge $Q = 1$ and $Q = 0$), respectively of a single particle, the nucleon. The pions π^+, π^0, and π^- are regarded as charge states of a single particle, the pion, having charges $Q = 1$, $Q = 0$, and $Q = -1$. In visualizing the symmetry corresponding to the charge independence of the strong nuclear interaction, it is useful to represent the charge states of the nucleons as vectors in an abstract, two-dimensional complex space (one spatial coordinate is complex; *i.e.*, its magnitude is multiplied by the imaginary quantity $\sqrt{-1}$, called the charge space, or isospin space). Charge independence, or isospin invariance, is then the statement that the strong nuclear interaction is invariant under arbitrary rigid rotations of the axes in this abstract space. The three operators that generate these rotations have the same algebra as the three operators that generate the rotations in real space (angular momentum operators), hence the name isospin. One of them, I_3, is related to charge and hypercharge by the relation: $Q = I_3 + \frac{1}{2} Y$. The largest value of I_3 within a multiplet is called I, and there are $2I + 1$ charge states in the multiplet. The continuous symmetry transformations (called the Lie group, after Marius Sophus Lie, a Norwegian mathematician), which lead to the conservation of angular momentum in the real space and of isospin in the abstract space, belong to the special unitary group in two dimensions, $SU(2)$. As stated, the generator I_3 of this group is related to the value of the charge, whereas I_1 and I_2 are related to the transformation from one charge state to another.

Isospin space

SU(3) and higher symmetries. By considering a three-dimensional abstract complex vector space and the group of rotations in this space, $SU(3)$, larger family groupings can be accommodated, which may be called charge–hypercharge multiplets. In this larger space, there are eight generators of rotations: the three isospin operators I_1, I_2, I_3, the hypercharge Y, and four others related to the change of strangeness. Although isospin multiplets may contain any number of members, $SU(3)$ multiplets are limited to sets, representations of the group, containing prescribed numbers of members. According to the presently accepted $SU(3)$ scheme (proposed in 1962 by a U.S. physicist, Murray Gell-Mann, and an Israeli physicist, Yuval Ne'eman), known as the "eightfold way," the simplest nontrivial representation employed by the hadrons has eight members. For example, a certain meson octet contains three pions, four kaons, and an eighth, a neutral meson, the eta. The baryon octet contains one proton, one neutron, one lambda, and five other strange baryons. About a dozen or more such representations have been identified.

In contrast to the well-established charge independence of the strong nuclear force, $SU(3)$ invariance is not an excellent symmetry, for it appears to be violated not only by the weak and electromagnetic interaction but by a part of the strong interaction force itself. The strong symmetry-breaking interaction is not well understood, but its presence is indicated by the fact that the average mass difference between isospin multiplets that make up an $SU(3)$ representation is far larger than the electromagnetic mass differences within the isospin multiplets. The eta mass, for example, is four times that of the pion.

A great deal of theoretical research has been concerned with the attempt to understand the breaking of $SU(3)$ symmetry and to explore still larger symmetry groups such as $SU(6)$, with the object of relating the properties of larger sets of particles to each other and to understanding the dynamics of the strong interactions. Particular emphasis has been placed upon the algebra of the conserved and partially conserved currents, related to the generators of the symmetry groups just as the electric current is related to the electric charge.

Quark models. As the properties of atomic nuclei are related to the constituent protons and neutrons, so the properties of hadrons would find a natural explanation if they could be regarded as composites of more elementary constituents. $SU(3)$ symmetry would require a triplet of half-spin constituents, called quarks. The quark triplets, q_1, q_2, and q_3, with their quantum numbers, are given in Table 2. There are also postulated three anti-

Hadrons composed of quarks

Table 2: Quantum Numbers for Quarks

	I	I_3	S	B	Y	Q
q_1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
q_2	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$
q_3	0	0	-1	$\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{1}{3}$

quarks, the quantum numbers of which are the negatives (except for I) of those of the quarks. The quarks q_1 and q_2 form an isospin doublet (like proton and neutron) and q_3 is an isospin singlet (like lambda). In the strong nuclear interaction, it is assumed that each variety of quark is separately conserved in the same sense as baryon number and lepton number (*i.e.*, number of quarks minus number of antiquarks of a given variety). Conservation of strangeness is explained as the conservation of the strange quark q_3, whereas weak interactions occur by a transformation of quark type.

Because each quark is assigned baryon number $B = 1/3$, baryons ($B = 1$) are made of three quarks, and mesons ($B = 0$) are made of one quark and one antiquark. All hadrons so far observed can be constructed from the three quarks and their antiquarks and their masses, mass differences, and interaction properties can be qualitatively understood in terms of their quark structures. The values of electric charge Q given in Table 2 are derived only from theoretical relationships. Intensive searches for particles of nonintegral electric charge

have failed so far to provide conclusive evidence for their existence, which suggests that the quarks may be extremely heavy particles, if indeed they exist.

BIBLIOGRAPHY. HERMANN WEYL, *Symmetry* (1952), a beautifully illustrated book on the mathematical, aesthetic, physical, and biological aspects of symmetry; EUGENE P. WIGNER, *Symmetries and Reflections* (1967), a collection of essays by a Nobel Prize winner, the first five of which discuss conservation laws and symmetry: CHEN NING YANG, *Elementary Particles* (1962), an elegant little book by another Nobelist on the symmetries of elementary particles; STEPHEN GASIOROWICZ, *Elementary Particle Physics* (1966), a well-written textbook on elementary particles, at graduate student level: HARRY J. LIPKIN, *Lie Groups for Pedestrians* (1965), a "simple" approach to *SU* (3) symmetry for those familiar with angular momentum theory, at first-year graduate level; OLIVER E. OVERSETH, "Experiments in Time Reversal," *Scient. Am.*, 221:88–94 (1969); Victor F. Weisskopf, "The Three Spectroscopies," *Scient. Am.*, 218:15–29 (1968), two articles for the intelligent layman; LAURIE M. BROWN, "Quarkways to Particle Symmetry," *Physics To-day*, 19:44–47 (1966).

Additional references at a more elementary level include: EUGENE P. WIGNER, "Violations of Symmetry in Physics," *Scient. Am.*, 213:28–36 (1965); R.P. FEYNMAN, *The Character of Physical Law* (1965); M. GARDNER, *The Ambidextrous Universe* (1964); and T. GOLD, "The Arrow of Time," *Am. J. Phys.*, 30:403–410 (1962).

(L.M.B.)

Conservation of Natural Resources

Although the history of conservation is as old as mankind, the use of the word in the contemporary context is relatively recent. Over the years conservation has acquired many connotations: to some people it has meant the protection of wild nature, apart from man; to others the sustained production of useful materials from the living resources of the Earth—*e.g.*, fisheries, wildlife, forests, agricultural croplands—along with the wise use of fuels and minerals. But by the 1970s achievement of the highest sustainable quality of living for mankind by the rational utilization of the environment had become the definition of conservation that was most widely accepted by such organizations as the International Union for the Conservation of Nature and Natural Resources. The emphasis was on man; by stressing the quality of living rather than the sustained production of commodities, however, the scope of the definition was broadened so that conservation now includes, in addition to the protection of wild nature to enrich the life of man, the control or elimination of environmental pollution in its many manifestations.

Because conservation is a social attitude as well as a social movement, it is necessarily partisan. It advocates practices that will perpetuate the resources of the Earth on which man depends or in whose continued existence he takes an interest and is opposed to the view that resources may always be used in the short run for personal profit or for the immediate benefit of living generations. Conservation derives its tenets from a knowledge of ecology, the science concerned with interrelationships between living things and their environment. Because ecology encompasses a body of scientific knowledge as well as the associated attitudes and techniques for acquiring further knowledge, it is one of the many disciplines employed by the conservation movement in its efforts to protect the environment.

This discussion of the many scientific and technological aspects of conservation, as well as its various social, economic, and political ramifications, is divided into the major sections indicated below. Other articles of interest include ECOSYSTEM; TERRESTRIAL ECOSYSTEM; AQUATIC ECOSYSTEM.

This article is divided into the following sections:

I. Concepts important to conservation
- Man's need for natural resources
  - Primary and secondary needs
  - Rational use of resources
  - Multiple use and restoration
- The importance of conservation
  - Values to mankind
  - Conflicting attitudes and issues

II. The history of conservation
- Early beginnings
  - Ancient conservation practices
  - Conservation during and following the Renaissance
  - Rise of the modern conservation movement
- Recent history
  - Spread of modern conservation practices
  - New conservation problems and approaches

III. Types of natural resources
- Renewable resources
  - Plants and animals
  - Ecosystems
  - Solar energy, air, and water
- Nonrenewable resources
  - Fossil fuels
  - Nuclear fuels
  - Minerals

IV. Management of natural resources
- Managing nonliving resources
  - Soils
  - Water
  - Air
  - Subsurface deposits
- Managing living resources
  - Natural communities
  - Wildlife and fisheries management
  - Multiple-use management
  - Intensive wild-land uses
  - Agricultural management
- International problems of resource management
  - Effective international agreements
  - Territorial limits and marine resources
  - International trade in animals and plants
- The pollution of natural resources
  - Water pollution
  - Land pollution
  - Noise pollution
  - Chemical pollutants
- The future of conservation
  - The role of population, industry, and technology
  - Problems in need of solutions
  - Extraterrestrial conservation

I. Concepts important to conservation

MAN'S NEED FOR NATURAL RESOURCES

Like "conservation" itself, the term natural resources has also undergone an expansion in meaning as a result of man's greater understanding of his relationship with the world he inhabits. Early in the 20th century natural resources were viewed primarily as sources of useful commodities. They were the raw materials in the environment that were used or capable of being used by man for some purpose: minerals and fuels, forest and grazing resources, wildlife, fisheries, and the like. In a restricted sense, the term is still used in this way. More recently, however, the concept of natural resources has been broadened to include the total natural environment—the entire surface layer of the planet—because all parts of the Earth's surface are of use and of value to man in that they contribute to the production of the necessities and amenities that people require or demand. Thus, when considered in this respect, the atmosphere, oceans, deserts, and polar regions have all become valuable resources that must be managed with care to provide for the future of man.

Primary and secondary needs. Certain primary needs or natural-resource requirements are necessary for man's existence. These include energy in the form of organic foods that are digestible, are capable of being assimilated, and contain adequate amounts of proteins, fats, carbohydrates, vitamins, and minerals; water with a relatively low content of dissolved salts and free from toxic or injurious substances; air that contains an adequate quantity of oxygen but no harmful materials; and, in most climates, an external source of energy for heating purposes as well as various materials from which clothing and shelter can be fashioned to provide warmth in cold weather and coolness during excessively hot weather. These basic human needs were supplied to primitive man by wild vegetation and animal life, by springs and streams, and by the atmosphere and the Sun. Later, the discovery of fire made possible the heating of shelters and also made available a greater choice of foods—those that

could be rendered palatable or more palatable by being heated.

The primary needs of man are the same today; however, with increased populations and depleted supplies of wild resources, a secondary category of needs has developed. These include those materials or energy sources needed to provide greater quantities of the primary materials from the spaces available for their production (*e.g.*, fertilizers to increase crop yields). In addition to such needs, man also has a wide range of natural-resource "wants." These include the materials, experiences, or space needed to make existence more enjoyable.

Natural-resource "wants"

Agricultural and urban development. The development of agriculture enabled man to produce greater amounts of food on a more reliable basis and from smaller areas of land than previously had been possible. He lost his direct dependence on the availability of wild food materials. Greater supplies of food made it possible to provide for greater numbers of people in the agricultural regions, and these numbers soon exceeded the capacity of the original or the still-existing natural environment to supply their primary needs. Thus, the first secondary needs developed—farming tools and, later, domestic animals to use the tools more effectively and, for the latter, the food supplies necessary to keep them alive. In time, to keep the agricultural soils productive, the need for fertilizers of various kinds developed. Increasing dependence upon foods that could not be eaten raw generated the need for materials from which cooking and eating utensils could be fashioned. Thus, requirements for a wide variety of nonliving natural resources developed along with the rise of agricultural lands and settled villages.

With the growth of civilization and the concentration of people into cities, natural-resource requirements increased as secondary needs expanded. It became essential to organize and direct agriculture over vast areas in order to provide for large numbers of urban inhabitants. Effective transportation from farmlands to cities became essential, as did metals and all kinds of other minerals, stones, and timber suitable for the construction of buildings, ships, and vehicles; in addition, greater numbers of domestic animals (cows, hogs, sheep, etc.) were required. Human wants were further increased as the greater leisure of civilized life enabled part of the population to look beyond the problem of mere survival. Thus, a desire for contact with wild nature developed as urban man became increasingly separated from it through urban existence.

Industrial and technological growth. The greatest expansion of human requirements for natural resources followed the Industrial Revolution during the latter half of the 18th and first half of the 19th centuries and the scientific and technological revolutions that succeeded it in the 20th. Resources that were of no value only a relatively few years ago are now used—beryllium for rockets and uranium for nuclear fuel. Although neither of these yet qualifies as a secondary need, uranium is rapidly moving into that category. Man now consumes enormous quantities of coal, natural gas, and petroleum—resources that were scarcely used only a century ago. The quantities of food now demanded require an enormous input into the agricultural economy of chemical materials and sophisticated farming implements as well as an input of fuel energy that in some areas exceeds the energy value of the food raised in these areas. Moreover, because both the demand for luxuries and the degree of wastefulness are excessive, not all consumption of resources is related to the supplying of needs.

Excessive resource demands and waste

In the confusion of needs, wants, and waste related to the current use of natural resources, it is notable that the primary requirements of man are still the same as they were in primitive times. It would be possible for mankind to survive, although in greatly reduced numbers, by using only wild vegetation and animal life to meet his resource needs. This is because the living resources of the Earth contain all the basic requirements for human survival. The need for other resources is the result of man's desire to live in greater numbers and at a standard of living considerably higher than that enjoyed by his forebears. By reducing population growth in the future, it would be possible to enjoy a highly developed technology, a high material standard of living, and a wide range of wants and luxuries while still placing little strain upon the Earth's available resources. But, with the steadily growing human population, with an expanding technology that becomes increasingly more demanding, and with the growing demands for material goods, the pressure on the Earth's natural resources increases daily. Whether or not available quantities of these resources are sufficient to meet man's growing wants and needs is uncertain.

Rational use of resources. In its present usage, the conservation of natural resources includes a wide range of subsidiary concepts. One such concept is that of the rational use of the environment, which includes the preservation of certain areas or resources in an essentially undisturbed condition because they either are of scientific interest, have aesthetic appeal, or have recreational value. Preservation also serves an ecological purpose by maintaining the function of the total environment, such as the protection of forests to assure a sustained yield of water into urban reservoirs or the protection of estuaries in order to perpetuate an ocean fishery. But the preservation or the protection of natural resources is not the only concern of conservation; rational use also implies the direct use of resources for their commodity or recreational values. Thus, the harvesting of forest crops, the grazing of grasslands by livestock, the catching of fish, and the hunting of wild animals can be considered a legitimate part of the rational use of natural resources when they are carried out in such a way that the resource is perpetuated and not endangered. Such activities involve another concept, that of sustained yield, the understanding, in other words, that hunting or fishing should take only the annual surplus of individuals so as not to endanger the breeding stock of game animals or fishes and, similarly, that the cutting of trees or the grazing of grasses should remove either their annual increment or that portion realistically capable of being replaced over a period of years through the operation of natural processes with man's assistance when needed.

Sustained yield

Multiple use and restoration. Also important to the concept of conservation is the recognition that natural resources have multiple values. In addition to its value as livestock forage, grass, for example, also supports wild animal life, holds soil in place, maintains the productivity of soil, keeps soil and water relationships in proper balance, and helps guarantee streamflow or yields of water to underground channels. Grasslands, moreover, have aesthetic, recreational, and scientific values. All of the many values of grassland must be considered before a decision is made to use a grassland for a particular purpose. Ideally, an area of land can serve many purposes simultaneously or sequentially; *i.e.*, can have multiple uses.

Another of the more hopeful aspects of conservation is the concept of restoration. A forest that has been cut or burned can, with care, regenerate itself. Areas that have been mined and left barren often can be revegetated with reasonable expenditures of money and effort. Depleted animal or plant populations will recover their original abundance if suitably protected in an adequate habitat. The restoration of natural vegetation depends upon the ecological process of succession, in which plants with varying degrees of tolerance to extreme conditions in an environment will invade a disturbed or barren area and replace one another until a stable, self-perpetuating community is achieved. Restoration is possible, however, only as long as species are protected and the genetic diversity of life is maintained. When species become extinct the restoration of past conditions becomes impossible.

THE IMPORTANCE OF CONSERVATION

Values to mankind. Conservation is essential to the survival of man. Because life depends upon the proper functioning of the biosphere—the relatively narrow zone of air, water, soil, and rock in which all life on Earth

The role of conservation in maintaining the biosphere

exists—the ultimate purpose of conservation is to maintain the biosphere in a healthy operating condition. Although it is known that green plants supply oxygen to the atmosphere, that plants and animals recycle nutrients, and that plants and animals help maintain the fertility of soils, many of the elements that contribute to the proper functioning of the biosphere have not yet been identified. Because mankind lives with such environmental uncertainties, an attitude of care and protection toward the Earth's living resources has come to seem necessary.

Certain aspects of conservation, however, such as the prevention of pollution, have a more narrow and immediate importance to man's welfare. There are numerous illustrations of the serious impact of pollutants in air, water, or soil on human health and survival; for example, the accumulation of sulfur dioxide in the air of London during the 1950s led to many deaths that probably would not have otherwise occurred. The dumping of mercury-containing wastes in waters around Japan caused the death of many people and destroyed the health of others, and continuing accumulations of such toxic metals as lead, cadmium, and arsenic in air and water threaten widespread damage to human health.

Economic value. Unless viewed in terms of human survival, the economic value of conservation is sometimes difficult to demonstrate. Although the floating plants of the ocean, the microscopic phytoplankton, are of little direct economic value to man, for example, their elimination from the food chain of animals would quickly destroy the world's marine fisheries, which are a major source of human food; in time, even the world's oxygen supply would be severely depleted. Similarly, the economic value of pollution control as a factor in maintaining good health can be measured only indirectly in such crude terms as labour lost and cost of medical and hospital care.

As explained below, much of the apparent conflict in the economics of conservation results from the difference between short-term or individual interests and the long-term interests of groups or of all mankind. The long-term economic community benefits to be derived from stable and productive farmlands and forests are considerable when compared with farms that are exploited, eroded, and abandoned or with forests that are cut, burned, and allowed to deteriorate. Short-term economic considerations, however, may lead individuals or communities to exploit their farms and forests for maximum profit at minimum cost and then move on, leaving the deteriorated lands behind.

The appreciation of wild nature

Aesthetic and recreational value. Appreciation of wild nature as a source of aesthetic pleasure and the use of wild lands and wild-animal resources for recreational enjoyment have long been recognized as among the more important values of conservation. Outdoor-based recreational activities, such as fishing, hunting, boating, swimming, picnicking, sunbathing, hiking, and skiing, are related to the continued existence of natural or near-natural environments as the sites for these activities. Although it is almost impossible to evaluate aesthetic and recreational values in terms of their psychological or sociological importance, because they may vary from one culture to another, evidence indicates that, as personal affluence and the freedom from the sheer struggle for survival increase, the demand for outdoor recreation and outdoor space also increases. Even in the absence of government activity, it has become financially attractive for private investors to provide facilities and opportunities that exploit outdoor recreational resources.

Scientific value. Conservation is also of great scientific value. Because relatively little is known about the past, present, and possible future of the biosphere, natural outdoor laboratories, including areas of undisturbed nature, must be maintained in order to conduct the studies needed to acquire knowledge. Moreover, there are many natural resources with undiscovered scientific and technological values. If, for example, all apparently worthless sources of high-grade uranium had been destroyed prior to the 1940s, the development and exploitation of nuclear energy in recent decades would have been seriously impeded. Because each wild plant and animal contains a storehouse of genetic and biochemical information, the loss of a single species might result in the loss of information that could ultimately have great value for man's welfare or survival.

Conflicting attitudes and issues. Although the importance of conservation may seem obvious, most of the world's people live too close to the margin of existence to exercise concern for anything more than their immediate survival and well-being. Planning for the future becomes difficult when the present itself is in doubt, and activities that could help tomorrow's generations may seem quixotic to those for whom survival is at stake. Thus, while conservation has made great strides in some areas of the world, it is still too soon for man to have any feeling of security about the future of his environment.

Short-term versus long-term views. There would seem to be no apparent need for any conflict between the attitude of conservationists and those of other responsible humans toward the environment, yet conflict does exist. As mentioned above, this conflict derives in part from the difference between short-term and long-term viewpoints. It is often regarded as essential to the survival or the enrichment of an individual or a group to use resources in such a way as to realize immediate gains or profits. Such shortsighted activities, however, may impair the future productivity of an area of land, exterminate a species, or destroy the usefulness of a site for any other purpose. In such a situation the short-term, private view conflicts with the long-term, public view. Though many would argue that the public view should be more conservation oriented, emphasizing proper safeguards to prevent deterioration of the environment, there are, nevertheless, times when governments take the short-term view in the face of real or imagined economic or political crises; they may, for example, authorize widespread destruction of resources as a temporary expedient to achieve a military goal or to strengthen the public treasury. But crises tend to become self-perpetuating if the destruction of resources weakens the country ecologically and economically. Thus, continued, unrestricted population growth in a country poorly equipped to manage its natural resources creates a continuing sense of crisis, because ever-expanding immediate needs are commonly met at the cost of future productivity and environmental stability.

Conflicting governmental roles

As long as human populations were small and the pressures upon the environment were limited, conflicts between long-term and short-term interests made little difference. Deteriorated lands could be abandoned and new lands found, because there was sufficient time to permit natural repair of environmental damage. Presently, however, with great and increasing numbers of people on a planet of limited capacity, conservationists are insisting that the difference between short- and long-term points of view be resolved in favour of actions that guarantee the survival of mankind.

Technological issues. Technological progress is another important reason for many conflicts in matters pertaining to conservation. Although technology can be a boon to man in most instances, it can also be poorly related to environmental realities. Through the use of technology great environmental changes can quickly be brought about. Although these changes are usually intended to be beneficial, they frequently occur in natural environments in which all things are ecologically related to one another. As a consequence, the changes may produce side effects that were not anticipated or that were discounted as being of little importance, thereby disrupting other human activities or the environment as a whole. Examples of such situations include the polluting effects of certain industries or the spread of waterborne diseases following the construction of major irrigation projects.

Use of global resources. A further area of conflict lies in attitudes toward resources that are held in common, such as the atmosphere and oceans. In instances in which the use of such resources is essentially free to the user, and the power to control usage does not rest with any

recognized authority, the resource inevitably deteriorates. Although each fisherman may feel that his individual activities have very little effect on the resources of the ocean, the effect of the activities of all fishermen may threaten the existence of those resources. Similarly, each automobile driver does not feel that he is contributing much to the pollution of the global atmosphere, but all automobiles throughout the world contribute a total level of pollution that most critics feel cannot long be tolerated. When such situations exist, a recognized controlling authority is usually seen to be necessary.

II. The history of conservation

EARLY BEGINNINGS

It is a familiar myth, expanded in the philosophical writings of Jean-Jacques Rousseau and others, that primitive man lived in a happy state of balance with nature; because of the direct relationship between early man and his environment, he purposefully developed conservation practices that guaranteed the continued health of his natural surroundings. The available evidence does not fully support this view, although there is no doubt that certain peoples and cultures did develop some conservation-oriented practices. For the most part, however, primitive man survived because his numbers were small and his technology limited, and he was unable to do any major harm to his environment. Even so, the only powerful technique that he possessed, the use of fire, was used virtually without discrimination and greatly modified many areas of the Earth. Animals were exterminated by effective, indiscriminately applied hunting techniques. Thus, the elimination of many large mammals from North America—*e.g.*, the mammoths, mastodons, camels, horses, and giant sloths—has been attributed to primitive hunters, as has the extermination of moas, the giant flightless birds that once inhabited New Zealand. When the Plains Indians of North America received the horse from the Spanish, they quickly developed an exploitative economy that threatened the continued survival of the American bison. Eskimos, when given modern firearms, quickly joined in the slaughter of Arctic wildlife, exceeding any immediate need.

Primitive hunting and agricultural abuses

Even agriculture was not immune to land abuse by early man. When practiced in humid forested regions, primitive agriculture did not cause serious degradations, because the people could not clear extensive forest areas with their available tools; in addition, forest soils tended to lose their fertility in a relatively short period, making them undesirable for continuous cultivation. Because the less fertile sites were abandoned before any severe damage had been done, they were revegetated quickly by forest plants and gradually restored to a productive condition. But, when grassland soils and drier forest areas were invaded by agriculture, the consequences were different. In the Asiatic homelands of Western agriculture there is widespread evidence of serious soil damage and loss during ancient times. Destruction of vegetation and the spread of deserts followed the rise of early civilizations in the Middle East and in northern Africa.

Ancient conservation practices. Certain conservation practices did develop in the earliest of times, however. Some species of animals were protected by religious taboos; religious sanctions prevented the destruction of forest groves and sacred mountains. The use of organic fertilizer to maintain soil fertility is found among many more recent primitive peoples and has had a long history in Western agriculture. The Bible is filled with various injunctions governing the use of land and resources that have a conservation function. Early civilizations such as those of the Phoenicians and the Incas developed sophisticated techniques of terracing to prevent soil erosion on hillsides and to make more effective use of water for irrigation. The earliest civilizations also show evidence of the creation of reserves or parks to protect wildlife or natural areas. Although they were hunting preserves for the use of royalty, they also served a conservation function.

As civilization developed, the accumulation of human experience led to increasingly sound land-use practices, evidence of which is found in the written descriptions of Roman agriculture and, later, in the well-tended irrigated fields and gardens developed during the height of Muslim culture. The agricultural landscapes of preindustrial western Europe, Japan, and China reflected great skill in the conservation of soil resources. Irrigated lands in the Nile Valley and volcanic soils in tropical Southeast Asia have been kept fertile and productive over thousands of years.

Early development of sound land use

In pre-industrial times, however, concern over wild nature was not widespread, largely because it was viewed as vast and inexhaustible relative to the domain, the numbers, and the power of man. This view was a justifiable one, because the 500,000,000 people who inhabited the world in 1600 lacked the energy sources and the machinery to effect great environmental changes. Moreover, most of the Earth's surface was uninhabited or sparsely settled.

Conservation during and following the Renaissance. Starting with the voyages of discovery in the 15th century, the influence of European man was spread over the world. By the 17th century Europeans were equipped with an increasingly powerful technology and a growing ability to modify large areas of the Earth and to subdue less sophisticated peoples. During this period the attitudes of explorers and colonists were oriented more toward immediate personal aggrandizement of the lands they visited and settled than toward any concern for the long-term health and productivity of the newly discovered countries. Soil erosion as well as the destruction of natural vegetation and wildlife accompanied the spread of European colonization in the Americas and later in Australia and Africa. Nevertheless, during the same period, various conservation ideas and practices were being promoted. Forest conservation, for example, developed sound beginnings in 17th-century England and France, in part because of the disappearance of natural forests as a result of the increasing demand for wood fuel for industrial uses. As early as the 18th century in eastern North America, such men as Thomas Jefferson already had sound ideas for land management and conservation, and a general interest in and concern for wildlife was developing.

The 19th century, however, witnessed unusually severe environmental depredations. In Australia, for example, livestock populations were allowed to increase to levels far above what the natural forage could support. Although millions of animals died during drought periods, the process of overforaging damaged the range lands to such a degree that they have not yet recovered. In southern Africa many forms of wildlife were hunted to extinction, and most of the larger mammals were reduced to numbers that endangered their survival. It was in North America, however, that the changes were most dramatic. The great herds of wildlife that inhabited the plains and prairies vanished as the numbers of bison, elk, antelope, and deer were reduced by hunters. Even the larger predatory animals were nearly exterminated, and some of them—varieties of grizzly bear, cougar, and wolf—subsequently became extinct. Many types of birds that once had occurred in great abundance—*e.g.*, the passenger pigeon, Carolina paroquet, and heath hen—were wiped out. Logging and fires combined to menace the once luxurious forests of New England, the states surrounding the Great Lakes, and the South. The grasslands were overgrazed, and in some areas such as California native vegetation was eliminated over most of its range and replaced by species of European and Asian origin. It is in the West Indies and other islands throughout the world that changes were most marked. Native plant and animal species were eradicated and replaced by exotic invaders. By contrast, in the long-settled areas of Europe and Asia, changes were much less marked, as conservation-oriented systems of land management persisted.

Environmental depredations in the 19th century

Rise of the modern conservation movement. It could have been predicted that the modern conservation movement would have its beginnings not in the settled lands of the Old World but in those areas of the New World where, within the memory of a single generation, there

had been extreme changes in the landscape and in the abundance of wildlife. The reaction to the destruction of natural resources in these areas precipitated the formation and growth of the conservation movement. As early as 1832, George Catlin, a U.S. artist and author, first proposed the idea of national parks encompassing major areas in which Indians and wild country could both be preserved. In the same decade the botanist William Bartram and the ornithologist John James Audubon were arousing an interest in wildlife and its conservation. A little later, writers Ralph Waldo Emerson and his friend Henry David Thoreau presented strong arguments concerning the importance of the continued survival of wild nature to the psychological well-being of man. Thoreau became one of the first literary advocates of wilderness conservation. The 1860s in America saw the first textbook on conservation, *Man and Nature*, by George Perkins Marsh. In 1864 the designation of the Yosemite Valley in California as a national park served as the forerunner of the world system of national parks. In the same period the author and naturalist John Muir settled in California and became a leading advocate of wilderness preservation. In 1872 the United States Congress not only proclaimed the Yellowstone region of Wyoming as a national park but also established for the first time a national-government role in the protection and administration of such areas. (Yosemite, the first national park in the United States, was administered by the state of California.) In 1891 the first of the U.S. forest reserves, forerunners of the system of national forests, was proclaimed in the area surrounding the Yellowstone National Park. The way was thus prepared for Theodore Roosevelt and his associates to establish in America the concept that the conservation of natural resources is an important concern of national government.

The establishment of national parks

RECENT HISTORY

The recent history of conservation has been marked by a great expansion of government roles in protecting the environment and by a growth of public interest in and support for this process. National-park systems, dedicated to the preservation of wild nature and to the provision of outdoor recreation space, have grown rapidly from their early beginnings. National-forest systems, dedicated to the multiple use of wild-land resources, also became firmly established. In the United States the conservation of wildlife became a cause of national interest and led to the establishment of a far-ranging system of wildlife refuges and the gradual restoration of most wild animal species to levels approaching, in some cases exceeding, their primitive abundance. On private lands, however, and on government or public-domain lands not specifically reserved as national forests, parks, or refuges, deterioration continued, reaching a peak in the 1930s, when it became widely recognized that those range lands in the public domain had been disastrously overgrazed and that many privately owned farmlands had been depleted or exhausted. Firm control over the management of lands in the public domain and federal intervention to establish soil conservation on privately owned lands were accepted as appropriate activities for the national government.

Spread of modern conservation practices. Conservation ideas spread widely, being most readily accepted by those countries that had experienced sudden environmental changes. By the 1920s national parks were to be found on all continents. In 1924 the Soviet Union established the first of its now extensive system of natural reserves (*zapovedniki*). Conservation-oriented management of forest lands, which grew more from its origins in Europe than from practices in the United States, also became more widely accepted throughout the world. The scientific basis for the management of wild grazing lands for the sustained production of forage for livestock was established in U.S. national forests in 1913 and soon spread to other countries. Aldo Leopold in the United States, in 1933, wrote a textbook on game management, in which the conservation and management of wild animal life for such recreational purposes as sport hunting and fishing and for direct commodity values on a sustained basis received particular emphasis. Leopold's work drew heavily on earlier studies of animal ecology by Charles Sutherland Elton in England; in fact, the establishment in Europe of wildlife reserves and protective laws as well as the managing of lands to produce sustained crops of wildlife long preceded even Elton's work. Subsequently, the management of wild animals in extensive wilderness areas made major strides in Africa, which possesses unusual wildlife resources, and in the Soviet Union, which retains large areas of wild land.

The production of sustained "crops" of wildlife

New conservation problems and approaches. After World War II the field of conservation expanded as new problems arose and as some older approaches proved to have been inadequate. With growing populations and increasing pressures on land and resources, planning for their use by taking into account only a single factor or a few factors at the most was found to be highly unsatisfactory. One such instance was the development of more effective synthetic pesticides for use in the control of disease-carrying insects as well as those that prey most heavily upon agricultural crops. The initial results were remarkable. In such countries as Ceylon, where the insecticide DDT was used to control malaria-bearing mosquitoes, the disease was reduced from being an important cause of human illness and mortality to a low and manageable level. Similarly, agricultural pests were drastically reduced, and crop yields soared in many regions. Eventually, however, it was discovered that the pesticides had unexpected and severe consequences on the environment, and by the 1970s their use anywhere for any purpose was open to serious debate.

All forms of pollution also became a matter of major significance as populations and industrial activities increased after World War II. Air in major cities became toxic; water supplies in many heavily populated areas were contaminated. Nuclear radiations had become a major cause for concern during the 1950s and early 1960s when it was found that radioactive materials from test explosions of atomic and hydrogen bombs spread throughout the entire biosphere instead of being confined to the immediate areas in which the tests were conducted.

In response to the need for a much more integrated approach to environmental problems and to natural-resource management than existed at the time, many countries established ministries for the environment or their equivalent, and in 1969 the United States, by the National Environmental Policy Act, established a national Council on Environmental Quality to oversee and help coordinate those activities of government departments that could have an impact upon the environment.

International environmental problems

By 1970, however, the problems of the environment had become international in scope. The oceans were seriously polluted, and no single country could control the situation. Pesticides and other toxic materials spread by air and water currents throughout the world were causing or threatening to cause environmental damage everywhere. But the need for an international approach to conservation problems found most nations generally unprepared to cope with the situation. Conservation-oriented recommendations aimed at controlling the use of radioactive materials, heavy metals, toxic pesticides, or the dumping of petroleum at sea could not be enforced internationally. The need to regulate the exploitation of marine resources was widely acknowledged, but such regulation was ineffective in the absence of an empowered international authority. In recognition of these problems many international conferences were held, new treaties and conventions were proposed, and the need for regulatory power over the environment at an intergovernmental level was stated frequently. The World Health Organization and the World Meteorological Organization began a global program to monitor pollution levels. The United Nations Educational, Scientific and Cultural Organization launched a major scientific program directed toward the problems of "Man and the Biosphere," and an international conference on environmental problems was held in Stockholm in June of 1972. But many

critics feared that, until the nations of the world were more willing to delegate greater authority to international organizations and to support them financially, little progress toward the solution of global problems could be expected. In existing conditions of international relations, this left each nation to attempt to do what it could within its own boundaries.

III. Types of natural resources

In classifying natural resources it has been traditional to distinguish between those that are renewable and those that are nonrenewable. The former were once considered to be the living resources—*e.g.*, forests, wildlife, and the like—because of their ability to regenerate through reproduction. The latter were considered to be nonliving mineral or fuel resources, which, once used, did not replace themselves. In practice, however, this separation is not entirely satisfactory, for reasons that will be dealt with in a later section. There are, nevertheless, certain aspects of conservation that apply specifically to nonliving resources. They are enumerated below.

1. Beneficiation is the upgrading of a resource that was once too uneconomical to develop. It usually depends upon technological improvements, such as those that make possible the concentration of a dispersed fuel or mineral so that it can be more easily handled, transported, or processed.
2. Maximization is the aggregate of those measures that avoid waste and increase the production of a resource.
3. Substitution involves the use of common resources in place of rare ones, as, for example, the use of aluminum in place of less abundant copper for a variety of products.
4. Allocation is the determination of the most appropriate use for a resource and the assignment of the resource to that purpose. Normally, allocation is controlled by economic factors—the higher the price a resource will bring when used for a particular purpose, the more likely it is that the resource will be used mostly for that purpose—but in government-controlled economies a resource may be reserved only for what are considered to be its most important uses.
5. Recycling, one of the most promising methods for conservation of mineral resources, involves the concentration of used or waste materials, their reprocessing (if this is required), and their subsequent reutilization in place of new materials. If carried out in an organized and consistent manner, recycling can greatly reduce the drain on supplies of minerals. It is also appropriate for products derived from living resources, such as the reuse of wood and paper as well as the reclamation of organic fertilizers from sewage.

Because all natural resources form a continuum, from those that are most renewable in the short term to those that are least renewable, they do not readily lend themselves to a single system of classification. It is useful, therefore, to examine the various types of natural resources in relation to their cycling time; *i.e.*, the length of time required to replace a given quantity of a resource that has been utilized with an equivalent quantity in a similarly useful form. From this point of view, renewable resources can be considered as those with short cycling times and nonrenewable resources as those with very long cycling times. Any resource can be nonrenewable, however, if the demand and rate of utilization exceed its cycling capacity.

The concept of cycling time

Two kinds of natural resources, pasture grass and coal, can be used to illustrate the concept of cycling time. When grass is grazed by livestock or mowed, a crop of it is removed. If provision is made to protect the fertility and structure of the soil and to leave enough seed or adequate roots and vegetative parts to produce new growth, then a grass crop can be removed from a pasture each year for an indefinite period of time. Removal of one year's crop does not diminish the supply available for the next year if the land is cared for properly. The cycling time for this resource may be one year in areas in which climate limits growth, or it may be less than a year if growth can be continuous.

By contrast, the coal resources of the Earth were built up over millions of years. Most were laid down during the Carboniferous Period of geological time (from 345,000,000 to 280,000,000 years ago), when climates were warm. Extensive swamp forests covered large areas of the Earth, and conditions were favourable for plant debris to accumulate in extensive deposits without decomposing and breaking down organically. Subsequently, heat and pressure generated by the deposition of other materials on top of the organic debris and by movements of the Earth's crust transformed the plant remains into coal. Organic debris is still being produced in swamps and marshes, and over millions of years this, too, could become transformed into coal. The time scale is so great, however, that, for human purposes, coal can be considered as a nonrenewable resource. Thus, only the supplies presently available in the Earth's crust can be counted on for future use.

RENEWABLE RESOURCES

Plants and animals. The most clearly recognizable renewable resources are those consisting of or produced by living things. Agricultural crops, animal forage, forest crops, wild and domestic animals—all can continue to reproduce and regenerate their populations as long as environmental conditions remain favourable and an adequate seed source or breeding stock is maintained. Moreover, all can be cropped or harvested without diminishing their supply, provided that the cropping does not exceed the reproduction or growth rate. If it does, the resources will be depleted; and, if the rate of cropping continuousuly exceeds the rate of replacement or regrowth, the resource ceases to be renewable, and the species involved are reduced to the point of extinction. A renewable resource thus can be said to be "mined"—that is, it is removed at a rate that does not permit renewal. The renewability of a living resource is further endangered if the environment required by that resource is allowed to deteriorate or disappear. Sheep in a mountain pasture are a renewable resource only as long as the pasture produces vegetation that will nourish and support the sheep. If the pasture is overgrazed, the vegetation destroyed, and the soil eroded, sheep cease to be a renewable resource in that locality.

Species renewability. The renewability of a living resource varies with the species and with the areas involved. Thus, annual plants, from which a high percentage of cultivated field crops are derived, grow to maturity each year and then die back. They are annually renewable and can be cropped at a relatively high rate. Perennial plants, such as fast-growing poplar trees, may have a much slower rate of renewability, although this depends upon the purposes for which they are used. If seedlings are in demand, they can be cropped annually, and a new supply can be grown from protected seed sources. More realistically, however, the cycling time for these trees depends on the length of time required for them to grow to maturity and to produce seeds from which a new crop can be grown. Certain conifers, such as the Monterey pine, can reach a size adequate to yield useful timber and other wood products in less than 30 years. Other conifers, desirable because of the quality of their mature wood, may be cropped only on a 100-year cycle. Old-growth redwood trees can almost be considered a nonrenewable resource, because the time required to produce their equivalent may be from 500 to several thousand years, well beyond the limits for which people are prepared to plan. Redwood forests used for timber production are managed on the basis of a much shorter cycle and the cutting of younger trees. Such management does not provide for the replacement of 1,000-year-old specimens.

Renewability of annual and perennial plants

Landscape renewability. A distinction should also be made between renewable species and the communities or landscapes they occupy. Although it takes about 100 years to replace a mature coniferous tree, certain types of coniferous forests that are managed for timber production can be logged almost indefinitely if the annual level maintains a sustained yield. When all the inter-

relationships among such factors as soils and plant and animal life are considered, a natural forest that has not previously been disturbed by man may be far less renewable in its totality than the individual species within it. In other words, although the procedures employed in the harvesting of timber may assure a sustained yield, they may, nevertheless, be disruptive to other forms of life in the forest community, in which case species that are intolerant of such disturbance may disappear. It may be exceedingly difficult, therefore, to regrow a new community that resembles the original primitive forest if the area is to be disturbed periodically by timber cutting. It was this consideration, among others, that engendered the need to protect national parks, wilderness areas, and undisturbed research reserves. From certain viewpoints a wilderness is a nonrenewable resource; if it is seriously disrupted by human activity, the time required for the landscape to recover its appearance and for the restoration of the various natural combinations of plant and animal life that contributed to its original wilderness value could involve hundreds of years. The redwood-forest wilderness, as previously noted, could require many thousands of years for complete recovery following a major disturbance. Certain tropical rain forests that were disturbed more than 400 years ago still have not regained their original balance of species and do not resemble undisturbed, primary rain forest in the same region.

Ecosystems defined as resources

Ecosystems. Resources that contain a combination of interacting living and nonliving components are called ecosystems. It is impossible to separate an ecosystem into its living and nonliving components, because the whole constitutes a dynamic system in which there is a flow of energy from sunlight, gases from the atmosphere, and minerals and water from the soil. As a natural resource, soil, in turn, is also a combination of living and nonliving components: it consists of atmospheric gases, water, living and dead organic materials, and more or less finely divided mineral substances. Moreover, soil is a product of the interaction between the living and the nonliving environment. The living components of soil conform to the definition of renewable resources, within the limitations that have been noted, and the mineral components conform to the definition of nonrenewable resources. As long as the living components of soil remain healthy and continue to function, the mineral components are recycled from the soil, through the organic life within it (*e.g.*, bacteria and other micro-organisms), and back to the soil following the decay and breakdown of dead organic materials. Because most forms of terrestrial life are dependent upon it for their continued existence, soil must be maintained in a renewable state. Mining soil, or using it in such a way that its fertility is exhausted and it is washed or blown away by too-rapid erosion, reduces the likelihood that life can continue to exist in the area affected.

Solar energy, air, and water. Solar energy, air, and water are sometimes classed as renewable resources, but solar energy is an inexhaustible resource relative to human time scales and the uses that man can make of it. The supply of solar energy is not affected by any human activities, and the potential lifetime of the Sun is in the hundreds of millions of years, barring any cosmic accidents. The actual amount of solar energy reaching the surface of the Earth, however, is determined by the condition of the atmosphere, which can be and is affected by man.

Air is also an inexhaustible resource in the sense that the uses made of it by man and other living organisms have little effect on its total quantity. The quality of air, however, as measured in terms of its chemical composition or its physical state, is subject to human interference. For life to exist on Earth there must be a proper balance among the nitrogen, oxygen, carbon dioxide, water vapour, and other components of the atmosphere. A layer of the gas ozone, for example, must be maintained in the upper atmosphere to screen out damaging ultraviolet light from the Sun. The accumulation of toxic materials in the air must be kept to a minimum, and the concentration of solid and liquid particles in the atmosphere must not be allowed to reach a level such that it interferes with the influx of solar radiation. All of these factors are affected by human activity and by the effects of this activity on other forms of life.

Water may also be considered an inexhaustible resource, because the total supply of water in the biosphere is scarcely affected by the activities of man or other living organisms. Water is not destroyed by most human uses, although it may be held for a time in combination with other chemicals. But, when needed in a particular place and of a quality useful to man, water must be regarded as a renewable resource, with cycling times dependent on its location and use. Water that falls from the atmosphere as various types of precipitation and then runs off the land surface to form streams and rivers that eventually reach the ocean generally operates on a one-year-renewal cycle. From the ocean the water is evaporated by solar energy and returned to the atmosphere, from which it again falls as rain or some other form of precipitation. In certain locations, however, water has a much longer cycling time; after entering the ground from rainfall, it may percolate slowly through underground channels until it reaches underground reservoirs. In certain arid regions the total water supply may be underground water that accumulated during past ages, when the climate of the region was more humid; since that time there may have been little or no addition to this supply because of the existing climatic conditions. Because its cycling time may be extremely long and dependent upon the frequency with which wet and dry climates alternate in a particular region, such a water resource can be virtually nonrenewable. The total movement of water from oceans to air to land and back to oceans, with all of the various pathways involved, is known as the hydrologic cycle. It is the renewal cycle that determines the amount of water available for human use and for all other purposes in any particular location.

Hydrologic cycle as an annual renewal cycle

NONRENEWABLE RESOURCES

As just noted, renewable resources include resources with widely different cycling times, some so long as to make the resources essentially nonrenewable. Those resources that are usually classified as nonrenewable—*e.g.*, fossil and nuclear fuels and minerals—also exhibit a wide range of properties that affect their management. Fossil fuels, such as coal and petroleum, are the least renewable of such resources because they are effectively exhausted by use and because their rate of formation is exceedingly slow. Most minerals, on the other hand, are not destroyed by use; thus, in a sense, they are renewable and inexhaustible because they can be recycled for further use. But useful supplies of these minerals in accessible locations are exhaustible, and thus they are nonrenewable for human purposes.

Fossil fuels. Fossil fuels are those organic materials that have been converted from their original form by physical and chemical processes within the Earth's crust into a solid mineral state (coal), a liquid (petroleum), or a gas (natural gas). If these substances are completely burned (oxidized) when used as fuel, the end products are carbon dioxide, water, and heat energy. These cannot be reconstituted into organic substances without either elaborate synthesis in a chemical laboratory or the natural photosynthetic processes of green plants. Thus, burning destroys fossil fuels as useful energy sources available to man.

On the basis of existing knowledge of the amount of fossil fuels in the Earth's crust, it has been predicted that supplies of petroleum and natural gas may be exhausted by 2070 if used at the rates anticipated. Although coal supplies are greater, projected rates of use indicate that they cannot be expected to last for more than a few centuries. These predictions can be changed, of course, if rates of use change, which is expected to happen with, for instance, the further development of nuclear-power sources. Fossil fuels are used for purposes other than fuel, however. Coal and petroleum are used industrially for the manufacture of a wide variety of carbon-con-

Possibility of exhausting fossil fuels

taining materials, such as plastics, synthetic fibres, medicines, and food.

Nuclear fuels. Although nuclear fuels are inorganic substances, like fossil fuels they are destroyed when used in the production of heat energy. Unlike fossil fuels, however, they are also destroyed by spontaneous disintegration, through natural radioactivity. Uranium, for example, ultimately changes to lead, but the rate of change is slow; its half-life (the length of time it takes for half the atoms of a given amount of a radioactive substance to disintegrate) is 7,600,000,000 years. Of the naturally occurring nuclear fuels, uranium and thorium, only uranium-235 can be used directly in a nuclear reactor to produce power. The more common form of uranium, uranium-238, must first be converted into plutonium before it can be used as a nuclear fuel. Similarly, thorium must be transformed into uranium-233 before it is usable for fuel purposes.

Although supplies of uranium and thorium are relatively abundant, they are exhaustible and nonrenewable. By replacing nuclear fission with nuclear fusion as a power source, however, deuterium (an isotope of hydrogen) can be used as the fuel. Because it occurs in substantial quantities in seawater, deuterium is practically an inexhaustible resource for human purposes.

Minerals. Certain minerals, such as iron and aluminum, are so widely distributed throughout the crust of the Earth that the amounts exceed any foreseeable human needs. Other minerals, such as the precious metals (*e.g.*, gold, platinum, silver), are much more limited in their distribution and quantity. The usefulness of a mineral to man, however, depends upon its accessibility and concentration; therefore, minerals that are highly dispersed throughout the Earth are essentially unavailable, even though their total quantity may be great.

Ores and reserves. Most efforts to obtain minerals are directed toward finding mineral ores, which are deposits in which the concentration and quantity of a mineral are such that it can be extracted profitably. Mineral reserves are those that are known to exist or can reasonably be inferred to exist from geological evidence. It follows that certain deposits of minerals that are not considered either ores or reserves today may become so if the technology for their extraction improves, if the supplies of energy available for their extraction increases, or if their economic value increases.

It is obvious from past experience that ore deposits can be exhausted. Because the gold mines around Virginia City, Nevada, and the tin mines of Cornwall no longer can produce significant quantities of minerals, they can be considered as virtually exhausted resources. The rich iron ores of the Mesabi Range in Minnesota have largely been depleted, and mining activity in this area has shifted to lower grade iron deposits. The supplies available from any mineral deposit, at least on dry land, are exhaustible and nonrenewable, because the geological processes that led to the formation of these deposits operate slowly and over long periods of time.

Some mineral deposits, however, are renewable. Manganese ore, for example, is relatively scarce on dry land but is continuously being formed in nodules on the ocean floor, as are cobalt, nickel, and copper. The rate at which the nodules of manganese, cobalt, and nickel are growing through chemical precipitation from seawater currently exceeds the rate at which these minerals are being used. Although the nodules are not yet being collected, the technology for doing so is being developed. Then these metals will be considered as renewable natural resources only so long as the rate of use does not exceed the rate of formation.

Oceanic mineral deposits

Recycling of minerals. The use of most metals does not destroy them, although rusting may reduce their quantities by a small amount when they are in use. As commercial products, some metals are found in such large quantities in urban areas that their new concentration may exceed that which existed while they were in the ground. Cities, therefore, may be considered as ore deposits for certain minerals. At present, however, it is cheaper to mine new ores than to recycle used or waste metals, but this economic balance does not take into account the cost of disposing of the metallic wastes that accumulate in urban regions. Thus, it is likely that sometime in the future many metals now considered as exhaustible, nonrenewable resources will be treated as recyclable resources.

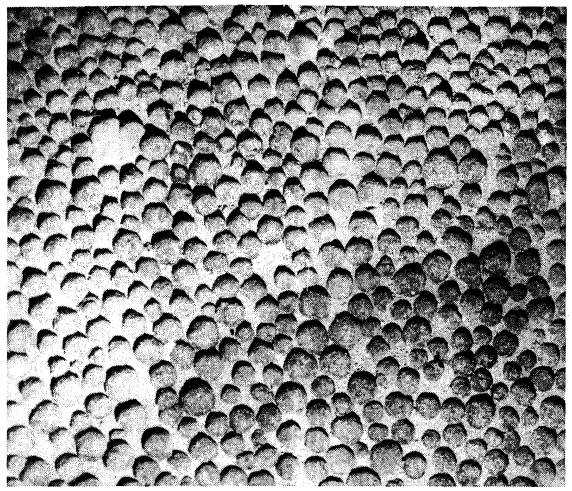


Figure 1: Manganese nodules on the southern Pacific Ocean floor.
By courtesy of the Lamont-Doherty Geological Observatory, Columbia University

Not all minerals can be recycled under existing conditions, however. The concentrated phosphates that are used in fertilizers and detergents, for example, are dispersed widely over the farmlands and waters of the Earth, from which they enter the life cycles of various organisms and eventually, through erosion or in wastes, reach the oceans. Because these phosphates are virtually irretrievable and because the rate at which available reserves are being used probably exceeds the rates at which new reserves are formed, such minerals are considered as nonrenewable and exhaustible resources.

IV. Management of natural resources

MANAGING NONLIVING RESOURCES

Soils. *Formation of soil.* Soils are the basis of support for most terrestrial life and a source of nutrients for freshwater and marine life. As noted above, soil is formed over time as the result of interaction between the living and the nonliving environment—climate, organisms, and the physical surface of the Earth. Rocks are broken apart by the action of sun, wind, rain, snow, sleet, and ice. With the aid of wind and water movements as well as gravity, rock particles from high elevations are deposited on mountain slopes or in valleys, where they are further acted upon by the local climate, by plant and animal life, and by such other environmental factors as fire until they become soil. Nitrogen from the atmosphere, formed into nitrates by the action of lightning and atmospheric water vapour, may enter the soil with rainfall. Other nitrates may be added by the action of such living organisms as soil bacteria and various algae that can convert atmospheric nitrogen into the nitrates required for plant growth. These and other chemicals in the soil eventually become part of the living tissue in plants and animals. The chemicals are returned to the soil as organic wastes and litter that form humus, which is partly decomposed organic material. As humus continues to decompose, the chemicals within it enter the soil for further use by plants and animals.

Soils vary from place to place depending upon the rocks and minerals from which they are derived, the nature of the local climate, and the kinds of organisms that live in or on them, as well as the amount of time that these factors have been operating. Developmental soils—*i.e.*, those still being modified by climate and organisms—reveal the nature of the parent materials from which they are derived; mature soils, those that have achieved a balance among the various forces op-

Developmental and mature soils

erating on them, show in particular the influence of the climate and vegetation in which they develop. Soils also differ greatly in their inherent fertility and in their ability to support life. Those derived from quartz sand, for example, may be naturally deficient in calcium, magnesium, and other elements essential to plant growth. The surface layers of those soils developed in humid, forested regions are often heavily leached, as rainwater containing weak organic acids percolates through them and dissolves the more soluble minerals.

Because soils are essential to man for such purposes as growing crops, forage, and timber, it is important that they not be allowed to wash or blow away more rapidly than they can be regenerated, that their mineral fertility not be exhausted, and that their physical structure remain suited to the continued production of desired plant materials. The objective of soil management, therefore, is to keep soil in place and in a state favourable to its highest possible productive capacity.

Soil erosion. In the past and, to a considerable degree, even now, soils have not been managed effectively. Those exposed through cultivation to the erosive effects of wind have been blown away; those laid bare on sloping ground have been washed downhill by rainfall. Although soil erosion has long been recognized as a major conservation problem, erosion as such—and its converse, the deposition of eroded soil particles—is not a problem but a normal and natural process leading to both soil development and maintenance. Soils exist only because of past erosion and deposition. The conservation problem involved in soil erosion is the accelerated erosion that occurs when soil cover in the form of living or dead plant material is removed. In such cases the soil then erodes at a rate faster than it can be replaced by normal deposition of particles on the soil surface or by the breakdown of rocks and minerals. In severe cases, such erosion leads to the formation of deep gullies that cut into the soil and then spread and grow until all the soil is removed from the sloping ground. Under severe wind action, the finer particles of surface soil are blown away and form drifts and dunes, leaving only the coarser sands and gravels on the soil surface.

Although measures to stop soil erosion are now widely used in most technologically advanced countries, the problem remains a major one in the developing nations of the world. It is particularly severe in the tropics, where high rainfall and steeply sloping ground favour the rapid loss of any soil exposed by agriculture, and around the edges of the world's deserts, where destruction of natural plant cover by cultivation or livestock grazing causes soil loss through wind action and the spread of desert-like conditions.

Erosion by wind and water

To prevent wind erosion, shelter belts of trees have been planted to break the force of the wind. The practice of covering soils with plant litter (mulch) when they are not actually covered with growing plants also helps to hold them in place. Cultivating at right angles to the direction of the wind further serves to prevent wind erosion.

Water erosion on sloping ground may be prevented by terracing on steep slopes or by contour cultivation on more gentle slopes. In the latter a slope is plowed along horizontal lines of equal elevation. Strip-cropping, in which a close-growing crop is alternated with one that leaves a considerable amount of exposed ground, is another technique for reducing water erosion; the soil washed from the bare areas is held by the closer growing vegetation. In the tropics a shelter over the ground serves as a means for breaking the force of raindrops, thus reducing their erosive power, and also to screen out direct sunlight. In addition to causing damage to certain crops, sunlight can accelerate the breakdown of organic materials in the soil at a rate that is faster than is desirable.

Soil fertility. A conservation problem equally as important as that of soil erosion is the loss of soil fertility. Most agriculture was originally supported by the natural fertility of the soil; and, in areas in which soils were deep and rich in minerals, farming could be carried on for many years without the return of any nutrients to the soil other than those supplied through the natural breakdown of plant and animal wastes. In river basins, such as that of the Nile, annual flooding deposited a rich layer of silt over the soil, thus restoring its fertility. In areas of active volcanism, such as Hawaii, soil fertility has been renewed by the periodic deposition of volcanic

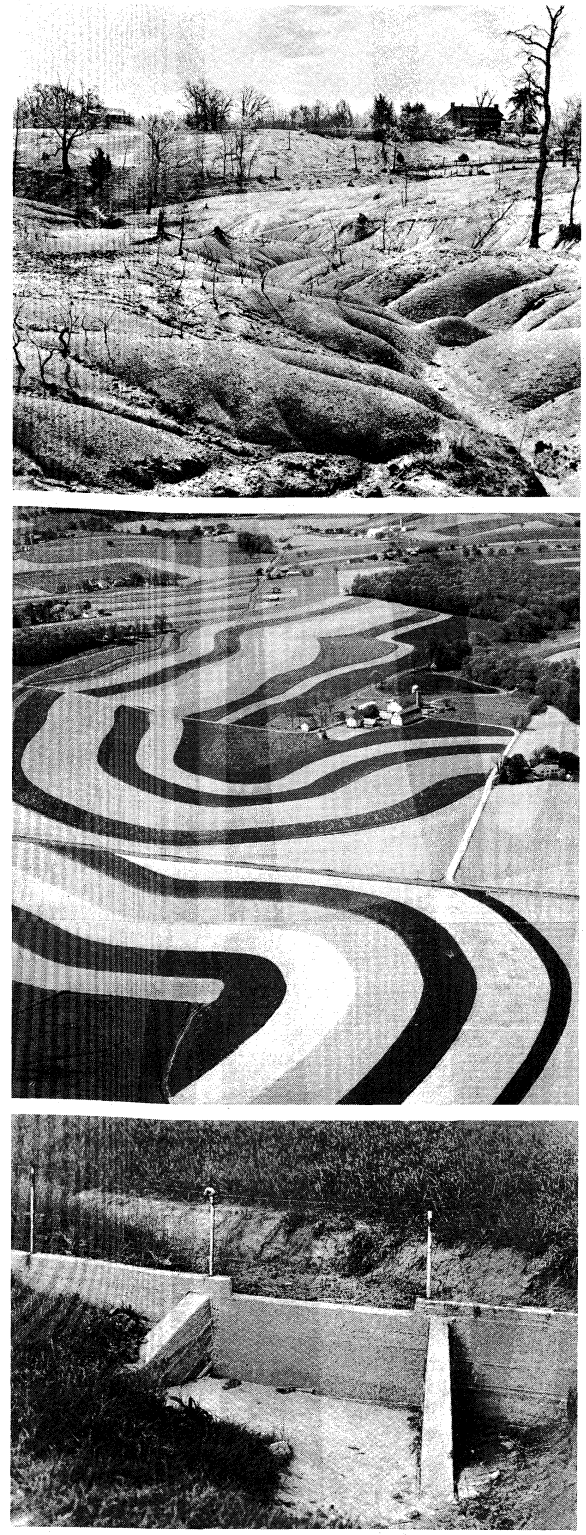


Figure 2: ***Soil erosion and ways to prevent it.***
(Top) Eroded farmland in west Tennessee, resulting from tenant–absentee-owner relationship. (Centre) Contour farming in Pennsylvania. (Bottom) Concrete dam built to protect field from erosion at Delavan, Illinois.
By courtesy of (top) the U.S. Department of Agriculture; photographs, (centre) Grant Heilman, (bottom) J.C. Allen and Son

ash. In other areas, however, natural fertility has been quickly exhausted. This is true of most forest soils, particularly those in the humid tropics. Because continued cropping in such areas caused a rapid decline in fertility and therefore in crop yields, fertility could be restored only by abandoning the areas and allowing the natural forest vegetation to return. Over a period of time the soil surface would be rejuvenated by parent materials, new circulation channels would form deep in the soil, and the deposition of forest debris would restore minerals to the topsoil. Primitive agriculture in such forests was of a shifting nature: areas were cleared of trees and the woody material burned to add ash to the soil; after a few years of farming, the plots would be abandoned and new sites cleared. As long as populations were sparse in relation to the area of forest land, such agricultural methods did little harm. They could not, however, support dense populations or produce large quantities of surplus foods.

Fertilizers and crop rotation

Starting with the most easily depleted soils, which were also the easiest to farm, the practice of using various fertilizers was developed. The earliest fertilizers were manures, but later larger yields were obtained by adding balanced combinations of those nutrients (*e.g.*, potassium, nitrogen, phosphorus, and calcium) that crop plants require in greatest quantity. Because high yields are essential, most modern agriculture depends upon the continued addition of chemical fertilizers to the soil. Usually these substances are added in mineral form, but nitrogen is often added as urea, an organic compound.

Early in agricultural history it was found that the practice of growing the same crop year after year in a particular plot of ground not only caused undesirable changes in the physical structure of the soil but also drained the soil of its nutrients. The practice of crop rotation was discovered to be a useful way to maintain the condition of the soil and also to prevent the buildup of those insects and other plant pests that are attracted to a particular kind of crop. In rotation systems a grain crop is often grown the first year, followed by a leafy-vegetable crop in the second year and a pasture crop in the third. The last usually contains legumes (*e.g.*, clover, alfalfa), because such plants can restore nitrogen to the soil through the action of bacteria that live in nodules on their roots.

Salinization of soil. In irrigation agriculture, in which water is brought in to supply the needs of crops in an area with insufficient rainfall, a particular soil-management problem that develops is the salinization (concentration of salts) of the surface soil. This most commonly results from inadequate drainage of the irrigated land; because the water cannot flow freely, it evaporates, and the salts dissolved in the water are left on the surface of the soil. Even though the water does not contain a large concentration of dissolved salts, the accumulation over the years can be significant enough to make the soil unsuitable for crop production. Effective drainage solves the problem; in many cases, drainage canals must be constructed, and drainage tiles must be laid beneath the surface of the soil. Drainage also requires the availability of an excess of water to flush the salts from the surface soil. In certain heavy soils with poor drainage, this problem can be quite severe; for example, large areas of formerly irrigated land in the Indus basin, in the Tigris–Euphrates region, in the Nile Basin, and in the Western United States have been seriously damaged by salinization.

Watershed soil. The soils of wild lands in all areas that yield water to streams and rivers are as important to man's welfare as are those in which he raises agricultural crops. If these soils are kept in place and in good condition, they support trees, forage, and wild animal life and yield clear water for human use. If, however, these soils are damaged physically by being compacted or are allowed to erode, they lose not only their capacity to support vegetation but also their capacity to hold and slowly yield useful water to streams or springs. Furthermore, eroded soils cause siltation of waterways; they will also accumulate in lakes and reservoirs, filling them and thereby reducing the useful purposes that these bodies of water serve.

In order to maintain soils in watershed areas, their vegetative cover must be retained. This requires avoiding excessive disturbance of forest vegetation and soil cover through logging, avoiding the excessive grazing and trampling of pasture lands, and preventing the kinds of wild-land fires that destroy plant cover and expose the soil. Often, the success of intensive land and water use downstream depends on the care taken of the soils in the less intensively used forest or range watersheds upstream.

Water. Life originated in the oceans, and the chemical composition of body fluids in land animals reflects their primeval origin. The dependence of life on water is complete; it is the major constituent of plant and animal cells. Most of the major groups of animals still live in water; a relatively small number have adapted to life on dry land.

Dependence of life on water

Uses of water. Man requires water for a variety of purposes; water for drinking is still paramount, and such water must be relatively pure. Water is also required by domestic animals and plants; and, if it is not supplied in sufficient amounts through precipitation, it must be supplemented by irrigation systems. Irrigation, however, is one of the most wasteful uses of water in areas in which it is scarce, because great quantities are lost through evaporation in both storage areas and transport. In many regions irrigation is, nevertheless, essential for human survival.

Water for transportation has always been important to man, as indicated by the fact that most major cities are located on the shores of oceans and other large bodies of water or along rivers and other types of navigable waterways. Despite recent advances in ground and air transportation, water transportation has an economic advantage for the movement of goods that have a relatively low value per unit of weight or volume, such as raw mineral ores, fuels, and various types of construction materials. Water for urban use other than drinking serves a multitude of purposes, such as fire fighting, street cleaning, sanitation, and sewage disposal. Steel mills, pulp mills, chemical factories, and most other industrial processes that involve the conversion of raw materials into finished products require water. Next to agriculture, one of the most extravagant uses of water is as a cooling fluid in the generation of power from fossil and nuclear fuels, with the latter consuming far greater volumes. Water has been used directly as a source of power since the time of the first sailboat and the first waterwheel. A small but important part of the world's electrical supply now is generated by hydropower, in which the force of falling water is used to turn turbines that produce electricity.

Husbandry of water supplies. Although water is a renewable resource, the many demands for water of a desired quantity and quality in a particular place require careful husbandry of the supply. After reaching the surface of the Earth as rain, water enters a supply system either by penetrating the ground and moving through subsurface channels, known as aquifers, or through runoff into streams and rivers. As mentioned above, the supply and quality of water depend in part on the management of the vegetation and soil in the watershed areas. Also involved is the control of streamflow or the control of pumping from underground sources. In many parts of the world, where rainfall is seasonal, streams run at flood levels during the wet season but are extremely low or completely dry at other times of the year. River-basin-management techniques attempt to equalize this variable supply for human purposes, in part through watershed management and in part through the capture of water by dams and its storage in reservoirs.

Consequences of mismanagement of water

When water is mismanaged, a high percentage is lost through evaporation in watersheds. Moreover, as a result of poor management of watersheds, the seasonality of water flow is more acute: floods that destroy lands in the river basins become more frequent during the wet season, and there is an increase in the frequency of droughts

during the dry or low-rainfall season. Soil eroded from watersheds impairs the functioning of dams, reservoirs, and other structures downstream. Furthermore, because of mismanagement water becomes polluted at various stages in its movement from atmosphere to land and thence to the oceans.

Effective water management starts when precipitation first reaches the ground, after which the quality and quantity of water must be protected at every critical point along the hydrologic cycle. Hence, although it may have been practical to use streams, lakes, and the oceans as dumping areas when populations were low and water was abundant, this practice becomes untenable when populations increase and supplies of water decrease relative to human needs. Thus, except in sparsely populated areas of the world, population and technological growth have made necessary the prevention of erosion, the recycling of wastes so that nutrients are restored to the soils and useful minerals are reclaimed for reuse, and the reuse of water to the maximum possible degree.

"New" sources of water. Lack of water of proper quality and quantity has been a major factor affecting urban and industrial growth. To overcome this problem, water has been transported great distances—*e.g.*, the channelling of Rocky Mountain water from the Colorado River to the city of Los Angeles. One project, now only projected, involves the reversal of Siberian rivers to meet the demands of urban and rural areas along rivers flowing into the Caspian and Black seas.

Desalinization of seawater

The use of the oceans as sources of fresh water is being developed in many areas. The nation of Kuwait, a desert amirate in Arabia, now receives much of its water supply through the desalinization of seawater, as do a number of small communities and several large urban centres elsewhere in the world. With the further development of nuclear energy as a power source, it is expected that seawater will be used to an even greater extent as a source of fresh water. Morever, the materials reclaimed from seawater could, if power is available for their separation and concentration, help in meeting many of the world's mineral needs. It seems unlikely, however, at least with foreseeable sources of power, that desalinized ocean water will be extensively pumped to inland regions. Meeting the growing needs of such areas will require the purification of waters polluted by urban or industrial use or of waters that have become salinized through their use in irrigation. The reuse of such waters could go far toward reducing the need for new water by inland communities.

Air. Concern about the quality of air is a relatively recent development, although polluted air has been a problem of urban communities for many centuries. As early as the 13th century in London, a royal decree forbade the use of soft coal for heating because it was a source of obnoxious fumes. The decree was not enforced very long, and by the 19th century the widespread use of soft coal for heating had become the cause of London's infamous "black fogs." In addition to emitting particles of smoke and soot that contribute to black fogs, the burning of soft coal also produces sulfur dioxide, a gas that, when combined with water vapour, forms sulfurous acid. Because sulfur dioxide poisoning was believed to have been responsible for the deaths of thousands of persons in London during the decades before 1960, stringent measures to curtail air pollution were enacted. As a result, air pollution from industrial sources and space heating has been greatly reduced. At the same time, however, the less visible but equally dangerous pollution from automobile engines has increased.

Air pollution results from a variety of causes, not all of which are man's responsibility. Dust storms in desert areas and smoke from forest and grass fires contribute to chemical and particulate pollution of the air. Forest fires that swept the state of Victoria, in Australia, in 1939 caused observable air pollution in Queensland, over 2,000 miles (3,000 kilometres) away. Dust blown from the Sahara has been detected in West Indian islands. The discovery of pesticides in Antarctica, where they have never been used, suggests the extent to which aerial transport can carry pollutants from one place to another. Probably the most important natural source of air pollution is volcanic activity, which at times pours great amounts of ash and toxic fumes into the atmosphere. The eruption of such volcanoes as Krakatoa, in the East Indies, and Katmai, in Alaska, has been related to measurable climatic changes.

Air pollution may affect humans directly, causing a smarting of the eyes or coughing. More indirectly, the effects of air pollution are experienced at considerable distances from the source, as, for example, the fallout of tetraethyl lead from urban automobile exhausts, which has been observed in the oceans and on the Greenland ice sheet. Still less directly experienced are the possible effects of air pollution on global climates.

Urban air pollution. It is the immediate effect of air pollution on urban atmospheres that is most noticeable and causes the strongest public reaction. The city of Los Angeles has been noted for both the extent of its air pollution and the actions undertaken for control. Los Angeles lies in a coastal plain, surrounded by mountains that restrict the inward sweep of air and that separate a desert from the coastal climate. Fog moving in from the ocean is normal to Los Angeles climate. Temperature inversions characterized by the establishment of a layer of warm air on top of a layer of cooler air prevent the air near the ground from rising and thus effectively trap pollutants that have accumulated in the lower layer of air. In the 1940s, Los Angeles air became noticeably polluted, interfering with visibility and causing human discomfort. Attempts to control pollution, initiated during the 1950s, resulted in the successful elimination of such sources of pollution as industrial effluents and the outdoor burning of trash and debris. Nevertheless, pollution continued to increase as a result of the increased number of motor vehicles. Exhaust fumes from the engines of automobiles contain a number of polluting substances, including carbon monoxide and a variety of complex hydrocarbons, nitrogen oxides, and other compounds. When acted upon by sunlight, these substances undergo a change in composition producing the brown, photochemical smog for which Los Angeles is well known. Efforts to reduce pollution from automobile engines and to develop pollution-free engines may eventually eliminate the more serious air pollution problems. In the meantime, however, air pollution has driven many forms of agriculture from the Los Angeles basin, has had a serious effect upon the pine forests in nearby mountains, and has caused respiratory distress, particularly in children, elderly people, and those suffering from respiratory diseases.

The air pollution problem of Los Angeles

Los Angeles is neither a unique nor the worst example of polluted air. Tokyo has such a serious air-pollution problem that oxygen is supplied to policemen who direct traffic at busy intersections. Milan, Munich, Rio de Janeiro, and Buenos Aires face similar problems. The small town of Knapsack, in Germany, was completely abandoned by its citizens in 1971 as a protest against industrially caused air pollution. Although New York City produces greater quantities of pollutants than Los Angeles, it has been spared from an air-pollution disaster only because of favourable climatic circumstances.

The task of cleaning up air pollution, although difficult, is not believed to be insurmountable. Use of fuels that are low in pollutants, such as low-sulfur forms of petroleum; more complete burning of fossil fuels, at best to carbon dioxide and water; the scrubbing of industrial smokestacks or precipitation of pollutants from them, often in combination with a recycling of the pollutants; and the shift to less polluting forms of power generation, such as nuclear fuels in place of fossil fuels—all are methods that can be used for controlling pollution. The example of London, as well as of other cities, has shown that major improvements in air quality can be achieved in ten years or less.

Climatic effects of polluted air. Less obvious than local concentrations of pollution but potentially more important are the climatic effects of air pollutants. Thus, as a result of the growing worldwide consumption of fossil

fuels, atmospheric carbon dioxide levels have increased steadily since 1900, and the rate of increase is accelerating. Now the output of carbon dioxide is believed by some to have reached a point such that it may exceed both the capacity of plant life to remove it from the atmosphere and the rate at which it goes into solution in the oceans. In the atmosphere carbon dioxide creates a "greenhouse effect." Like glass in a greenhouse, it allows light rays from the Sun to pass through, but it does not allow the heat rays generated when sunlight is absorbed by the surface of the ground to escape. An increase in carbon dioxide, therefore, can cause an increase in the temperature of the lower atmosphere. If allowed to continue, this might conceivably cause melting of the polar ice caps, raising of the sea level, and flooding of the coastal areas of the world. There is as yet, however, little evidence that such an increase in temperature is taking place.

The "greenhouse effect"

Counterbalancing the effect of carbon dioxide is the increase of particulate matter in the air, a result of the output of smoke, dust, and other solids associated with human activity. Such an increase might, in turn, increase the reflectance, or albedo, of the atmosphere, causing a higher percentage of solar radiation to be reflected back into space. This, in time, could cause a lowering of the Earth's surface temperature and, potentially, a new ice age. It is not known, however, which of the two effects is more likely to occur, because until relatively recently there existed no monitoring system able adequately to measure the necessary parameters of the atmosphere and to detect the changes that are taking place globally. Such a system is now in operation; after measurements have been made over a period of years, scientists will be better able to predict the consequence of man's activities on the composition of the atmosphere.

Radioactive contamination of the atmosphere. During the 1950s the effects of atmospheric testing of atomic and hydrogen bombs became a source of major concern. The danger of radioactive pollution of the air and the fallout of radioactive particles to the surface of the Earth stimulated serious investigation, resulting in the discovery of some potentially dangerous conditions. It was observed, for example, that radioactive materials of many kinds, such as radioactive iodine and strontium, are concentrated in living tissue and can cause damage even when the general level of environmental contamination is low. Fortunately, atmospheric testing of nuclear bombs was stopped, and radioactive fallout has not become the worldwide menace that was anticipated. Concern is still expressed, however, over possible air pollution from nuclear-power plants, and scientists disagree on its importance. Further study and free public access to all of the facts are essential to remove fears connected with the use of nuclear energy for any purpose.

Subsurface deposits. Several types of conservation activities are associated with the use of fossil fuels and minerals. First are those that involve making the available fuel and mineral reserves serve the most worthwhile purposes for the longest period of time. Second are the problems associated with extracting fuels and minerals from the ground and their subsequent transport, processing, and manufacture, all of which can directly affect the quantity or quality of other resources or the general environment. Third are the side effects, which, because they involve pollution, are considered in the next section.

Conserving exhaustible fuel and mineral resources. As has already been noted, fossil fuels and minerals fall into two categories: those that are destroyed by use and those that retain their physical and chemical characteristics during use and are capable of being reclaimed for other uses. The outlook is most bleak with respect to the conservation of those fuels and minerals that are destroyed by use. If petroleum continues to be used at its present rate, the supply will be exhausted or reduced to a level at which further use may be restricted by the high cost of the remaining resource. Because the situation is generally recognized, conservation need involve only allocating the resource to those uses for which it is best suited and restricting its use in those cases in which other materials, less likely to be exhausted, can be equally well substituted. Thereafter, it is necessary to maximize the available resource. This can be done by avoiding waste in its extraction, transportation, and processing and by utilizing fully all that can be made available. Thus, several oil companies no longer pump competitively from the same oil field, leaving only unobtainable oil in the ground. Instead, fields have been unitized; companies now cooperate in bringing out oil by using only strategically located pumping units. The practice of burning the natural gas from oil fields has been replaced by tapping and piping the gas for use as a fuel. Unrestrained gushers are a thing of the past; more important are the new methods of drilling and of injecting gas or water under pressure to force out oil that previously would have been left in the ground. Improved methods for refining petroleum have eliminated much of the waste that once occurred in this process. Yet accidental losses still do occur. The leaking well that caused such extensive damage to wildlife and recreational resources in the Santa Barbara, California, channel in 1969 resulted in part from failure to use available safety devices. A similar failure was involved in an oil well off the Louisiana coast in 1970. Great quantities of oil are still wasted and waters polluted because of improper navigation of oil tankers and because some such tankers are so poorly constructed that excessive quantities of oil can escape as a result of relatively minor damage to the hull.

Maximizing available resources

Beneficiation, the concentrating of relatively dispersed or low-grade resources into a form in which they can be handled economically, is another means of extending the supply of exhaustible petroleum resources. There are, for example, great quantities of petroleum available in oil shales and tar sands in various parts of the world. As long as petroleum was plentiful in a more concentrated form in oil fields, it was not economical to extract the more dispersed petroleum in such shales and sands. Now, with the prospect of a shortage of petroleum, techniques are being devised for concentrating these lower grade supplies. One method has involved the use of underground nuclear explosions; the resulting heat and pressure force oil and gas into cavities created by the blast.

Conservation in the case of recyclable minerals will involve reuse. For this to be accomplished, incentives and methods may be necessary to encourage the gathering of used materials for reprocessing. Provisions have been made for the collection of junked automobiles, waste cans, bottles, and other containers and for the reuse of building materials from demolished structures. Although much of this is done more to prevent pollution than to reclaim the materials, it does serve both purposes. Another conservation measure is waste-processing technology and the growth of industries concerned with the recycling of wastes.

Just as applicable to the conservation of minerals as they are to the conservation of petroleum are such other techniques as the allocation of scarce resources to their most essential uses, the substitution of other resources for those that have become scarce, and the maximization of supplies. Moreover, there has been great progress toward the development of composite materials, in which relatively small amounts of metal are used in combination with plastics and ceramics. Metallic alloys that minimize the need for scarce materials (as in the substitution for copper or silver in coins of various cheaper metallic alloys and the substitution of aluminum alloys for copper or steel wire, for roofing materials, or for tin in cans) are in many cases more effective than those used for the same purposes. Beneficiation has been used increasingly; thus, low-grade taconite and jasperite ores are now being mined and pulverized, and their iron components separated and pelleted before being shipped to smelters. Such techniques make possible the use of iron deposits that previously would not have been considered economical sources of ores.

Because of the many ways for making better use and reuse of the available supply of metals and other minerals and because the extent of new discoveries cannot be foretold, it is difficult to predict the likelihood of de-

pleting any particular mineral resource; for example, the wasteful use of metals and minerals in most affluent societies creates a drain upon supplies that is entirely unnecessary. Predictions of resource use based on the continuation of wasteful practices are likely to prove unrealistic if these practices are changed voluntarily or forced to change by a scarcity of materials.

Predictions of depletion of a resource

Conservation problems caused by mining. Some of the most serious conservation problems are associated not with the use of minerals or fuels but with the methods used to extract these resources. Mining for coal has created widespread devastation in the Appalachian Mountains, in Bohemia (Czechoslovakia), and, in the past, in England and Germany because deposits in these regions are found near the surface. In the removal of the coal, soil and all living resources have been destroyed, leaving behind a barren, denuded, and eroded wasteland. Because mining wastes often contain much sulfur, their watery runoff contains sulfuric acid, which destroys aquatic life in streams. Similar problems are associated with the surface mining of nickel in such places as New Caledonia and Australia. In the former, as much as one-quarter of the island may be destroyed if current practices continue. Dredging for tin in Malaysia has also created widespread conservation problems. The mining of phosphates has brought destruction to many Pacific islands; one-half of the island of Nauru has already been devastated, and the other half is now being mined. Titanium mining in the sands along the Australian coast has destroyed natural vegetation on beaches, sand islands, and dunes and opened pathways for dune movement that destroys still greater areas of vegetation. Dredging for gold has damaged many streambeds and riverbeds because it is destructive to aquatic life and water quality.

The means for preventing damage by surface mining do not exist, but damage can be controlled and minimized, in part by the creation of erosion-control structures and the prevention of water pollution. The rehabilitation of mined-over land may involve, among other things, the removal and safe storage of topsoil before mining begins and its restoration after mining operations have been completed. It may also involve the shaping of mining wastes into landforms that can be covered with soil, replanted, and revegetated. In England and Germany attractive recreational landscapes have been created on land that once would have been left a devastated wasteland. Of equal importance, however, is the decision not to mine areas that have high surface-resource values—*e.g.*, highly productive forest or farmlands, certain urban areas, and areas with natural qualities of scenery, plant life, or wildlife that make them suitable to be maintained as parks or reserves. The ores in such areas are not mined unless the need for their minerals is most urgent.

MANAGING LIVING RESOURCES

Any area of land and water not yet modified by man can be managed in a number of ways. Certain choices must be exercised early, however, because even minor deterioration of an area through unplanned use may make it unusable for its intended purposes. Other choices, particularly those involving modification of the natural features of the landscape, may be exercised at virtually any time; any previous changes in the character of the living resources in a region to be modified have no effect on the ultimate structural modification of that region. In planning for the use of any undeveloped area, therefore, those purposes that have the most exacting requirements must be considered first; they often depend upon the continuance of relatively undisturbed conditions as well as upon the maintenance of the full variety of wild species and the natural environment within the area.

Natural communities. The idea that biological communities should be protected for their own intrinsic value is of relatively recent origin. Although natural communities have been protected since ancient times, the reasons for doing so have not been related to the value of the community per se but to some special feature that was of value to man. Thus, hunting preserves were protected in ancient Mesopotamia, in China, and in England, where the New Forest was set aside by William the Conqueror. While such preserves protected natural areas, their major purpose was to provide a setting for royal hunting. Temple gardens have been preserved over the centuries in China and Japan; the cedars of Lebanon were maintained around holy places. But, again, such preservation was fortuitous rather than intentional.

The idea of preserving wild areas for their own value had its origin in the United States with Catlin, Thoreau, Muir, and others of similar mind. Only more slowly has this concept been accepted in other countries. Appreciation of wild nature is acquired along with scientific knowledge, particularly ecological knowledge. It is a sophisticated taste not usually to be found among those who earn their livelihood in close contact with the wild. Nevertheless, by the 1970s the concept of preserving wild nature for its own value had been widely accepted, although the means for implementing it were not necessarily available in most countries.

Value of natural communities

Natural communities, little affected by man's activities, are thought to be worth preserving for a variety of reasons. First, perhaps, is the scientific benefit to be derived from studying them, particularly concerning the functioning of the biosphere. From studies of undisturbed ecosystems much can be learned about the behaviour of those systems modified by man for the production of useful materials. Also, the value of wild species to man has been little explored; in their totality they are known to be essential to the function of the biosphere, but the importance of individual species is little understood.

Past experience has demonstrated that wild species of little apparent value may prove to be of major importance to medical research and human health. Sea urchins, for example, are used in studies of embryology; nonhu-

By courtesy of (right) the U.S. Department of Agriculture, Soil Conservation Service; photograph (left) Grant Heilman

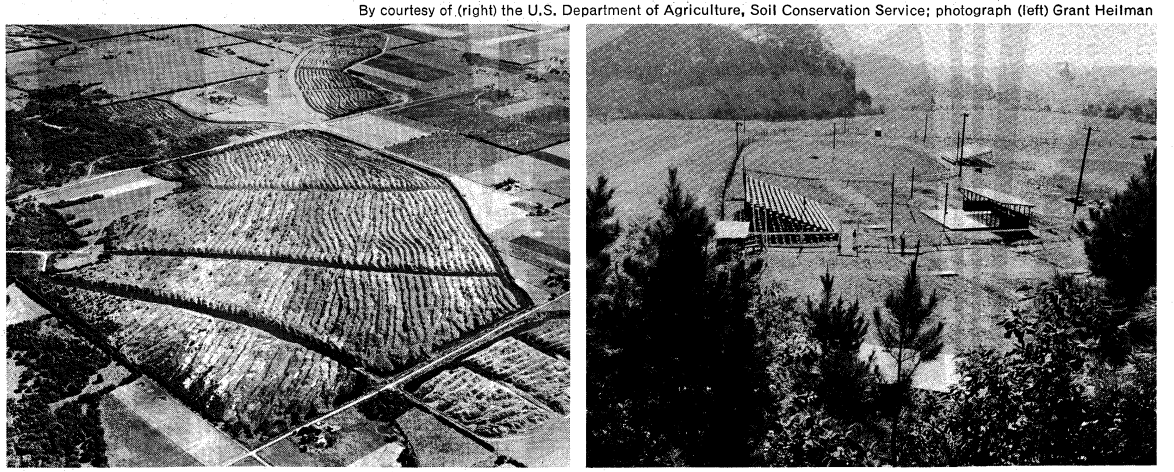


Figure 3: *Strip mining and its consequences.*
(Left) Strip-mining coal in Missouri. (Right) Wise County Fair Grounds constructed on coal strip mine spoil in Virginia.

man primates (*e.g.*, monkeys and apes) are used for many studies of human functions and diseases; and a great variety of wild plants are used as sources of drugs and medicines. Much of man's knowledge about the growth of populations and social behaviour under various conditions of crowding has come from the study of wild mammals. Furthermore, it is known that more or less undisturbed natural communities are important to the continued operation of those systems that man has created. Watershed forests are protected so as to maintain streamflow and to avoid siltation of reservoirs; estuaries are protected so as to guarantee the continued production of forms of marine life important to man for food or other purposes. Finally, there are aesthetic and recreational values attached to wild areas and wildlife. It would appear that outdoor activities in a natural setting or contact with plants and animals in a wild state are important to man's psychological well-being, because people of all races and cultures seek such experiences when they achieve the affluence that enables them to do so.

Protection of natural communities

Biological communities can be protected in a variety of ways, depending upon the desired objectives. The most difficult and exacting task is the protection of unmodified natural communities, with their full array of wild species, for use in scientific research. Because such communities are becoming increasingly rare, the major efforts to protect them are undertaken at an international level, as well as at the national level. The International Biological Programme, a worldwide research effort, has focussed attention on the many kinds of natural communities that require protection. The International Union for the Conservation of Nature, a semigovernmental international agency, devotes an important part of its activities to the establishment of reserves and parks for the protection of natural communities. The United Nations, through its Food and Agricultural Organization and the UN Educational, Scientific and Cultural Organization, has contributed to the establishment of many parks and reserves in developing nations. Yet, despite such activities, certain kinds of natural communities will be irrevocably lost unless there is greater effort toward their conservation.

The danger to natural communities and wild species comes from many causes. One long-standing problem derives from the exploitation of wild species that have commercial value, as, for example, the widespread removal of mahogany trees and other valuable timber from the forests around the Caribbean area. The uncontrolled hunting of whales and other sea mammals has brought some species to the point of near-extinction, and at least one, the Steller sea cow, has been exterminated. The demand for such high-fashion products as shoes and handbags made from crocodile skins or coats and wraps made from the skins of tigers and leopards has endangered the continuing existence of these animals.

Control of those species considered detrimental to man's welfare has led to an unreasonable warfare against predatory animals. As a result, the wolf, cougar, lion, lynx, eagles and hawks of various species, and other carnivorous animals have been eliminated from the vicinity of human settlements or from pastoral lands, even though the evidence that some of these species do any appreciable damage is not well substantiated. But the greatest single cause for the depletion of natural communities and wild species has been the desire to use land for more productive purposes. This has led to extensive clearing of forests and woodlands, burning of vegetation, and the cultivation of previously undisturbed land for crop production. Many of the lands that are cleared eventually prove to be poorly suited for the purposes for which they are intended and are ultimately abandoned; however, it is then no longer possible to protect the natural communities that previously existed in these areas.

Programs for the protection of natural communities must involve, first, rational planning for the use of land and control over its exploitation by agencies charged with such responsibilities. Adaptation of land use for commodity production by using only those sites best suited for such purposes is a first step toward protecting other lands that are now being cleared unwisely. Although the establishment and proper management of parks and reserves permits the survival of certain species in certain areas, a more general program of rational management and use of all lands and species is essential to the long-term survival of wild nature throughout the world.

Strict nature reserves. The decision to maintain an area in a more or less unmodified condition usually is determined by its overall scientific and aesthetic value. It may also be determined by the contributions the protected area can make to the region as a whole, as, for example, the regulation of water yield and streamflow or perhaps as a reservoir of wild species that subsequently can move or be moved into other modified regions. The most restrictive category of land use is that of the scientific reserve, which is also known as a strict nature reserve. Such areas may be selected because of their unique geological or biological features. Or they may be selected because they are representative of widespread biological communities that will be transformed elsewhere for the needs of commodity production, recreational use, or other more intensive purposes. Retaining representative areas in an unmodified condition provides standards by which the health and productivity of the modified sites can be tested.

Strict nature reserves and other types of scientific reserves often occupy only a relatively small area of land. The amount of land occupied, however, depends to a large extent on the biological requirements of the species to be protected and the interactions of the area with the surrounding region. Thus, in the rolling taiga country (swampy coniferous forests) of northern Europe, Asia, and North America, the decision to maintain a bog in a natural state may involve the protection of only a small area—the bog and those immediately adjacent slopes that drain into it. On the other hand, the decision to protect the natural habitat and population of migratory caribou in the same region could involve an area several hundred miles in length and more than 100 miles in width—a sizable area that might include taiga and tundra. The decision to protect the natural habitat of a species of migratory waterfowl would be even more far-reaching. In addition to protecting the tundra breeding ground of such birds, protection of their resting and wintering grounds, which could lie in a different hemisphere, might also be involved. It is always easier to protect rooted plants than mobile animals. And the more mobile the species, the more difficult the task; for many migratory species, international action is required.

Difficulty in protecting migratory species

Any decision to protect an area in an undisturbed condition, however, must take ecological reality into account. The experience of Everglades National Park in the United States, which protects only the lower end of an extensive watershed, is illustrative. The stormy history of this park has involved efforts to control water and land use in areas outside the park because its future existence depends upon the flow and quality of water from these areas. By contrast, high mountain reserves, such as those surrounding Mount Kinabalu, in Borneo, Mount Kenya, in East Africa, or Glacier National Park, in Montana, offer no such difficulties.

The designation of an area as a strict nature reserve is often proclaimed by law and recognized only by the appropriate governmental authorities, with no cognizance of the action by others. Such areas must have boundaries that are clearly demarcated and identifiable to prevent accidental intrusion and modification.

Because a strict nature reserve is set aside for scientific purposes, its use for recreation or any other purpose may disturb its natural integrity. Even scientific use itself can be a disturbing factor; hence, the use of such reserves is regulated by scientists. Usually a scientific advisory committee must rule on the appropriateness of any proposed research to the long-term future of the area.

The decision to protect a natural area for scientific purposes is not a simple one. In the case of most of the less affluent countries, in which neither money nor technical expertise is available, it is virtually impossible to proclaim a strict reserve and have it maintained as such except in the most remote areas. Usually international as-

sistance is required, in money or manpower, if such reserves are to be established.

National parks. The establishment of national parks in the United States represented one of the first national efforts to protect wild nature. Yet, in establishing Yellowstone National Park, Congress made clear that it was viewed as "a pleasuring ground" for people and not as an area intended only to safeguard communities of plants and animals. It was not until the formation of the U.S. National Park Service in 1916 that the concept of managing parks so as to maintain their natural qualities was accepted. Nevertheless, the practice of killing predatory animals as "undesirable" elements of wild nature continued in U.S. national parks into the 1930s and lasted in some African national parks as late as the 1960s.

Unlike a strict nature reserve, a national park may be made available for various purposes but usually only for those forms of recreational use that do not create great changes in or require significant modifications of the natural environment. National parks usually are selected on the basis of their unique qualities, outstanding natural beauty, unusual geological formations, or remarkable array of wild animal or plant life. They may also be selected, however, to protect areas of anthropological, archaeological, or historical importance along with the natural or artificially modified landscapes that surround them. In the United States, national parks are dedicated solely to recreational activity. National parks in England may protect cultural as well as natural landscapes, in that some may be dedicated to the preservation of traditional forms of land use that are disappearing elsewhere. Some national parks, such as in Peru, protect primitive peoples along with their hunting and gathering grounds.

Thus, exactly what constitutes a national park varies according to the nations and people involved. The dedication of an area as a national park is everywhere a highly restrictive form of land use, in which all incompatible activities are prohibited. Hunting, logging, mining, commercial fishing, agriculture, and livestock grazing are excluded from most such parks, as are urban and industrial uses not directly related to recreation. There is much debate as to whether tourist facilities should be within or outside national parks; because of their disruptive effects, the trend is to locate such facilities outside.

Facilities needed in national parks

National parks, at a minimum, require equally extensive boundary demarcation and perhaps policing and patrolling as are necessary for strict nature reserves. They also require the careful planning of trails, roads, and other means of human access in order to channel the activities of visitors in ways that will not disrupt the resources or landscapes. Not only must certain fragile areas be set aside and protected from visitors, but visitor use must be concentrated in those places in which human activities will do a minimum of harm. The trend has been to divide national parks into zones that range from areas of intensive public use at one extreme to the most remote wilderness or strict nature reserves at the other.

Usually a considerable amount of money and manpower must be invested in the planning and management of a national park. This is often beyond the resources of the less affluent countries unless international assistance can be provided. Because of their attraction as sites for outdoor recreation and their appeal to tourists, however, national parks often more than pay for themselves in a short period of time. In East Africa, for example, national parks are the major source of foreign exchange of the countries in which they are located because of their unique wild animal life. As a result of the body of expertise that has developed in the planning and management of national parks and because of their growing economic importance, expert direction in the establishment and maintenance of national parks is now available throughout the world whenever it is requested. Within the United Nations such assistance is offered to developing countries by UNESCO and the Food and Agricultural Organization; outside the United Nations, assistance is available from the International Union for Conservation of Nature and from the World Wildlife Fund.

Refuges and sanctuaries. A less restrictive use of an area of wild land is that directed toward the protection of certain species or groups of wild animals or plants. This type of land use includes wildlife refuges and sanctuaries as well as various kinds of botanical reservations. In such areas, use of the land that could interfere with the well-being of the protected species is excluded. Other forms of land use may be encouraged, however, if they are not in conflict or if they actually assist with the creation of a suitable environment for the protected forms of life.

Refuges and sanctuaries often are established for the preservation of endangered species of wildlife or plants, particularly those whose numbers and distribution have been seriously curtailed; examples include the Umfolozi Game Reserve, in South Africa, in which the southern species of the white rhinoceros is protected, and the Mountain Zebra National Park, in the same country. The refuges for the California condor and the Torrey pine in California are of a similar nature. Refuges and sanctuaries may also be provided for more abundant species that require protection at certain periods of their life cycle or in certain areas where they gather in order to reproduce. Many wildlife refuges and sanctuaries in the United States and Europe are of this type; they provide protected sites for the resting, breeding, or wintering of wildlife species (particularly waterfowl) that otherwise are hunted outside of the refuge. In such places management measures are necessary to enhance the habitat for the protected species and to remove any competitors or predators that might interfere with the breeding and survival of the young. Such measures would be inappropriate to parks or strict nature reserves, because their purpose is to protect a total biological community without favouring a particular species.

Refuges as protected sites for abundant wildlife species

The various measures employed for the protection and management of living resources on land also apply to water areas. Marine and other aquatic parks and reserves have been established in many parts of the world to protect various forms of saltwater and freshwater plant and animal life. Australia, for example, has reserves that protect important areas of the Great Barrier Reef; Kenya and Tanzania have marine parks and reserves on their coasts. Lake Baikal, in the Soviet Union, is now included in a major national park designed to protect not only its unique freshwater life but also its watershed areas.

Wildlife and fisheries management. The protection of most wild animal life cannot be accomplished solely through parks and reserves. Consequently, wildlife conservation and management represent an activity that extends into other areas. As has been mentioned previously, one form of wildlife conservation with a long history of laws, regulations, and proclamations dating back to ancient centres of civilization was the protection of royal game. Wildlife conservation is also closely related to the management of fisheries, because both are directed toward the preservation of wild species and their habitats and toward increasing the productivity and yield of these species. Increased productivity in wildlife or fish may contribute to a commercial harvest, or it may make contributions to human well-being through sport hunting or fishing, wildlife viewing, or some other form of recreation.

The basis for wildlife and fishery management includes research into the ecological requirements and breeding potentials of the species involved. Protective laws and regulations are necessary to control the allowable commercial or sport take in order to guarantee that it remains within the sustainable yield. Means for enforcing such laws should include the employment and deployment of adequate protective forces, preferably specially trained wildlife and fisheries wardens who are capable of identifying species likely to be killed or captured and who are familiar with the ways of hunters and fishermen. Management activities intended to maintain or improve the habitat for the wild animals concerned are also important. Special refuges of the kind already described may be essential both to control hunting and fishing pressure and to allow animals to reproduce, rest, or winter

without being disturbed. In addition to these activities, wildlife conservation also involves the management of the requirements needed by wildlife in the areas in which the land is used for other purposes.

Commercial value of fish and wildlife

Commercial fishing, which is of great economic importance in countries with extensive inland waters or with access to the seas and oceans, contributes a significant share of world food, particularly protein.

It has been estimated that ocean fisheries yield more than 55,000,000 tons of fish to commercial markets, and future sustained yields in excess of 110,000,000 tons have been projected. The problems of fisheries conservation are many, however. It is essential first to locate the fish and derive some estimate of their abundance. Much of the recent increase in yield by world fisheries has come from the discovery of previously little known and unexploited sources of fish. It is also necessary to determine the maximum sustained yield of the fish population. Finally, the most difficult task is the supervision and control of the fishery. Even in national waters, fisheries have been overexploited and depleted to a level such that the species concerned no longer has any commercial value—the sardine fishery of the California coast is an example. In international water it is much more difficult to regulate and control the activities of fishermen.

The commercial use of terrestrial wildlife, although not in the same economic category as fisheries, is nevertheless significant to some countries. Wild animals are killed for their meat, hides, furs, fats, oils, bones, ivory, antlers, and other by-products. Live animals are captured for zoos, for medical research, and for the pet trade. The exceedingly high prices paid for certain species, such as various primates used for medical research or those that are valued for their furs, has been of great importance in causing the reduction of some species to a dangerously low level. Despite the high commercial value of some wildlife, however, in many parts of the world their recreational or sporting value has an even greater economic importance. In the United States, for example, the commercial hunting and sale of game animals is generally prohibited, as is the commercial harvesting of certain game fish, because such wildlife is reserved for recreational use.

In technologically advanced countries it has proved practical and economically desirable to manage wildlife and fisheries so that no species becomes severely depleted. But, in those countries in which a high percentage of the people are illiterate or live near a poverty level, the protection and management of wildlife and fisheries is much more difficult to achieve. In such nations protective laws are not easy to enforce because trained wardens are not widely available and because the laws are neither understood nor accepted by the general population. Thus, the depletion of wildlife and fisheries through hunting and fishing continues in these countries, and growing numbers of species are now endangered.

Multiple-use management. The use of an area for wildlife conservation may include its use for other purposes. Throughout the world many areas of public wild land are managed on a policy of multiple use; that is, they can provide either at the same time or sequentially a variety of wildland products or services. National forests, for example, are used for recreation and to produce timber and range forage for domestic livestock, wildlife, and fish; and at the same time they stabilize watersheds and yield sustaining water supplies to natural bodies of water or to man-made reservoirs.

The decision to manage an area for multiple use, however, necessitates minimization of conflicts between the various forms of use. Thus, removing the timber from an area involves the construction of logging roads, landings, and other facilities; and, while the area is being logged, most other uses are restricted. Furthermore, during and following logging, care is taken not only to provide a site and seed trees suitable for growing the next crop of trees but also to protect the natural reproduction of trees, young seedlings, and saplings. In cases involving more intensive management, seeds or seedlings are planted and given protection during the period when they are becoming established. Moreover, the value of the site for recreation or for other purposes may be greatly reduced during this process, particularly when certain systems are used for cutting timber; thus, clear-cutting, which involves the total removal of all large trees from an area, creates an appearance of greater devastation than is seen in selective cutting, in which only a few trees of a particular species and age class are removed. Obviously, clear-cutting then favours wildlife species that prefer open areas at the expense of those that require dense forest cover. Generally speaking, however, logging of timber involves the exclusive use of an area for this purpose over a period of time; other uses may not be fully restored until the vegetation or animal life has recovered from the effects of timber removal.

Conflicts in multiple use

(Left) Ray Atkeson, (right) Josef Muench

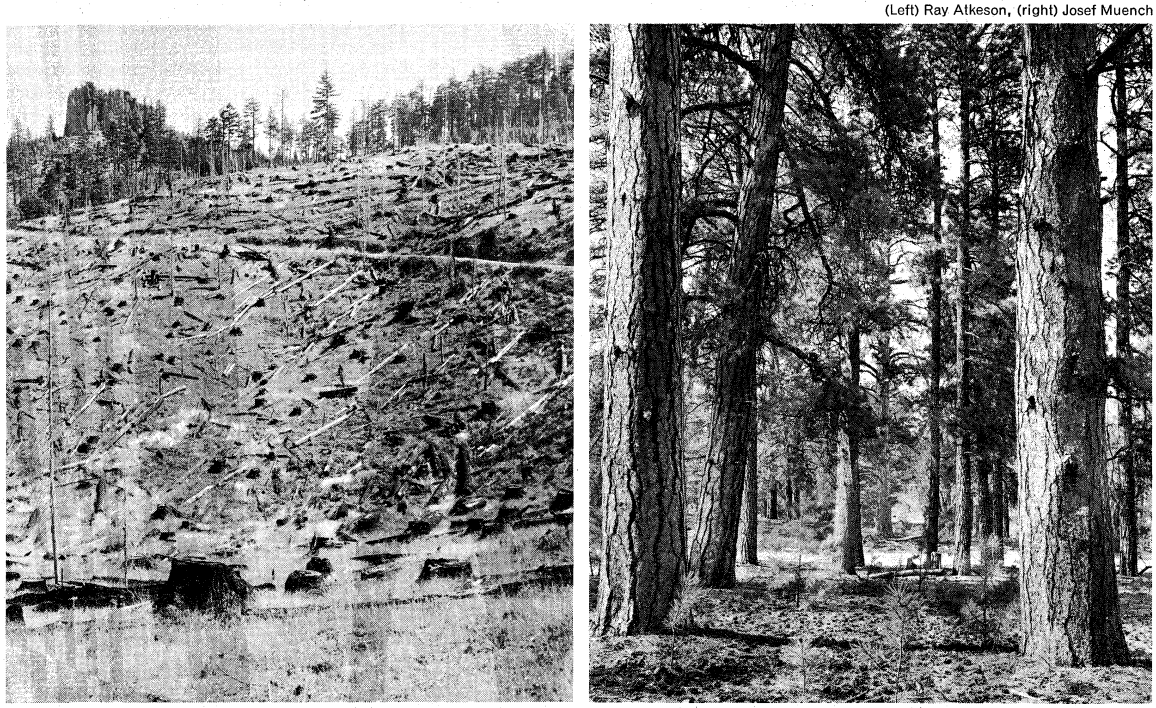


Figure 4: *The cutting of timber.*
(Left) Clear-cut timber in Oregon. (Right) Selective cutting of timber in the Chuska Mountains, New Mexico.

Development of an area for grazing by domestic livestock often includes fencing; the demarcation of livestock trails for the transfer of animals from one area to another; the provision of watering points, salt grounds, and other items essential for the welfare of the animals; and sometimes the construction of special corrals or handling chutes. When livestock are in an area, they cannot be disturbed too frequently. Livestock owners also demand the removal of any predatory animals that are likely to attack their stock. Such needs obviously restrict the development of the same site for purposes that could conflict with its use as a livestock range.

Factors involved in development for water power

Intensive development of an area for water power (*e.g.*, the construction of dams, power stations, pipe lines, canals, and power lines) entails complete change of the area affected and removes much of its wild or natural quality. Although such a location is obviously less suited for wilderness or wild-country recreation, it often is enhanced for certain forms of mass recreation, such as fishing, boating, and water sports in reservoirs. Recreational development on any intensive scale also involves considerable modification of an area. The construction of tourist and visitor facilities of various kinds—trails, ski runs, campsites, roads and parking areas, lodges, and picnic grounds—necessarily restricts the use of such an area for other purposes.

Harvey Caplin

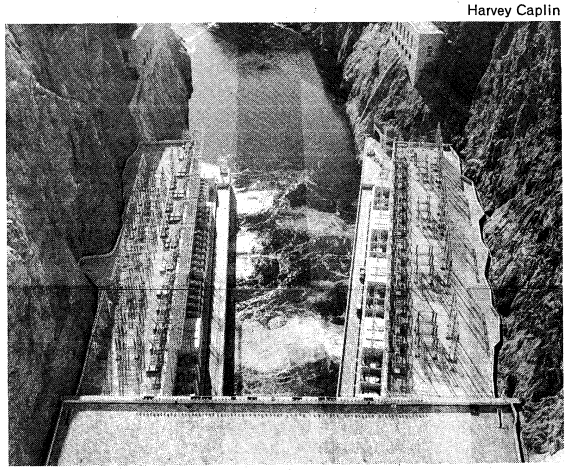


Figure 5: Hoover Dam, Colorado River, on the border of Arizona and Nevada, used for irrigation and hydroelectric power.

The concept of multiple use thus applies mostly to the management of large areas of wild land, although within such areas various special uses may from place to place and from time to time be emphasized. Nevertheless, the governing policy of multiple use is valuable because it prohibits the exclusive use of a particular area for the benefit of any one segment of society and, instead, forces accommodation to the greatest variety of uses that best suit a particular site.

Intensive wild-land uses. *Timber production.* Loss of the natural character of an area usually accompanies the decision to manage it for maximum production of such crops as timber or range forage. Thus, if a forested area is to be managed for maximum timber production, establishment of fast-growing species of trees that will make the most efficient use of the soils, water, and climate is necessary. Crops of timber or other wood products are cut and removed regularly and other trees planted in their place, usually from nursery stock. Roads and other facilities are constructed to facilitate the removal or protection of the forest products. The ultimate objective is usually a high-yielding plantation or tree farm, often using exotic species of trees or specially developed varieties that combine high-yielding capabilities with resistance to insect pests or diseases. Seldom do such forests have any value for other purposes, except possibly for soil protection and water yield. Aesthetically, such a forest may even lose its wild aspect and have all the regularity and artificiality of an orchard or cornfield. Recreational uses often are discouraged; even if they are not, the recreational appeal of such areas usually is slight. Few if any of the larger forms of wild animal life can find a suitable habitat in a forest plantation, and those that do become established often are regarded as pests if they interfere in any way with tree production. Such forests, however, produce far more timber, pulp, and other kinds of wood products than can be obtained from an equivalent area of wild forest. Moreover, intensive development of production forests in suitable locations can spare areas of natural forest for other purposes. Privately owned forests in North America are increasingly managed in this way, a practice that increases with the demand for forest products; government-owned production forests and forest plantations in other countries receive similar treatment. Such intensive development, however, is economically feasible only at sites with soils and other habitat conditions favourable to high yields.

Livestock production. Management of range and pasture lands for livestock production can become almost a single-purpose use of the land. A range, any area of wild land that is covered with natural vegetation, is usually located in the more arid mountainous or colder regions of a country; such land is suitable for the production of the larger grazing or browsing animals. Pastures, on the other hand, are grazing areas with better growing conditions than ranges because they are located either in regions of higher rainfall or in areas suitable for irrigation. Because of the nature of pasture, more intensive management is economically practical, including regular cultivation and planting to replace the natural vegetation with higher yielding domestic, exotic, or artificially bred varieties of forage plants. Livestock production on pastures is a single-purpose use of the land, although pastures are often included as part of a rotation system in the raising of farm crops. The intensive use and management of pastures effectively removes such areas from the category of wild land.

Range and pasture as single-purpose use of land

Management of the forage on range lands requires skillful manipulation of natural processes of plant succession and the balancing of livestock numbers to the carrying capacity of the vegetation; that is, the number of animals the vegetation can support without deteriorating and endangering its future productivity. Efforts are made to avoid overgrazing, which reduces the vegetation to a less productive state, and excessive trampling by animals, which can compact the soil and interfere with its water relationships and productivity. A flexible method of livestock handling is essential to success on ranges: animals are not allowed to concentrate excessively in any one area, and their use of vegetation must be consistent with continued production of the more nutritious and palatable forage plants. Moreover, during dry periods, when there is low forage production, the number of livestock using a range must be reduced in order to avoid damage to the range; otherwise, feed lots are provided or the animals are moved to irrigated pastures.

As a primary use of range land, livestock production requires the construction of various facilities—fences, water holes, corrals, roads, and troughs for salt or other forage supplements. On productive range lands it may be economically profitable to provide fertilizers, usually by spraying or dusting from aircraft, to encourage higher forage yields. Undesirable vegetation is sometimes burned, crushed, or sprayed with herbicide to provide more space for the better forage plants; ranges also may be seeded with higher yielding or more disease-resistant forage plants. Such activities necessarily restrict use of the land for other purposes. Nevertheless, except in cases of intensive management, range lands can provide recreation and often are highly suited to various forms of wildlife management. There is considerable evidence from many parts of the world that some combination of domestic animals and wildlife can provide greater yields and profits from drier, colder, or steeper range lands than can be obtained from such lands when used exclusively for domestic-livestock production.

Agricultural management. The most intensive forms of rural land use for agricultural purposes are those con-



Figure 6: *Livestock production on range and pasture land.*
(Left) Rangeland in southwestern Washington. (Right) Angus cattle on bluegrass pasture land in Kentucky.
(Left) Ray Atkeson, (right) J.C. Allen and Son

cerned with the raising of harvestable crops or with the production of animal products. Unlike primitive agriculture, which involved only the temporary removal of natural vegetation and depended for a short period of time on natural soil fertility, modern agriculture requires large inputs of chemicals, energy, and technical skills to produce increased yields of crops or animals. In the technologically advanced countries food production is now so much greater than population growth that it is possible to retire former farmlands from use and to produce crops acording to demand without approaching the maximum yields obtainable. The so-called Green Revolution of recent years has been based on the spread of improved farming skills to less developed nations of the world. It has been made possible by the breeding of high-yielding forms of grain specifically adapted to the ecological conditions of the countries involved.

Green Revolution

The decision to use an area of land for high-yield agriculture essentially rules out its use for other purposes. The intensive production of farm crops in an agricultural region may also have undesirable side effects; as has been previously noted, these may include the pollution of other areas when the pesticides, herbicides, or other agricultural chemicals blown or washed from farmlands affect vegetation and animal life elsewhere. Nevertheless, committing an area to intensive agricultural production does not rule out its future restoration for other uses. As long as the soils are well cared for, such areas can be converted quickly to other purposes if it is not necessary to keep them in farm production. Abandoned farmlands in the southern United States, for example, are now highly productive forest areas, and former farming lands elsewhere are being used to support wildlife and outdoor recreation. In general planning for conservation of natural resources, intensive use and high production in those areas best suited for farming must be encouraged—provided, of course, that the polluting effects of these activities on the general environment are avoided. Such concentration can spare the destruction of other resources through attempts to use inadequate lands for marginal farming activities.

INTERNATIONAL PROBLEMS OF RESOURCE MANAGEMENT

The management of living resources requires a high degree of international cooperation and a willingness on the part of nations to agree to some forms of international control. This is particularly true for the management of those aquatic animals that occupy international waters; it is equally true for migratory animals that move from one country to another. Furthermore, the animals and plants that are commodities in international trade must also be protected through international agreements. Finally, when an international agency or an agency from a particular country works with the resources of a less developed nation in an effort to help that nation improve its economy, care is taken to avoid any adverse effects on the conservation of that nation's environment.

Despite general recognition of these problems, few nations have shown a willingness to forgo any of their sovereign rights or to cede authority over their affairs to international bodies. International laws and regulations remain, for the most part, gentlemen's agreements among nations. A country can and often does ignore the rules when it is economically advantageous for it to do so. Only international public opinion or, more rarely, the threat of force by one nation or by a group of nations can serve as an effective deterrent in compelling a country to stop certain activities that are endangering a resource. Yet, despite the generally chaotic state of international rules governing the management of environmental resources, there are some excellent international agreements concerning this matter. Such agreements provide hope that the general field of conservation may serve as an area in which nations can learn to work together more effectively for their mutual benefit.

Problems in enforcing international conservation regulations

Effective international agreements. *The Antarctic Treaty.* The Antarctic continent remained unknown and unexplored during the period when most other parts of the world were being claimed and colonized by European powers. Uninhabited and, until recently, of little or no value to any country, Antarctica never became an area for international dispute. Although segments of the continent and sub-Antarctic islands have been claimed by various nations, including the United States, the Soviet Union, Great Britain, Australia, New Zealand, Chile, and Argentina, the validity of these claims has never been tested. It was not until the latter half of the 20th century that the potential future importance of Antarctica and the knowledge to be gained from scientific research in and exploration of the continent were recognized. As a consequence, nations with an interest in Antarctica signed the Antarctic Treaty in 1961. This could be considered de facto establishment of the Antarctic continent as the world's largest strict nature reserve. The treaty provided for joint scientific research to be conducted by the signatory nations, with the results to be shared by all. It also prohibited the exploitation by any one nation of living Antarctic resources as well as any activities that would cause deterioration of those resources. The treaty has been respected, Antarctic research has gone forward, and the wild animals and sparse plant life of the continent have been protected.

Whaling agreement. Unfortunately, agreements covering the Antarctic have not included the oceans surrounding the continent. The value of the animal resources in these waters has been known since the 18th century, and exploitation of them was begun during the 19th century. Of particular importance were the huge herds of whales that congregated in the Antarctic seas to feed upon the abundant plankton in these nutrient-rich waters. (One component of this plankton is the krill, a shrimplike crustacean; it occurs in such large quantities that, according to calculations, its exploitation alone, if properly conducted, could have increased the 1970 world yield of ocean fisheries by more than half.) The early whaling ships concentrated their efforts on the larger

whales, which could yield the greatest quantities of whale oil, whalebone, and other useful products. The most sought after of these creatures, the blue whale, was the world's largest mammal. It is estimated that when whaling began there were 200,000 blue whales in the Antarctic; by 1965, however, the population of blues had been reduced to about 2,000. Preceding the period of concentration on blue whales, whalers had pursued the various species of right whales in the world's oceans and had reduced their numbers to a similar low level. The gray whale of the Pacific was also brought to a point of near-extinction. The efforts of Antarctic whalers then shifted to smaller species; the humpback, once abundant, was reduced to an estimated 1,000 in 1962. Fin whales and sperm whales also bore the brunt of whaling pressure for a time; when they become scarce, the still smaller sei and Minke whales were killed.

Slaughter of blues and other whales

Despite the efforts of conservationists, attempts to restrict the activities of whalers met with little initial success. As whale populations dwindled and the danger of extinction became evident, however, the principal whaling nations agreed to sign the International Whaling Convention in 1946. This led to the establishment of the International Whaling Commission, which was authorized to sponsor scientific studies of whale populations and to recommend to the whaling nations limitations on harvest that were necessary to perpetuate whale populations and the whaling industry. Because the nations involved did not give the commission any firm power or enforcement authority, any nation could dispute the recommendations of the commission's scientific advisers and insist upon higher quotas than had been recommended. Everything depended on the willingness of the nations to obey the rules and to report honestly the number of each species of whales taken during the whaling season. Although there is no evidence that the nations involved deliberately disobeyed the recommendations finally agreed upon by the commission or that they failed to provide the commission with anything but accurate figures, the number of whales continued to decline. Sustained pressure from conservation interests, however, finally accomplished results. After 1967 the endangered whales—the blue, right, gray, and humpback—were given complete protection. And, starting in the 1970s, quotas for other species were reduced to limits that enabled the whales to maintain their populations at certain levels. Thus, the record of international cooperation provided by the International Whaling Convention is a mixed one. Although it did not accomplish all of the desired results, neither did it fail entirely. Indeed, the record of whale conservation would be much worse had it not been for this international agreement.

Protection of fur seals and migratory birds. Other activities intended to control the international exploitation of environmental resources have had some successes and many failures. Among the most successful treaties has been one to protect the northern fur seal, a species that breeds in the Pribilof Islands, in the Bering Sea. During the 19th century fur seals were reduced to a dangerously low level as a result of the heavy slaughter at their breeding grounds to obtain skins for the manufacture of fur coats and various other sealskin products. In 1911, Canada, Japan, Russia, and the United States signed a treaty to limit the annual harvest of seals to a quantity that would not only sustain the population but also increase it annually. In addition, the profits from sealing were to be divided proportionately among the signatory nations. The treaty was adhered to; the seals have increased and now stock the available breeding grounds.

The fur seal treaties

A good example of an international agreement governing the conservation of species that migrate between two or more nations is the International Migratory Bird Treaty. Established in 1918 between the United States and Canada, this treaty was subsequently extended to include Mexico. The treaty, which has been adhered to by the nations concerned, limits the kill of migratory waterfowl. It also provides for the protection of migratory species in their breeding grounds, along their migration routes, and at their wintering areas.

Territorial limits and marine resources. For the most part, however, international control of species occupying habitats within several national jurisdictions, particularly when the habitats are in international waters, continues to be a serious problem. Agreements have yet to be reached on what constitutes national and international waters. The traditional three-mile territorial limit was loosely agreed upon by coastal nations for military reasons; it evolved in the days when that distance was beyond the range of shells fired from guns aboard naval vessels offshore. Today, however, this limit has become useless for defense and is not widely recognized for any other purpose. Disputes over fishing rights have led to a more general acceptance of a 12-mile (19-kilometre) limit to the seaward boundaries of national territories, but even this is not acceptable to all nations. Many Latin American countries now insist on jurisdiction over waters to a distance of 200 miles (300 kilometres) from their shores.

In 1970 it was proposed that resources lying on or beneath the floor of the continental shelf (the zone of shallow ocean floor that gradually slopes from the seashore to an abrupt drop-off) be under national jurisdiction out to a point at which the mean water depth is 200 metres (650 feet). Beyond that depth, which generally marks the limit of the continental shelf, the resources of the continental slope down to the floor of the ocean would be internationally owned but under the management and control of the coastal state involved. National exploitation of these resources would be permissible but with a payment of royalties to an international body. Farther out at sea, exploitation of the resources on and under the floors of the ocean basin would be entirely governed by an international agency. The proposal, which has not yet been generally accepted, is intended primarily to govern the exploitation of mineral and fuel resources as well as other uses of the ocean floor. It does not propose a change in jurisdiction over the ocean waters or their fisheries resources.

The problems concerning the conservation of marine resources have already received much international attention, and they probably will receive more both in meetings held under the auspices of the United Nations and in those conducted by other international agencies. It is apparent, however, that, until nations are willing to agree to international regulation and to provide the means for enforcing such regulation, there can be no great hope for solving those problems concerned with the rational use and conservation of international marine resources.

International trade in animals and plants. Special problems are involved in the regulation of trade and commerce in living animals and plants as well as in the products derived from them. Demands by wealthy nations for certain animal and plant products create particularly severe problems in less affluent countries. As mentioned previously, the trade in endangered species of wildlife is illustrative. The demand for furs and skins of rare animal species is artificially created in the fashion centres of the world. Prices paid by wealthy people for these items in affluent countries exceed the lifetime income of most people in the countries from which the leopards, crocodiles, tigers, and other wild species come. Poachers go to great lengths to obtain these animals wherever they can be found, including inside national parks and reserves. Because effective policing is virtually impossible, legal and illegal trade in wildlife begin to overlap, and both become firmly established. Exporters of wild animals and their products are the end links of profitable business chains that include far greater numbers of hunters and trappers in remote areas. Furthermore, for each animal or skin that reaches a foreign market, many more are destroyed in hunting, trapping, and transporting.

Furs and skins for the fashion market

Only the purchasing countries can control the illegal trade in wildlife. This can be done in several ways: elimination of the demand created by changing fashions; refusal of purchasers of live animals (*e.g.*, zoo directors) to buy species that are endangered; or, least effectively,

by controlling their clearance through customs offices of international airports and seaports. In 1971 the International Union for Conservation of Nature prepared for consideration by governments a draft convention intended to accomplish some degree of control. Adopted by numerous countries in following years, this convention prohibits trade in endangered wild animal species or their products. Its provisions can be extended to include endangered plants (such as rare orchids that are now being removed from even the most remote tropical forests), fish, shellfish, and other invertebrates.

Perhaps even more serious than the trade in wild animal species is the international exploitation of other living resources, particularly the tropical forests of the world. These forests, which contain many hundreds of species of trees growing in diverse mixtures, were spared from exploitation in earlier decades because of their inaccessibility, the relatively low value of most of the trees for timber purposes, and the limited world demand. Heavily exploited for special uses were a few species of high value, such as teak, ebony, sandalwood, mahogany, and other furniture woods. Most tropical forests were not greatly disturbed, however. This situation has changed, and a wide variety of woods previously considered worthless are used for pulp, chipboard, and fibreboard or as cellulose for plastic production. With new machines and better transportation, it has become profitable to remove trees from previously remote areas and to ship logs, bolts, wood chips, or other partially processed materials to foreign markets. Faced with a high demand for their forest products, most developing nations have been willing to sign over timber rights to foreign companies, hoping thereby to increase their national incomes and to advance the general material welfare of their people. Unfortunately, most of these timber contracts contain few or no provisions for conservation. Forest industries that have excellent management and conservation records in their home countries behave differently in other lands. Great areas of tropical forest have been laid waste, soils bared to erosion, and the wildlife within them destroyed. Because no laws are violated in either the exploited or the home country, there is no effective redress. General international agreements governing the conservation of such living resources would provide an answer to this problem, but they are unlikely to be implemented in time to prevent the devastation of large areas of the tropical world.

Unintended side effects of international assistance

Equally if not more unfortunate have been the side effects of some well-intentioned international development projects. These are sometimes sponsored by international agencies concerned with such affairs and sometimes by the foreign-assistance departments of individual donor nations. Usually the projects are intended to benefit one segment of the economy of the recipient nation; but, because ecological advice generally is not sought and because of the broad impact of the proposed development on other resources or on the total environment, the side effects of some of these activities often far outweigh any benefits that are derived. An example is the Aswān High Dam of Egypt, where the need to increase the supply of water for irrigation and power was considered paramount. The environmental side effects, however, have been enormous and include the spread of the disease bilharzia (schistosomiasis) by snails that live in the irrigation channels, loss of land in the delta of the Nile River from erosion once the former sediment load of the river was no longer available for land building, and a variety of other consequences. Furthermore, water-development projects in the semi-arid lands of Africa, sponsored by international agencies and funded by various donor organizations, have quite frequently been far more destructive than constructive. Failure to take ecological and social factors into consideration has resulted repeatedly in overgrazing and desertification in an area intended to be improved by the expenditure of money. The responsibility of agencies concerned with international development to seek the best environmental advice is now generally accepted; implementation of this responsibility has been slow, however.

THE POLLUTION OF NATURAL RESOURCES

Although various problems related to pollution of the environment have already been mentioned, the following is a more general discussion of pollution as a phenomenon. Pollution may be defined as the addition of any substance or form of energy (*e.g.*, heat, sound, radioactivity) to the environment at a rate faster than the environment can accommodate it by dispersion, breakdown, recycling, or storage in some harmless form. A pollutant need not be harmful in itself. Carbon dioxide, for example, is a normal component of the atmosphere and a by-product of respiration that is found in all animal tissues; yet in a concentrated form it can kill animals. Human sewage can be a useful fertilizer, but when concentrated too highly it becomes a serious pollutant, menacing health and causing the depletion of oxygen in bodies of water. By contrast, radioactivity in any quantity is harmful to life, despite the fact that it occurs normally in the environment as so-called background radiation.

Pollution has accompanied mankind ever since groups of people first congregated and remained for a long time in any one place. Settlements of primitive man can be recognized by their pollutants—shell mounds and rubble heaps. But pollution was not a serious problem as long as there was enough space available for each individual or group. With the establishment of permanent human settlements by great numbers of people, however, pollution became a problem and has remained one ever since. Cities of ancient times were often noxious places, fouled by human wastes and debris, and, in the Middle Ages, unsanitary urban conditions favoured the outbreak of decimating epidemics. During the 19th century, water and air pollution and the accumulation of solid wastes were largely the problems of only a few large cities. But, with the rise of advanced technology and with the rapid spread of industrialization and the concomitant increase in human populations to unprecedented levels, pollution has become a universal problem.

The various kinds of pollution are most conveniently considered under three headings: air, water, and land. Air pollution involves the release into the atmosphere of gases, finely divided solids (particulates), or finely dispersed liquid aerosols at rates that exceed the capacity of the atmosphere to dissipate them or to dispose of them through incorporation into solid or liquid layers of the biosphere. The problems of air pollution have been discussed at length in a previous section (see above *Types of natural resources*).

Water pollution. Water pollution involves the release into lakes, streams, rivers, and oceans of substances that become dissolved or suspended in the water or deposited upon the bottom and accumulate to the extent that they interfere with the functioning of aquatic ecosystems. It may also include the release of energy in the form of radioactivity or heat, as in the case of thermal pollution. Any body of water has the capacity to absorb, break down, or recycle introduced materials. Under normal circumstances, inorganic substances are widely dispersed and have little or no effect on life within the bodies of water into which they are released; organic materials are broken down by bacteria or other organisms and converted into a form in which they are useful to aquatic life. But, if the capacity of a body of water to dissolve, disperse, or recycle is exceeded, all additional substances or forms of energy become pollutants. Thus, thermal pollution, which is usually caused by the discharge of water that has been used as a coolant in fossil-fueled or nuclear-power plants, can favour a diversity of aquatic life in waters that would otherwise be too cold. In warmer bodies of water, however, the addition of heat to a body of water changes its characteristics and may make it less suited to species that are considered desirable.

Thermal pollution of water

Pollution may begin as water moves through the air, if the air is polluted. Soil erosion adds silt as a pollutant. The use of chemical fertilizers, pesticides, or other materials on watershed lands is an additional factor con-

tributing to water pollution. The runoff from septic tanks and the outflow of manures from livestock feedlots along the watershed are sources of organic pollutants. Industries located along waterways downstream contribute a number of chemical pollutants, some of which are toxic if present in any concentration. Finally, cities and towns contribute their loads of sewage and other urban wastes. Thus, a community far upstream in a watershed may receive relatively clean water, whereas one further downstream receives a partly diluted mixture of urban, industrial, and rural wastes. The cost of cleaning and purifying this water for community use may be high, and the process may be only partially effective. To add to the problem, the cities and towns in the lower, or downstream, regions of the river basin contribute additional wastes that flow into estuaries, creating new pollution problems.

The output of industries, agriculture, and urban communities generally exceeds the biological capacities of aquatic systems, causing waters to become choked with an excess of organic substances and organisms to be poisoned by toxic materials. When organic matter exceeds the capacity of those micro-organisms in water that break it down and recycle it, the excess of nutrients in such matter encourage rapid growth, or blooms, of algae. When they die, the remains of the dead algae add further to the organic wastes already in the water; eventually, the water becomes deficient in oxygen. Anaerobic organisms (those that do not require oxygen to live) then attack the organic wastes, releasing gases such as methane and hydrogen sulfide, which are harmful to the oxygen-requiring (aerobic) forms of life. The result is a foul-smelling, waste-filled body of water, a situation that has already occurred in such places as Lake Erie and the Baltic Sea and is a growing problem in freshwater lakes of Europe and North America. The process by which a lake or any other body of water changes from a clean, clear condition—with a relatively low concentration of dissolved nutrients and a balanced aquatic community— to a nutrient-rich, algae-filled body and thence to an oxygen-deficient, waste-filled condition is known as accelerated eutrophication.

Eutrophication of bodies of water

Land pollution. Land pollution involves the deposition on land of solid wastes—*e.g.*, used cars, cans, bottles, plastic containers, paper—that cannot be broken down quickly or, in some instances, cannot be broken down at all by the action of organic or inorganic forces. (The term biodegradable is used to describe those materials that can be decomposed and recycled by biological action.) When such materials become concentrated within any one area, they interfere with organic life and create unsightly accumulations of trash. Methods of disposal other than recycling include ocean dumping, which creates water pollution and destroys marine habitats; landfill, which often requires the availability of low-lying ground and frequently involves the destruction of marshland or swamps that have high biological value; or burning, which increases air pollution. Obviously, none of these methods is entirely satisfactory, although using landfill to create artificial landscapes, which then are covered with soil and planted with various kinds of vegetation, is a possibility that remains to be fully developed. It is the great quantities of debris produced by urban communities, more so than a shortage of raw materials, that forces the development of more effective means for recycling wastes. Land pollution also involves the accumulation on land of substances in a dispersed solid or liquid form that are injurious to life. This has been particularly noticeable with those chemicals (*e.g.*, DDT) that are spread for the purpose of exterminating pests but then accumulate to the extent that they can do damage to many other forms of life.

Noise pollution. One form of pollution that is characteristic of industrial societies is noise. The intensity of sound is measured in logarithmic units known as decibels; a change from a level of ten decibels to one of 20 decibels actually represents a 100-fold increase in the sound level. At a level of 80 decibels, sound is annoying; permanent loss of hearing can be caused by steady exposure to noises in excess of 90 decibels, a level that is frequently exceeded by many common urban sounds, such as jackhammers, jet planes, and excessively loud music. In addition to causing loss of hearing, there is some evidence that noise can produce other deleterious effects on human health and on work performance.

Many large cities have taken measures to decrease the level of urban noise; the problem has received much attention with the advent of supersonic jet airplanes. These aircraft, which travel at speeds faster than the speed of sound, create sound waves (sonic booms) equivalent to those of major explosions and capable of damaging structures. The extent to which continuous exposure to sonic booms affects human health and functioning has yet to be determined. Nevertheless, in 1971 the United States Congress voted down appropriations to support the development of supersonic transport planes; Great Britain, France, and the Soviet Union, however, have developed such planes.

Sonic booms

Chemical pollutants. In 1971 a special committee established by the International Council of Scientific Unions reviewed the existing problems of global pollution and selected those forms believed to be of such a widespread and serious nature as to require a worldwide monitoring program. The pollutants selected for this attention were the chlorinated hydrocarbon pesticides, such as DDT, aldrin, and dieldrin; the polychlorinated biphenyls (PCBs), which are used in a variety of industrial processes and in the manufacture of many kinds of materials; and such metals as mercury, lead, cadmium, arsenic, and beryllium. All of these substances persist in the environment, being slowly, if at all, degraded by natural processes; in addition, all are toxic to life if they accumulate in any appreciable quantity.

The persistent pesticides have created serious ecological problems. As they move through successively higher organisms in food chains, they accumulate in increasingly concentrated forms at each level, causing damaging effects to the predators at the end of the chains—*i.e.*, they are present in low quantities in simple organisms but become more concentrated as these organisms are consumed by more complex ones, which are themselves consumed by predators. Among the species known to be adversely affected are such meat-eating birds as falcons, hawks, and eagles and such fish-eating birds as pelicans, petrels, cormorants, and egrets. The reproduction capacity of all of these birds has been affected by an accumulation of DDT or a similar compound in their tissues. This is manifested by an impairment in the ability of the females to form eggshells properly. As a result, some species lay soft-shelled or shell-less eggs that cannot be hatched, and there has been a general decline in the numbers of these birds in Europe, Japan, and North America. Although the effects of the same chemicals on mammals is less obvious and still a matter for investigation, some studies suggest that DDT can reduce the productivity of plant plankton, upon which all other marine life depends.

As the widespread biological effects of pesticides become better known, there is growing evidence that they lose the ability to control the pests they were designed to kill. Many insect species have developed immunity to a wide range of synthetic pesticides, and the resistance is inherited by their offspring. Furthermore, it has been observed that repeated use of such chemicals creates pest populations in areas in which none previously existed. This happens because the pesticides destroy populations of carnivorous, predatory insects that previously kept the plant-eating insects in check.

Among other materials that are harmful to most forms of life are such metals as mercury, lead, and arsenic. The increasing release of these substances into the biosphere by industrial processes has created conditions that are now generally viewed as harmful to human welfare. Studies have been conducted on metallic pollutants to determine the normal environmental levels, the levels that are toxic to man, and the extent to which industrial processes are responsible for the problem.

The ultimate control of pollution will presumably involve the decision not to allow the escape into the en-

vironment of the substances that are harmful to life, the decision to contain and recycle those substances that could be harmful if released into the environment in excessive quantities, and the decision not to release into the environment substances that persist and are toxic to living things. Essentially, therefore, pollution control does not mean an abandonment of existing productive human activities but their reordering so as to guarantee that their side effects do not outweigh their advantages.

THE FUTURE OF CONSERVATION

The Earth has supported civilized man for more than 5,000 years and agricultural man for twice that period. Before that, reaching back an unmeasured number of years into the past, human or near-human groups occupied various parts of the Earth, modifying them to some degree in the course of hunting, fishing, and food gathering. Patterns of land use were determined over many centuries of trial-and-error experimentation by people equipped with primitive tools who depended on the biological communities of the Earth for their energy supplies. Today, however, with abundant fossil fuels, growing amounts of nuclear energy, and sophisticated tools and machines, it is possible to modify quickly entire landscapes, changing long-established natural patterns into new ones with new purposes. The opportunity to enhance the material welfare and general well-being of great numbers of people is available, as is the opportunity to cause great damage and to impair the capacity of Earth to support life. What the outcome will be will depend on man's ability to plan properly for the use and conservation of the Earth's living and nonliving resources.

Changing patterns of land use

The role of population, industry, and technology. Uses of lands and resources are being modified in the expectation of continued population growth, industrial expansion, and accelerating technological change. Yet it is possible that in the future—perhaps early in the 21st century—uses of lands and resources will be planned with the expectation of population stability, little industrial expansion, and a technology directed toward a reorganization and a rearrangement of man's activities to achieve a better environmental relationship. Although certain countries of the world have already reached some degree of population stability—*e.g.,* Ireland, Hungary, East Germany, France, Sweden, Switzerland, and Japan —industrial expansion and rapid technological change continue in these countries, in part because of the demands made by other expanding nations. The existing expansionist phase of technological civilization cannot, however, be expected to continue indefinitely. The ecological limitations on any population growth in a limited space with limited resources lead most workers in the field to predict an inevitable end to this expansion, even if mankind fails voluntarily to limit its own growth.

Need for ecological considerations and advance planning. Current decisions made about land and resource use have important consequences for the future. If extensive areas of the Earth are badly damaged or their productivity destroyed by the present expansion of technological civilization, they will be difficult, if not impossible, to restore in the future. If a species becomes extinct, for example, it cannot be brought back. It is essential, therefore, that care be exercised in further modifying the planet to suit human purposes. Yet in many developing regions of the world, those where the greatest changes may be expected, little attention is being given to planning for and carefully controlling the use of land and resources. Thus, important tropical, semi-arid, and subpolar regions of the Earth—the three principal climatic belts that have not yet undergone major technological development—are now being changed drastically without much consideration for their environments. In many parts of the world, ecologically trained experts are not available; in others, because of strong economic pressures toward development, ecological advice is either not sought or ignored.

Climatic belts currently being exploited by man

Comparatively speaking, the failure to apply current ecological knowledge to the changing land and resource use taking place in tropical, semi-arid, and sub-Arctic lands is equivalent to the modification of the more temperate lands that took place centuries ago, when ecological knowledge was not available. In earlier centuries, however, the capacity to do irreparable damage was restricted by the lack of machinery, industry, and fuel energy. Today, capabilities are such that major destruction can be accomplished quickly. There is, therefore, a need to call upon environmental expertise during the process of economic development in any area of the world if natural resources are to be conserved and the future welfare of humanity is to receive due consideration.

Ecologists now largely believe that, although it was once possible to allow the development of a region to proceed more or less at random, based on individual wants, aspirations, and decisions about the use of lands and resources, such a process now holds too much risk for the well-being of society as a whole and for the future of the resources on which that society depends. Planning, they argue, must precede development, and regional planning is required if the use of major areas of land and its resources is to be brought into accord with environmental necessities and with the long-term needs of society. More detailed land-use and urban planning are necessary if all lands and resources are to be managed with a view to their conservation. Further, it would appear that population planning to limit growth in areas in which it is not desired and to encourage it in areas in which it can have beneficial effects will have to accompany the planning of resource and land use if such planning is to be effective.

In densely populated and technologically advanced nations, such as those of western and northern Europe, most of the land-use decisions that would affect large areas have already been made. Although changes do occur, mostly in relation to growing urbanization and increasing material wealth on the part of citizens, it seems likely that the remaining woodlands and fields will continue to be devoted to their present uses. In England the interest of the central government in the planning and control of land use and population distribution was marked by the passage of the Town and Country Planning Act shortly after the end of World War II. This legislation led to decisions to limit the growth of London and to develop instead a pattern of new towns outside a greenbelt of agricultural and recreational land that surrounds the metropolis. In France, where there are still large areas of open space, a system for regional planning and control of land use and development (*aménagement du territoire*) has been formulated. It has already resulted in the establishment of new cities and recreational sites in previously undeveloped areas along the Mediterranean coast.

Control of land use and population distribution

It is in the sparsely populated areas in the underdeveloped countries of Africa, Asia, and Latin America as well as in such technologically advanced countries as Canada, Australia, and the Soviet Union that the greatest range of options and choices for the future are available. Because these areas have yet to undergo drastic environmental change, the need for local and regional environmentally oriented planning for resource and land use is most urgent in these countries.

Areas of promise. Many of the current vexing conservation problems may be solved by technological developments. Perhaps the greatest need for the continuation of a highly technological society is an abundant and reliable source of cheap energy. Present promising research on nuclear fusion as a source of power indicates that this process may replace nuclear fission in the near future. If this is true, the use of nuclear energy to generate power will no longer depend on such relatively rare and expensive nuclear fuels as uranium-235 and plutonium. Instead, the almost limitless quantities of heavy hydrogen (deuterium) present in seawater can suffice as a fuel. Moreover, with abundant, cheap power, the desalinization of seawater and the reclamation and concentration of minerals dissolved within it would have possibilities both as a continuing source of abundant

fresh water and, more important, as a new source of the minerals now being depleted. Mineral exploitation could operate on a sustained-yield basis, just as the harvesting of living resources now can be conducted.

Apart from the development of major new sources of power, the greatest promise for the future of mineral resources and for the prevention of pollution of the environment lies in new technologies involving the recycling and reclamation of what are now considered waste products. Demands for new minerals will be greatly reduced when those already available in population centres can be reused more readily. Reclamation of sewage and other organic wastes and restoration of these materials to soils can help to arrest losses in soil fertility and structure and to reduce the need for new supplies of chemical nutrients for soil fertilization. If development of technologies for recycling and reutilization continues, many of the existing problems of environmental pollution will be solved.

In addition to the breeding of new strains of crop plants, the development of techniques for the control of diseases and pests of crop plants that do not involve the release of persistent, poisonous chemicals into the environment holds much promise for the production of greatly increased quantities of food and fibre from smaller areas of the Earth's surface. Much more intensive development of aquaculture (cultivation of the natural produce of water), perhaps utilizing coolant water from nuclear-power plants, can also produce much higher food yields from smaller areas than are now usually obtainable. As a result of these advances in intensive food production, agriculturally marginal lands and the wilder aquatic areas would be spared for the continued support of wild species as well as for the adventure and recreation of mankind, thus helping to solve one of the most troublesome of all conservation problems, the conservation of wild nature.

Possibility of aquaculture for food production

Problems in need of solutions. Considering the potential of industrial and agricultural technology and the accompanying advances in medical science and technology, it is possible to foresee a world in which a relatively stable human population lives at a high level of material affluence, wild nature continues to exist in abundance, and relatively undisturbed lands remain available for human enjoyment. But this scientific and technological optimism is not supported by existing world conditions. Because the scientific and technological knowledge now available is more than adequate to solve most of the world's major environmental problems, those of the 1970s are not problems of science and technology but of the arrangements and functioning of human institutions and of the attitudes of individuals. Thus, while research in forestry science continues in all the forestry schools of the world, tropical forests are being devastated in ways that suggest that forestry science and management never existed. Although the techniques for managing livestock on natural ranges and pasture lands have reached a high level of sophistication, overgrazing continues around most of the world's major deserts, and animals die of hunger, people suffer from deprivation, and the deserts spread. Obviously, the knowledge available to the technician does not reach or influence the behaviour of most of the pastoral people on Earth.

Population growth. During the 1970s population growth continued at a rate estimated to be approximately 2 percent per year. Whereas in 1970 there were over 3,500,000,000 people on Earth, predictions indicate that the population of the world will not be stabilized, even under the best of conditions, before it attains a level of 6,000,000,000–7,000,000,000 people. These predictions assume, of course, that there will be no major catastrophes—outbreaks of war, famine, or disease—that would cause drastic reductions in human numbers. There is little doubt that rapid population growth interferes with orderly economic development, leads to a deterioration of the human environment, places a severe strain on human institutions, and constitutes a growing threat to the survival of wild animal and plant life.

Although techniques for birth control are effective and well-known among educated people, they are virtually unknown, unavailable, and apparently unacceptable to those people having the most rapid rate of population growth—the ones who also live in the most precarious balance with their environment. This does not mean that the prospects for controlling population increase are poor; actually, they are better than at any time in the past. But ecologists insist that, in order to be effective, the materials needed to control births must be made more available; in addition, more education is needed to encourage people to limit the size of families, and the prospects for material and economic advancement for those who have fewer children must be made more obvious. Because all of these needs are expensive, they usually cannot be provided by nations with high rates of population increase and low economic productivity. Thus, control of world population growth is an international responsibility, but it is one toward which most nations give limited support at best.

Necessity for curbing birth rates

Pollution control. Next to population growth, the conservation problem of greatest magnitude is the control of pollution; it might even be argued that it is even more urgent and important than the control of population growth. The knowledge and technology needed to control pollution effectively are now available: pollution-free engines can be built, pollution-free factories have been put into operation, and techniques for controlling agricultural insect pests with a minimum use of persistent pesticides have been developed. For economic reasons, none of these measures, however, is being applied universally, and political and social pressures have not yet forced their application. Moreover, emerging nations have expressed fear that excessive concern over pollution could impede their economic development. Indeed, some of these countries have become sanctuaries for industries that find it less expensive to operate in areas with more lax standards. It is apparent that pollution control, regardless of the state of its technology, will become a reality only when people demand it and only when nations are willing to agree on appropriate international standards.

Control over use of resources. Equally as important to the future of world conservation as population growth and pollution is the failure of most nations to exercise adequate controls over land, water, and other resource use. Effective means for controlling land use do not exist in most countries; laws and regulations that permit governments to exercise such control, when existent, often cannot be enforced because of the danger of strong public resentment and resistance. Although it is essential for conservation that lands and all other resources be used with a view to preserving their future productivity, this view all too often conflicts with present needs or demands of the resource users. Unlike other conservation problems, however, the solution to this conflict is not within the scope of science or technology; instead, it is a legal, economic, and political matter as well as one for the social and behavioral sciences. Its resolution also requires a much more effective use of environmental education that has been practiced thus far in most countries.

Extraterrestrial conservation. Space exploration has begun at a time when man's willingness to manage the Earth's resources is in doubt. Experiences gained from early space efforts, however, have helped promote an awareness of planetary realities among a growing number of people. The need to provide the requirements of life to men confined for long periods of time in spaceships and the problems of waste disposal in such vehicles have cast new light on similar problems within Earth-bound communities. The view of Earth from outer space has engendered the concept that the planet is, in effect, a large spaceship with a limited capacity to support life and that it is highly vulnerable to damage from poorly planned human activities.

Spaceship Earth

Exploration of other planets has been undertaken in full realization of the dangers involved, both in the pollution of a satellite or planet being visited by Earthlings and in the potential danger to life on Earth by

materials brought back from other celestial bodies. Although there is as yet no reason to expect the existence of life on those planets nearest the Earth and every reason to doubt its existence on planets farther away in the solar system, the most rigid precautions are still necessary. Experience with the rapid spread of diseases in new locations has demonstrated the dangers to other forms of life if other planets were to be contaminated with organisms brought from Earth or if Earth were to be exposed to agents from other planets capable of contaminating life on this one.

If planets capable of supporting life are reached someday, they will offer an opportunity to science never before available. The study of life that has evolved under different conditions in a different world could add greatly to an understanding of the Earth's biological systems as well as provide a great intellectual challenge to mankind. Perhaps the hard lessons of conservation, which have yet to be completely understood or accepted on Earth, will be applied to other worlds before irrevocable damage is done.

At the present time, however, the available knowledge brought back from outer space indicates only that Earth is unique, the one planet where life can thrive. Its inhabitants, therefore, are vested with a special responsibility: to keep it functioning.

BIBLIOGRAPHY

General references: RAYMOND F. DASMANN, *Environmental Conservation*, 3rd ed. (1972), a general textbook on the conservation of natural resources, stressing their interrelationships and the use that men make of them; FRANK FRASER DARLING, *Wilderness and Plenty*, (1970), a series of lectures presented by the British Broadcasting Company in 1969 that had a great effect on environmental thinking in Europe; FRANK FRASER DARLING and JOHN P. MILTON (eds.), *Future Environments of North America* (1966), a multidisciplinary view of the present and future of the North American environment; FAIRFIELD OSBORN, *Our Plundered Planet* (1948), a classic in conservation, one of the first world views of the impact of human populations and land use on the natural resources of the earth; WILLIAM L. THOMAS, JR. (ed.), *Man's Role in Changing the Face of the Earth* (1956), a comprehensive presentation of man's past, present, and probable future roles on Earth that is rapidly becoming a classic in conservation; UNESCO, *Use and Conservation of the Biosphere* (1970), papers presented and conclusions reached by the first major international conference on problems of the human environment.

History of conservation: CLARENCE J. GLACKEN, *Traces on the Rhodian Shore* (1967), an account of the philosophies and ideas relating to man's attitudes toward nature from ancient times to the 18th century; GEORGE PERKINS MARSH, *Man and Nature* (1864), a classic of conservation and the first general view of how man changes the Earth and affects its living resources; RODERICK L. NASH (ed.), *The American Environment: Readings in the History of Conservation* (1968), a review of the important developments in the evolution of conservation in the United States with readings from the works of those who contributed to its development; HENRY DAVID THOREAU, *Walden* (1858, reissued 1961), a classic philosophical exploration of man's relationship with nature and the values to be found in living apart from civilization and its artifacts, by one of the first advocates of wilderness preservation; STEWART L. UDALL, *The Quiet Crisis* (1963), a review of the history of conservation and the growing crisis in conservation, with emphasis on public lands and living resources in the United States.

Ecology: CHARLES S. ELTON, *Animal Ecology* (1927), a classic in ecology and the forerunner to the application of animal ecology in the management of wildlife; EUGENE P. ODUM, *Fundamentals of Ecology* (1971), a standard textbook on general ecology, emphasizing the nature and functions of ecosystems with a discussion of the relationships between ecology and the environmental problems of today.

Pollution: RACHEL CARSON, *Silent Spring* (1962), a popular best-seller that first alerted the general public to the dangers inherent in the widespread use of persistent pesticides; BARRY COMMONER, *Science and Survival* (1966), a discussion of how science and technology can cause unanticipated changes in the environment, with particular reference to pollution resulting from the use of nuclear energy; COUNCIL ON ENVIRONMENTAL QUALITY, *Environmental Quality* (annual), a review of the state of U.S. environment with particular emphasis on problems relating to pollution; ROBERT L. RUDD, *Pesticides and the Living Landscape* (1964), a scholarly review of the ways in which pesticides enter into the ecological networks in natural and man-made environments, with a discussion of the dangers involved in their continued use.

Population, resources, and technological development: GEORG BORGSTROM, *The Hungry Planet* (1965), a review of the food needs of mankind and the prospects for meeting them, with particular attention to the use of ocean resources; ARTHUR B. BRONWELL (ed.), *Science and Technology in the World of the Future* (1970), generally optimistic accounts by selected experts of man's prospects in meeting his future needs for energy and mineral resources; M.T. FARVAR and J. MILTON (eds.), *The Careless Technology* (1971), a collection of case histories and related discussions concerning the unexpected side effects of economic and technological development; HANS H. LANDSBERG, *Natural Resources for U.S. Growth: A Look Ahead to the Year 2000* (1964), a review of the needs of society for natural resources and the prospects of meeting them from available or foreseeable supplies; WILLIAM VOGT, *Road to Survival* (1948), a classic on the relationship between populations and resources that is written from an emotional viewpoint in an attempt to draw attention to the coming population crisis, emphasizing the growing threat of world hunger as well as the impact of man on soils and renewable resources.

Management of living resources: RAYMOND F. DASMANN, *Wildlife Biology* (1964), an introductory college textbook reviewing the principles of wildlife biology and their application to the conservation and management of wild animals; S. HADEN-GUEST, J.K. WRIGHT, and E.M. TECLAFF (eds.), *World Geography of Forest Resources* (1956), a survey of the forest resources of the world, the extent to which they are being used, and the problems associated with their management; G.V. JACKS and R.O. WHYTE, *Vanishing Lands* (1939), a classic that reviews the nature of world soils and describes their destruction by erosion as a result of misuse; ALDO LEOPOLD, *Game Management* (1933), the first textbook of wildlife management that has become a conservation classic; *A Sand County Almanac and Sketches Here and There* (1949), the ethics and aesthetics of conservation with emphasis on land, wildlife, and wilderness; L.A. STODDART and A.D. SMITH, *Range Management*, 2nd ed. (1955), a textbook on the principles and practices affecting the use of range and pasture lands by grazing animals.

(R.F.D.)

Conservatism

The term conservatism, although it has had different implications in varying historical and geographical contexts, is best reserved to denote a preference for institutions and practices that have evolved historically and that are thus manifestations of continuity and stability. Political thought, from its beginnings, contains many strains that can be retrospectively labelled conservative, but it was not until the late 18th century that conservatism began to develop as a political attitude and movement reacting against the French Revolution of 1789. The noun seems to have been first used after 1815 by French Bourbon restorationists such as François-René, vicomte de Chateaubriand. It was used to describe the British Tory Party in 1830 by John Wilson Croker, the editor of *The Quarterly Review;* and John Calhoun, a formulator of conservative minority rights against majority dictatorship in the United States, also used the term in the 1830s. The generally acknowledged originator of modern, articulated conservatism (although he never employed the term) was the British parliamentarian and political writer Edmund Burke in his essay *Reflections on the Revolution in France* (1790). Pro-parliamentarian opponents of the French Revolution, such as Burke, believed that the violent, untraditional, and uprooting methods of the Revolution outweighed and corrupted its liberating ideals. More authoritarian opponents, such as the polemicist and diplomat Joseph de Maistre, also rejected the ideals themselves. The general revulsion against the course of events in France provided conservatives with an opportunity for restoring the pre-Revolutionary traditions, and a sudden flowering of more than one brand of conservative philosophy followed.

CONSERVATIVE ATTITUDES

Because Burke's case against radicalism and revolution has also influenced liberals, there is often no sharp dis-

tinction between liberals and conservatives in action. In philosophy, however, conservatism has maintained certain sharply nonliberal assumptions about human nature.

Whether intentionally or unconsciously, whether literally or metaphorically, for example, conservatives tend to assume in politics the Christian doctrine of man's innate original sin, and herein lies a key distinction between conservatives and liberals. Men are not born naturally free or good (conservatives assume) but are naturally prone to anarchy, evil, and mutual destruction. What the 18th-century French philosopher Jean-Jacques Rousseau denounced as the "chains" that hinder man's "natural goodness," are for Burkeans the props that make man good. These "chains" (society's traditional restrictions on the ego) fit man into a rooted, durable framework, without which ethical behaviour and responsible use of liberty are impossible.

The conservative temperament may be, but need not be, identical with conservative politics or right-wing economics; it may sometimes accompany left-wing politics or economics. Regardless of a conservative's politics or economics, however, it can be said that two characteristics of the conservative temperament are: a distrust of human nature, of rootlessness, of untested innovations; and a corresponding trust in unbroken historical continuity and in traditional frameworks within which human affairs may be conducted. Such a framework may be religious or cultural or may be given no abstract or institutional expression at all. In relation to the latter aspect, many authorities on conservatism—a minority in France and a majority in England—consider conservatism an inarticulate state of mind and not at all an ideology. Liberalism argues; conservatism simply *is*. When conservatism becomes ideologized, logical, and self-conscious, then it resembles the liberal rationalism that it opposes. According to this British approach, logical deductive reasoning is too doctrinaire, too 18th century. Whereas the liberal and rationalist mind consciously articulates abstract blueprints, the conservative mind unconsciously incarnates concrete traditions. And, because conservatism embodies rather than argues, its best insights are almost never developed into sustained theoretical works equal to those of liberalism and radicalism.

Conservatism as a state of mind

Conservatism is often associated with some traditional and established form of religion. After 1789, the appeal of religion redoubled for those craving security in an age of chaos. The Roman Catholic Church, because its roots are in the monarchic Middle Ages, has appealed to more conservatives than any other religion. Himself a Church of England Protestant, Burke praised Catholicism as "the most effectual barrier" against radicalism. But conservatism has had no dearth of Protestant and strongly anticlerical adherents also.

Conservatives typically view society as a single organism and condemn as "rationalist blueprints" the attempts of progressives to plan society in advance from pure reason instead of letting it evolve naturally and unconsciously, flowering from the deep roots of tradition. They dismiss a liberal society as "atomistic," meaning composed of disrupted elements held together merely mechanically. A society, they argue, has to be rendered whole by religion, idealism, shared historical experiences, commitment to its long standing political institutions, and by the emotions of reverence, cooperation, and loyalty; a society, they believe, can, to the contrary, be rendered atomistic by materialism, class war, excessive laissez-faire economics, greedy profiteering, overanalytical intellectuality, subversion of shared institutions, insistence on rights above duties, and by the emotions of skepticism and cynicism. Except for the German Romantic school, conservatives do not carry their conceptions of the organic wholeness of society to the extreme at which the individual becomes nothing, society everything, for they recognize that, at that extreme, one no longer has conservatism but totalitarian statism.

VARIETIES OF CONSERVATISM

The Burkean foundations. Burke did more than any other thinker to turn the intellectual tide from a rationalist contempt for the past to a traditionalist reverence for it. An Irishman, he loved England, including its established Anglican Church and its nobility, with an outsider's passion. In 1765 he became private secretary to Charles Watson-Wentworth, 2nd marquess of Rockingham, the head of the less liberal wing of the Whig Party. Against the untraditional tyranny of George III, Burke defended the American Revolution of 1776, which he viewed as being in defense of traditional liberties, but attacked the radical French Revolution of 1789 as tyranny by mobs and deracinated theorizers. At a time (1790) when the French Revolution still seemed a bloodless utopia, he predicted its later phase of terror and dictatorship, not by any lucky blind guess but by an analysis of its devaluation of tradition and inherited values.

Indeed, the core of Burke's thought and of conservatism is fear of rootlessness. Rousseau's *Social Contract* of 1762 had favoured a contract merely among the living, to arrange government for their mutual benefit. Burke, instead, argued:

Burke's conception of the social contract

> Society is indeed a contract . . . [but] as the ends of such a partnership cannot be obtained in many generations, it becomes a partnership not only between those who are living, but between those who are living, those who are dead, and those who are to be born. . . . Changing the state as often as there are floating fancies, . . . no one generation could link with the other. Men would be little better than the flies of a summer.

Burke's veneration of the past may be contrasted with the rationalist hostility of Karl Marx, the most influential social critic of modern times: "The legacy of the dead generations weighs like a nightmare upon the brains of the living." But for Burke the contract is with "the future" as well as with the past, and he thus urges improvement, as long as it is evolutionary: "A disposition to preserve and an ability to improve, taken together, would be my standard of a statesman."

Burke was defending not conservatism in the abstract but, rather, one concrete instance of it, the unwritten British constitution. His arguments, however, were not always consistent. Sometimes he justified that constitution by "natural rights"; more often by "prescriptive right." Natural rights meant a universal code external to any given constitution; prescriptive right, a local code authoritative (prescriptive) by virtue of its age and its links with the past, which are prima facie evidence of its value. Sometimes he argued that natural rights preceded the constitution and gave it "latent wisdom." But, when arguing against French rationalists, who would justify their own revolutionary constitution by natural rights, he argued instead, and more typically:

> Our constitution is a prescriptive constitution . . . [whose] sole authority is that it has existed time out of mind . . . without any reference whatever to any other more general or prior right.

Burke shocked his century by his brutal frankness in defending "illusions" and "prejudices" as socially necessary. In doing so, however, he was, in fact, being not so much a cynic as one of the few old-fashioned Christians among 18th-century intellectuals. He was an old-fashioned Christian in the sense of believing man innately depraved, innately steeped in original sin, and incapable of bettering himself by his feeble reason. So defined, man could be tamed only by following an ethically trained elite and by education in "prejudices," such as family, religion, and aristocracy. He called landed aristocrats "the great oaks" and "proper chieftains," provided they tempered their rule by a spirit of timely reform from above and remained within the constitutional framework. He defended the Church of England for its political as well as its religious function, "To keep moral, civil, and political bonds, together binding human understanding."

Coleridge and Wordsworth. After Burke, the English poets Samuel Taylor Coleridge and William Wordsworth were significant figures in the formulation and expression of conservative sentiment. They began, however, as utopian liberals supporting the French Revolution. Wordsworth spoke for a whole generation of European intellectuals with his famous salute to the new dawn in France:

"Bliss was it in that dawn to be alive, but to be young was very heaven." Disillusionment followed, and Coleridge and Wordsworth reacted against liberalism and rationalism and turned to traditional monarchy and the Church of England.

In 1798 Wordsworth and Coleridge jointly published their book of poems, *Lyrical Ballads*, marking the revolt of the human heart against abstract 18th-century rationalists and thereby helping to create a new philosophical climate. Conservatism was permanently influenced by Coleridge's prose works: *Lay Sermons*, 1816–17; *Biographia Literaria*, 1817; *Philosophical Lectures*, 1818–19; *Aids to Reflection in the Formation of a Manly Character, on the Several Grounds of Prudence, Morality, and Religion*, 1825; and his various *Letters* and *Specimens of Table Talk*. His public lectures exercised an indirect influence by molding the minds of university students who later became national leaders.

Coleridge's views on social classes

According to Coleridge, society divided its functions among different "class orders." Each class had its valuable function, but this did not necessarily include the right to vote and rule. That right was best left to an ethically trained aristocracy, functioning within the strict lawful limits of Parliament. All classes, Coleridge argued, must cooperate harmoniously within the organic unity of the constitution. His greatest influence on practical politics was through his disciple Benjamin Disraeli, later to be Conservative prime minister, and his disciple's disciple, Sir Winston Churchill. Coleridge considered businessmen often subversive, not conservative; they allegedly gnawed at the foundations of Christian monarchy by substituting a newfangled, un-Christian religion known as economic profit. Thus Coleridge, defining "shopkeepers" as "the least patriotic and the least conservative" class, fought against the Whig Reform Bill of 1832, which made "hucksters" the dominant voting group.

Maistre and Latin conservatism. It would convey an unbalanced picture of conservatism to present only the moderate and British brand founded by Burke and to omit the more extreme and Latin brand founded by Maistre (died 1821). Whereas Burkean conservatism is evolutionary, the conservatism of Maistre is counter-revolutionary. Both favour tradition against the innovations of 1789, but their traditions differ: the former fights against 1789 for the sake of traditional liberties, the latter for the sake of traditional authority. The former is not authoritarian but constitutionalist—and often parliamentary—whereas the latter, in its stress on the authority of some traditional elite, is often justifiably called not conservative but reactionary. To call it totalitarian, however, would be to go much too far, for its authority does not try to be "total," in the sense of taking over the total personality, the total culture, but is restricted to politics—and sometimes also religion. The distinction between the authoritarian and the totalitarian separates even the most reactionary conservative from the totalitarian Nazis and Communists.

After the breakdown of the French Revolution, Maistre became the most influential philosophical spokesman for the *ancien régime*. Against the slogan "liberty, equality, fraternity," he seemed almost personally to embody the slogan "throne and altar." His program consisted of a restoration of hereditary monarchy but a more religious and less frivolous monarchy than before. He was an international refugee after the French, during the Revolution, invaded his native Savoy—then a French-speaking province of the Italian-speaking monarchy of Piedmont-Sardinia. He became for 14 years Sardinian ambassador to Russia, where his restorationist faith was strengthened by the example of the absolute monarchy still functioning there.

Maistre on the role of monarchy

Both restorationist and evolutionary conservatives defended monarchy as a social cement needed to hold society together, to keep it "organic," not "atomistic." But, while the Maistre school (key source of conservative thought in Spain and Italy as well as France) defends monarchy as absolute, the evolutionary British school defends it merely as being "pragmatic"; that is, useful. Maistre and many continental monarchists carried their belief in the monarchy to the extreme of demanding "love" even for an "unjust" ruler, earthly or heavenly:

> We find ourselves in a realm whose sovereign has proclaimed his laws. . . . Some . . . appear hard and even unjust . . . What should be done? Leave the realm, perhaps? Impossible: the realm is everywhere. . . . Since we start with the supposition that the master exists and that we must serve him absolutely, is it not better to serve him, whatever his nature, with love than without it?

This chain of authoritarian reasoning reached its climax in a logical if inhuman paradox: "The more terrible God appears to us . . . the more our prayers must become ardent. . . ." Cruel as these arguments sound, the motive of the personally mild Maistre was humane: revolts against cruel authority would inflict even crueler sufferings on mankind. He drew from the French Revolution the lesson that submission to traditional authority, though admittedly a bitter pill, was Europe's cure for a still more bitter chaos.

Maistre's politics were a theological drama in which "order" (his key concept) was angelic, "chaos" diabolic, and "revolution" original sin. Seduced by the glittering *Social Contract* of Rousseau, giddy and inexperienced nations might lust after democracy or a plebeian Bonapartist dictatorship. But they would come to a perfectly dreadful end, which would serve them right for provoking the wages of sin: "Because she [Europe] is guilty, she suffers" (1810). From suffering, Maistre argued, Europe would learn that the purest order is a fatherly Christian monarchy. Even kings must avoid rocking the boat of order with liberal "innovations": Europe must "suspect" the word "reform." In *Du Pape* (1817; "Concerning the Pope"), he analyzed "order" further: its hierarchical pyramid logically required one supreme apex. That apex must be no earthly monarch, of which there were so many, but the union of earthly and spiritual power in the papacy.

The vast extent of the instability following the French Revolution surprised even its supporters, and the problem of how to restabilize society emerged as one of some practical importance. According to Maistre's *Soirées de Saint-Pétersbourg* (left unfinished 1821; "Evening Conversations in St. Petersburg"), the solution was more faith and more police. That combination he summed up in his own frank formula: "the pope and the executioner." The pope was the positive bulwark of order: he gave faith. The executioner was the negative bulwark: he suppressed disorder. Himself an intellectual, Maistre indicted intellectuals as "rebellious" and "insolent" fomenters of disorder.

Maistre, this very secular exalter of clericalism, resembled not the Church Fathers but the very rationalists he attacked. He arrived at his glorification of unreason and of divine authority not by mystic intuition—not even by unthinking acceptance of traditional authority—but by using his own mind independently, rationally, and with steps of deductive logic. Though Maistre would never have admitted it, he might be characterized as the last abstract rationalist of the whole Voltairean Age of Reason. Even more than the rationalist Voltaire and as much as the rationalist Jacobins, Maistre believed in pure and absolute ideas, although his idea was absolute authority rather than absolute reason. In Maistre the destructive deductive logic of the 18th century was carried so far that it destroyed even itself—pure reason committing suicide for the sake of pure order.

This division into Burke and Maistre wings does not mean both were equal in importance or influence. No work of Maistre or any other anti-Jacobin has approached the influence of Burke's classic essay. Burke, above all, was the first to formulate the rebuttal to the French Revolution; his arguments were borrowed, sometimes word for word, by all later conservatives, including the restorationists. Maistre's rigid hierarchical conservatism is in the latter part of the 20th century dying out, whereas Burke's more flexible brand is stronger than ever, permeating all parties of the West, emphatically including democratic Socialists with their increasing stress,

in Great Britain and Germany, on what a Fabian Socialist has called, in good Burkean language, "the inevitability of gradualness."

Later French conservatism

French conservatism after Maistre presents a diversified range of views, from the thought of Charles Maurras (died 1952), the far-right editor of *L'Action Française* who seemed more Fascist than conservative and became a Nazi collaborator, to the anti-authoritarian Alexis de Tocqueville (died 1859), author of *Democracy in America* and the most Burkean French critic of the Revolution and of plebiscitarian mass democracy. To some extent, however, Tocqueville, an evolutionary parliamentarian, can also be regarded as a liberal thinker. In between Maurras and Tocqueville come the great anti-Jacobin Hippolyte-Adolphe Taine; the philosophical novelist Maurice Barrès, more a nationalist than anything else but conservative in his stress on organic roots; and Louis-François Veuillot, the editor after 1843 of the newspaper *L'Univers Réligieux* and a clerical restorationist who ably re-adapted Maistre to the industrial modern world. An influential right-wing extremist, less clerical and more statist than Maistre and Veuillot, was Louis-Jacques-Maurice de Bonald (died 1840), the apologist for Napoleon's empire and then for the Bourbon Restoration.

Metternich and the Concert of Europe. The problems posed by the widespread social unrest of the Revolutionary and Napoleonic periods and their aftermath, and the insecurity of governments in the face of demands for constitutions and liberal reforms, provoked a reaction of more immediate and far-reaching consequence than the writings of conservative theorists. During the period 1815–48, Prince Metternich, a major influence in Austria and in Europe generally, devoted his energies to erecting an anti-revolutionary chain of international alliances throughout Europe in order to protect the multinational empire that he administered.

Metternich viewed the liberal revolutions of the 1820s and 1830s in Italy, Spain, and Germany as being unhistorical and unrealistic. Liberals were trying to transplant from England free institutions, which had no historic roots on the Continent. He retorted with Burkean arguments about the need for old roots and orderly organic development. Hence, his sarcastic comments on the liberal revolutions in Naples and elsewhere:

> A people who can neither read nor write, whose last word is the dagger—fine material for constitutional principles! . . . The English constitution is the work of centuries . . . . There is no universal recipe for constitutions.

Though his repressive Carlsbad Decrees of 1819 infringed inexcusably on basic liberties, his attitude was not always so negative. Just before his fall in 1848, he was at last winning acceptance from the archdukes of his sincere, thoughtful, and practical plan (postponed too long by the reactionary emperor Francis I) to convoke delegates from all the provincial estates to a representative body in Vienna.

Conservative objectives following the Napoleonic Wars

Metternich was a dominating figure at the Congress of Vienna, the international peace conference of 1815 after the Napoleonic Wars. The Vienna peace was based on certain conservative principles shared by the Austrian delegate Metternich, the British delegate Robert Castlereagh, the French delegate Talleyrand, and the formerly liberal Russian tsar Alexander I. These principles were conservatism, in reaction against Revolutionary France; traditionalism, in reaction against 25 years of rapid change; legitimism (the principle of hereditary monarchy as the only lawful rule); and restoration (the principle of restoring the kings ousted after 1789).

The European great powers also aimed at the enforcement of peace by subsequent conferences between kings, and those subsequent conferences gave rise to a period of international cooperation known as the Concert of Europe. As liberal democrats correctly pointed out, the weakness of that first successful attempt at a "United Nations" was its narrowly aristocratic base. But it did achieve the positive function—and important precedent—of peacefully arbitrating several disputes. The debit of the conservative Concert of Europe was its bigoted suppression of democratic social progress.

Goethe's spiritual conservatism. Johann Wolfgang von Goethe was Germany's greatest dramatist, poet, and personality. In his youthful "storm and stress" period of the 1770s, Goethe went through a phase of revolt and of nationalism. In his old age, however, he became Germany's greatest cultural influence for classical balance and for antinationalist cosmopolitanism, influencing many outside Germany, including, in England, Coleridge. In 1815 Goethe and Metternich both took pride in being "good Europeans," not German nationalists. After a friendly personal conversation with Metternich, Goethe wrote that Metternich "inspires with the assurance that reason, reconciliation, and human understanding will lead us out of present chaos." Later, in 1830, Goethe urged a mature synthesis between a conservative framework and liberal goals:

> The genuine liberal tries to achieve as much good as he can with the available means to which he is limited; but he would not use fire and sword to annihilate the often inevitable wrongs. Making progress at a judicious pace, he strives to remove society's deficiencies gradually without at the same time destroying an equal amount of good by violent measures. In this ever-imperfect world he contents himself with what is good until time and circumstances favor his attaining something better.

His rhymed credo "Nature and Art" (1802) expressed his conservative and classic stress on voluntary submission to law: "Only in self-restriction does the master reveal himself. And only law can give us liberty." His political drama *Die natürliche Tochter* (1803; *The Natural Daughter*) reflected his hostility to the French Revolution, radicalism, and mass movements. Much quoted by classicists, such as the United States' Irving Babbitt, was Goethe's definition: "The classical I call the healthy and the romantic the diseased." Yet his *Faust* drama (*Part I* published 1808, *Part II* 1832) retained the liberal-minded stress of his younger days on constant change, "constant striving," as salvation. His most unique achievement consisted of his being, so to speak, self-invented. By sheer strength of character, he remolded his naturally revolutionary and romantic temperament into what the world accepted as a conservative and classicist temperament.

Savigny and Ranke

Perhaps Germany's most mature conservative thought came from her great historians. Friedrich Karl von Savigny (died 1861) and Leopold von Ranke (died 1886) were outstanding as pupils of Burke in their reverence for history as organic growth. Savigny stressed that custom, operating over centuries, creates its own framework. On custom, Savigny founded an entire science of historical jurisprudence, denying the abstract, liberal "rights of man." Similarly, Ranke saw every society in terms of its own unique evolution. He opposed the universal generalizations of the 18th-century Enlightenment; every people, he wrote, "is related directly to God" in its own concrete way.

Tsarist and Dostoyevskyan conservatism. Whereas Western conservatism arose from reactions to the French Revolution, Russian tsarist conservatism had different and older origins. The practice of the absolute Tatar khans and the theory of Byzantine caesarism combined to produce an un-Western elephantiasis of autocracy. Nevertheless, two antiliberal traditionalists of Russia made such an impact on the West—the first by politics, the second by art—that their mention is indispensable: Konstantin Pobedonostsev (died 1907) and Fyodor Dostoyevsky (died 1881). The former was the tutor and chief ideologist of two tsars (Alexander III and, until the Revolution of 1905, Nicholas II). His book *Reflections of a Russian Statesman* (1898) denounced free press, trial by jury, parliamentary government, secular education, skepticism toward the divine mission of tsars, and, above all, intellectuals.

Dostoyevsky's disillusionment with his youthful radicalism resembled Coleridge's in its psychological as well as literary consequences. Both turned to an organic, religious, and monarchic society, to which they paid more homage via literature than via politics. Dostoyevsky attacked Socialism, liberalism, materialism, and atheism. He preached Greek Orthodox tsarism, Slavic traditional-

ism, and the redemption of mankind by "Holy Russia." His novel *The Possessed* (1871–72) pictured the idealistic ends of Socialists as corrupted by their terroristic means, and he boasted somewhat fawningly to Alexander III about the book's effectiveness against radicals. His novel *The Brothers Karamazov* (1880) contrasted a dry Western rationalism with a more deeply moving Russian mysticism. To the end he retained from his young Socialist days his characteristic compassion for what he called "the insulted and injured"; only now he expressed this in the more spiritual creed of Christian love. What influences many modern readers so compellingly is not his political but his cultural conservatism, exalting vision beyond external material progress.

American conservatism. The American Revolution owed many of its ideals to Burke's interpretation of the British heritage of 1688, the heritage of mature self-government. Burke favoured the Revolution as defending the traditional rights of freeborn Englishmen against newfangled royal usurpations. In that sense, one might describe it not as the Revolution but as the "Conservation" of 1776.

In *The Rights of the British Colonies Asserted and Proved* (1764) the American spokesman James Otis typically argued that the demand for no taxation without representation was an old British tradition. America, he said, was conserving "the British Constitution, the most free one on earth." "We claim nothing," added George Mason of Virginia, "but the liberty and privileges of Englishmen." Almost all other revolutions, colonial or otherwise, have been radical in the sense of demanding new or increased liberties and a new order. In contrast, the American demand of July 6, 1775 (*Declaration of the Causes & Necessity of Taking Up Arms*), was for conserving old liberties and the old order: "in defence of the freedom that is our birth right and which we ever enjoyed until the late violation of it." Such words promulgated no democracy, no abstract "Rights of Man"; rather, they promulgated what Burke called "prescriptive right. . . . considering our liberties in the light of an inheritance." Despite important exceptions, which should not be minimized, it was not until the election of the more truly "revolutionary" Andrew Jackson (1828) that the democratic doctrines of the pamphleteer Thomas Paine gained solid roots in the United States, dividing the nation between conservative and progressive traditions. Paine was the man whom the Burkean John Adams (president 1797–1801) came to loathe most—for eternally sloganizing about apriorist utopias. A leading historian, Daniel Boorstin, has observed in *The Genius of American Politics* (1953):

> The ablest defender of the Revolution—in fact, the greatest political theorist of the American Revolution—was also the great theorist of British conservatism, Edmund Burke. . . . Ours was one of the few conservative colonial rebellions of modern times.

Conservative doctrines of Hamilton and Madison

The spirit of the United States was partly molded by two masterpieces of Burkean conservatism, both published in 1787–88: *The Federalist*, by Alexander Hamilton, James Madison, and John Jay, and *Defence of the Constitutions of Government of the United States of America*, by John Adams. The achievements attributed by historians to the *Federalist* papers exceed those of any other series of newspaper articles in history, for they forged a close-knit unity during a separatist crisis. In the context of the Shays's Rebellion of 1786 against the judiciary, they saved government by law from government by mob and established minority rights against majority dictatorship. They based American liberty on the Burkean principle of historical roots, prescriptive right, and judicial precedent instead of on vague grand rhetoric about democratic utopias and the masses. Similar in thought and richer in historical background was the *Defence* by Adams, one of the most penetrating analyses of self-government ever written.

The U.S. Constitution was drawn up in Philadelphia by the U.S. Constitutional Convention of 1787. The objectives of many liberal democrats were: easy amendment; facilities for mass pressure and rapid change; unchecked popular sovereignty; universal manhood suffrage; a single parliamentary body; and the basing of liberty on a long list of universal apriori abstractions, such as Burke later criticized in the French Declaration of the Rights of Man and of the Citizen. But in the Constitution of 1787 the Federalists foiled each of these objectives. They made amendments slow and difficult, greatly reduced the number of voters by property restrictions, created a congress of two parliamentary bodies, and based liberty primarily, though not entirely, on the concrete, inherited precedents of British tradition. Except for the House of Representatives (a sop to democrats), the main cogs of government —president, Senate, justices—were not to be chosen directly by the people but, respectively, by the electoral college, state legislatures, and appointment, and not until 1913 did an amendment eliminate this intentionally undemocratic election of senators. The judicial branch (Supreme Court) continues to be a nonelective, nonremovable elite not responsible to democratic majorities. Yet it can veto as unconstitutional measures passed by a democratic majority of the two elective, removable branches of Congress.

The American Founding Fathers adopted a conservative constitution in reaction against current mob excesses and against the democratic–utopian rhetoric of the earlier Declaration of Independence (drawn up by Thomas Jefferson) with its grand abstractions about "life, liberty, and the pursuit of happiness." Yet the Constitution was the Burkean, not the reactionary brand of conservatism. Thus it defeated not only the liberal objectives but also the more extreme conservative ones, including a hereditary, titled aristocracy and Hamilton's notion of a president for life with absolute veto power.

The United States' only consistently conservative party was the Federalist Party of John Adams and Alexander Hamilton. Hamilton was perhaps too much the reckless commercial adventurer to be classified under conservative or any other principles, but Adams remains the closest New World equivalent to Burke. After the death of the Federalist Party in the early 1800s, two mutually hostile kinds of political conservatism emerged: that of the urban New England Brahmins and that of the Southern semi-feudal landowners. The latter received their most persuasive defense in the famous *A Disquisition on Government and Discourse on the Constitution and Government of the United States* of Calhoun, the closest New World equivalent to Maistre. This more extreme, very regional Calhoun conservatism still dominates much of the American south in the 1970s, typically cutting across Democrat or Republican party lines and still alien to New England conservatism.

Modern U.S. political parties, being pragmatic alliances of geographic patronage groups rather than matters of doctrine, cannot realistically be classified under "isms." It is nearer to reality to look for conservatism, instead, in the indirect diffusion—cutting across all party lines—of the above described restraining principles of the Constitution.

CONSERVATISM IN THE 20TH CENTURY

The 19th and, particularly, the 20th century (that is, the period since the 18th century Enlightenment) have in many ways been antithetical to conservatism, both as a political philosophy and as a program of particular parties identified with conservative interests. As described above, the consciously articulated conservatism of Burke was formulated in reaction to the French Revolution; similarly, the anti-liberal, anti-revolutionary policy that was a major factor in European international relations during the Metternich period (1809–48) was a reaction to the political discontent aroused by demands for liberal reforms and constitutions. The Enlightenment, in fact, had resulted in the propagation of certain attitudes and ideas that were to have far-reaching political consequences during the succeeding centuries, the most significant of which were a belief in the possibility of improvement in the human condition—a belief, that is, in the idea of progress—and a concomitant disposition to tamper with or discard existing institutions or practices in

pursuit of progress, a disposition that has been characterized as "rationalist." Such rationalist politics embrace a broad segment of the political spectrum, including much of liberal reformism, socialism of the welfare-state or mixed-economy variety characteristic of western Europe, and Marxist socialism. The changes that have been wrought under the banner of rationalist politics have thus been immense and point to what has been described as a dilemma of modern conservatism—the extent to which, in face of constant rationalist innovation, conservatives may be forced to adopt a merely defensive role, so that the political initiative lies always in the other camp.

The responses of conservatives to this predicament have naturally varied considerably in differing political contexts; an account of some of these responses is given below. An analysis of the role of conservatism in contemporary politics, however, cannot be confined merely to an account of the programs of political parties identified with the conservative cause, for conservatism makes its influence felt in a variety of ways less direct than through expression in party platforms. Conservatism in the 20th century has in fact been a pervasive force in the political life of those parliamentary democracies in which rationalist politics have seemed to hold sway, as well, of course, as in less liberal political climates, such as in Spain and Portugal.

Non-political manifestations of conservatism

Conservative influences operate indirectly (*i.e.*, other than via the programs of political parties) largely by virtue of the fact that, while man is undeniably a persistent innovator, there is also much in the human temperament that is naturally or instinctively conservative: among such conservative traits are the tendency to fear and avoid sudden change and the tendency to act according to habit. While these are traits of the individual, they may find collective expression in, for example, resistance to imposed political change and in a whole cluster of value preferences that contribute to the formation and stability of a particular culture. The tendency for value preferences to find expression in cultural forms and political institutions (the so-called pragmatism of the British, for example, in their unwritten constitution) constitutes a profound conservative influence in political life over and above any explicit articulation of particular conservative interests that may be undertaken by a political party, for it gives rise to practices and institutions that are products of a long process of social and political evolution and are closely related to other culture-related factors, such as religion and property relationships. The existence of such cultural restraints on political innovation constitutes in all societies a fundamental conservative bias, the implications of which have been aphoristically expressed by an English commentator, F.J.C. Hearnshaw: "It is commonly sufficient for practical purposes if conservatives, without saying anything, just sit and think, or even if they merely sit." Mere inertia, however, has rarely sufficed to protect conservative values in an age dominated by rationalist dogma and by social change related to continuous technological developments. The conservative reaction, however, is best analyzed in specific political contexts. Historians, it may be noted, cannot safely agree on there being more than four great political parties of the 20th century deserving of the name: the Conservative Party of England, the Christian Democrats of Italy and of Germany, and the Liberal Democrats of Japan.

In England, Disraeli's successor, Lord Salisbury, was prime minister in 1885, from 1886 to 1892, and from 1895 to 1902; Arthur Balfour succeeding him from 1902 to 1905. This longest era of Conservative rule was characterized by imperialism, high tariffs, and the gradual erosion of the party's working class vote, which Disraeli had so far-sightedly nurtured by extending the franchise to the workers in 1867. The party had thereby broadened its original class basis (landed aristocracy and established church) to outflank from below and above the new commercial class and its Liberal Party. It may be said that conservatism in Great Britain since Disraeli's time has veered between a passive and largely resigned acceptance of changes introduced by its Liberal and, later, Labour opponents and a more positive conservatism, the aim of which has been to foster a social environment in which the individual is encouraged to advance his own interests without undue hindrance from, or reliance on, the state —a policy descended from the liberal individualism of the 19th century, associated particularly with the Liberal Party. This positive conservatism of liberal individualism tinged with a strong sense of social conscience was given its earliest formulation by Disraeli, who combined a desire to mitigate harsh conditions suffered by the working class under conditions of unrestrained capitalism with a belief in the value of existing institutions such as the monarchy, the church, and the class system. Disraeli's foreign policy, which emphasized the need for Britain to act constructively as a "moderating and mediatorial" power and to maintain its interest in its empire, also reflected the view that conservatism must be a force shaping events rather than merely reacting to them. These three elements—the improvement of material conditions both by encouragement of individual initiative and timely reform of abuses, emphasis on the value of traditional institutions, and belief in the need for an active foreign policy—have been recurring themes of British conservatism in the 20th century. Later conservative thinkers have elaborated on the value of divergency of personality and attitudes, the role of property as an expression of individuality, and the central role of the family in providing a stable environment in which the individual may develop.

In its less positive periods (as, for example, during the interwar period), conservatism in Britain has been identified with the defense of class privileges and of the status quo, an unconstructive opposition to socialism, and, during the 1930s, a deal-making commercialist approach to the rising Nazi menace. Faced, however, with the introduction of a mixed economy and the vast extension of state welfare services by the Labour Party after 1945, the Conservatives, when returned to power in 1951, reversed very few of their socialist predecessors' innovations, emphasizing instead their claim to be more able to administer the welfare state more efficiently and to some extent outbidding their opponents, especially in areas of social policy related to their fundamental beliefs —the encouragement of a heavy program of house building being an example.

Western Europe. It is of significance that the British Conservative Party has been the more ardent of the two major British parties in championing British membership of the European Economic Community (EEC), reflecting an internationalism voiced by Sir Winston Churchill when, in 1940, he appealed for a Franco-British union and, in 1946, for a European union. Originally conceived as a means by which the economies of the European countries might be integrated—so that war between them would be impossible—the nascent community assumed significance during and after the Cold War as a means of strengthening western Europe against the threat of external Communist aggression and internal subversion. Together with the military North Atlantic Treaty Organization alliance, it thus assumed a role as a bulwark of parliamentary democracy and capitalism.

In the arena of party politics, conservatism in western Europe is generally represented by two or more parties, ranging from the liberal centre to the moderate and extreme right. Three types of party may be discerned: agrarian parties (particularly in Scandinavia), Christian democratic parties, and conservative parties linked strongly with big business interests and sometimes with a markedly nationalistic outlook. Such categories are very general and are not mutually exclusive.

The Christian democratic tradition

Among parties of the right, the Christian democratic tradition has the longest continuity, the predecessors of contemporary parties having emerged during the first half of the 19th century to represent supporters of the church and the monarchy against liberal elements. Especially after World War I, business interests became a third important element. The clerical interest is strongest in the Democrazia Cristiana (DC; Christian Democrat

Party) of Italy, which has dominated government since 1945. Through this party, Catholicism has set limits on policy concerning such church-related matters as divorce and contraception; in regard to other social questions, however, the party has never presented a coherent policy, largely because it comprises little more than an alliance of disparate and often conflicting interest groups.

In West Germany, a country divided between Catholics and Protestants, the church plays a far less significant role in the main conservative party, the Christlich-Demokratische Union (CDU; the Christian Democratic Union). After 1950, following debate within the party over economic and social questions, advocacy of a free-enterprise economy coupled with a strong commitment to maintain and improve social insurance and other welfare provisions became established policy. The conservative temper of the political climate in Germany since the beginning of economic recovery may be judged from the fact that since the early 1950s the main opposition party, the Sozialdemokratische Partei Deutschlands (SPD; the Social Democrats), has progressively eliminated the socialist content of its program, a congress at Bad Godesberg (1959) in fact going so far as to champion the profit motive.

France provides an exception to the general pattern of the representation of moderate conservative opinion by a Christian democratic party; the closest equivalent has been the Catholic, right-wing, Mouvement Républicain Populaire, moribund in the early 1970s. Instead, a large proportion of conservatives in France have supported Gaullist groups such as the Union pour la Défense de la République. Gaullist conservatism has been markedly nationalistic, involving assumptions concerning French leadership of a united Europe and emphasizing tradition, order, and the regeneration of France. Gaullists espouse divergent views on domestic social issues, however, as do non-Gaullist groups such as the Centre National des Indepéndants et Paysans. The number of conservative groups, their lack of stability, and their tendency to be identified with local issues defy simple categorization. Conservatism in France, however, as in Italy and Germany, has been the dominant political force since World War II.

Conservatism in Europe is thus revealed as a dominating political influence in the major states, finding expression in parties of very different character. These parties represent traditional bourgeois values and oppose unnecessary state involvement in economic affairs and any radical attempts at income redistribution. They are also characterized by an absence of ideology and often of even a well-articulated political philosophy, but this tends to be of little consequence in terms of their influence since they give political expression to the conservatism of temperament mentioned above as an important underlying bias in political conflict, as well as to persistent culture-related values that are of great importance in terms of continuity and stability.

Japan. The relationship between conservatism as an underlying bias related to psychological factors and cultural values and conservatism as an articulated political credo is illustrated by the history of party politics in Japan since its opening to Western influence in the middle of the 19th century. The political and social changes that took place following the Meiji Restoration (1868) were of major proportions, involving the abolition of feudal institutions and the introduction of such Western political ideas as constitutional government. But despite institutional innovations and the dislocations resulting from rapid industrialization, traditional loyalties and attitudes proved to be more important factors in shaping political developments.

Except for the period of intervention by the militarists during the 1930s and 1940s, Japan has been ruled by conservatives since the beginning of party politics in the 1880s. The conservative parties (the two most important of which merged to form the Liberal-Democratic Party in 1955) have been dominated by personalities rather than by ideology and dogma; and personal loyalties to leaders of groups within the party (factions) rather than commitment to policy have determined the allegiance of conservative members of the Diet. As one American scholar, Nathaniel B. Thayer, has described it, the factions

> have adopted the social values, customs, and relationships of an older Japan. . . . The old concepts of loyalty, hierarchy, and duty hold sway in them. And the Dietman (or any other Japanese) feels very comfortable when he steps into this world.

The Liberal-Democratic Party is intimately linked with big business interests, and its policies are guided primarily by the objective of fostering a stable environment for the development of Japan's free-enterprise economy; to this end, the party functions as a broker of conflicting business interests. Policy toward other Asian countries, national defense, and internal security are other conservative preoccupations.

The United States. It may be argued that the United States has no nationwide conservative or liberal parties but instead only two fluctuating, all-inclusive coalitions. Both the Democrat and Republican coalitions include interest groups occasionally labelled conservative—the racists among southern Democrats, for example, and such Republican off-shoots as the local New York Conservative Party. On a journalistic level the word conservative has been used loosely for a segment of the Republican party associated with Sen. Barry Goldwater and a governor of California, Ronald Reagan. A more historical approach, however, applies the more precise label Manchester liberal to so unconservative an outlook—so laissez-faire and so untraditional.

Modern American conservatism has been most influential not in the unphilosophic realm of politics but in the literary and religious realm in such masterpieces of conservative outlook as Irving Babbitt's *Democracy and Leadership* (1924) or the aristocratic traditionalism of the Nobel Prize-winning novelist William Faulkner. Indeed, such figures as the novelist Herman Melville and the theologian Reinhold Niebuhr, usually independent of each other and eschewing conservative labels, have performed the nation's spiritual arithmetic, calculating the spiritual price of material progress and of a robotizing technology. Unconsciously conservative in this sense, even when under radical slogans, is the impulse among young people in the 1970s to conserve ecology and environment against what Melville called "the impieties of progress." These unconscious young conservers sublimate the old class-based elitism into a new value-based elitism, open to all: thereby rescuing quality (the cultural as well as physical ecology) from the parvenu plutocrats of quantity (mass culture and robot technology).

The impact of the horrors allowed at Auschwitz has purged—in effect conservatized—many modern liberals out of their most unconservative axiom: the Rousseauist doctrine of the "natural goodness" of man and the masses. For many, the real battle for the future now seemed to be an alliance of such chastened liberals with conservatives in jointly defending their shared constitutional and ethical framework against extremist destroyers from a mirror-image right and left.

It is arguable that conservatism, whether its influence operates through political parties or through psychological, cultural, and institutional factors, is a far more persuasive influence in democratic societies than the rate of social and economic change and the welter of rationalist dogma would suggest. That it is often lacking in articulation and that, as critics of conservatism point out, there is a comparative lack of persuasive presentations of the conservative cause compared with the abundant literature of rationalist politics is in part a consequence of its underlying strength and in part a result of a certain coyness among the best conservative thinkers deriving from the fear that a conservatism that needs to present itself in the same terms as the doctrines it opposes is no longer conservatism or is a conservatism in retreat. In the latter part of the 20th century, however, many would say that it may be argued that the predilection of governments to extend their role in social life is so strong as to necessitate a more articulate, even ag-

gressive, conservatism. One particularly important task of conservatives will be to emphasize that the social sciences, particularly anthropology and psychology, so long enlisted in the cause of social engineering and liberal utopianism, also reveal much about the role of tradition, custom, and evolution in the survival of societies.

BIBLIOGRAPHY. Among "Burkean-conservative" works sharing an anti-extremist centre with moderate liberals are DANIEL J. BOORSTIN, *The Genius of American Politics* (1953); THOMAS I. COOK and MALCOLM MOOS, *Power Through Purpose* (1954); ERIC HOFFER, *The True Believer* (1951); ROSS J.S. HOFFMAN and PAUL LEVACK (eds.), *Burke's Politics* (1949); HENRY A. KISSINGER, *A World Restored* (1957); WALTER LIPPMANN, *The Cold War* (1947); REINHOLD NIEBUHR, *The Irony of American History* (1952), *Christian Realism and Political Problems* (1953), and *The Self and the Dramas of History* (1955); ROBERT A. NISBET, *Community and Power* (1962); CLINTON L. ROSSITER, *Conservatism in America*, 2nd ed. rev. (1962); GEORGE SANTAYANA, *Dominations and Powers* (1951); LEO STRAUSS, *Natural Right and History* (1953) and *What Is Political Philosophy? and Other Studies* (1959); FRANK TANNENBAUM, *A Philosophy of Labor* (1951); PETER VIERECK, *Conservatism Revisited* (1949; paperback rev. ed., 1965), *The Unadjusted Man* (1956), and *Shame and Glory of the Intellectuals* (1953; paperback rev. ed., 1965); ERIC VOEGELIN, *The New Science of Politics* (1952); and *Order and History*, 3 vol. (1956–58); FRANCIS G. WILSON, *The Case for Conservatism* (1951). For British Conservatives, see LEOPOLD AMERY, *The Forward View* (1935); ARTHUR BRYANT, *The Spirit of Conservatism* (1929); LORD HUGH CECIL, *Conservatism* (1912); HENRY FAIRLIE, *The Life of Politics* (1968); QUINTIN HOGG, *The Conservative Case*, rev. ed. (1959); F.J.C. HEARNSHAW, *Conservatism in England* (1933); MICHAEL J. OAKESHOTT, *Rationalism in Politics, and Other Essays* (1962); and PEREGRINE WORSTHORNE, *The Socialist Myth* (1971).

More extreme views, whether right-wing nationalist in politics or militantly business-oriented in economics (mainly by contributors to the contemporary New York periodical *National Review*), may be found in WILLIAM F. BUCKLEY, JR., *God and Man at Yale: The Superstitions of Academic Freedom* (1951) and *Up from Liberalism*, rev. ed. (1968); JAMES BURNHAM, *Congress and the American Tradition* (1959) and *Suicide of the West* (1964); MILTON FRIEDMAN, *Capitalism and Freedom* (1962); BARRY GOLDWATER, *Conscience of a Conservative* (1963); JEFFREY P. HART, *The American Dissent* (1966); NELLIE D. KENDALL (ed.), *Willmoore Kendall contra mundum* (1971); RUSSELL KIRK, *The Conservative Mind, from Burke to Santayana* (1953), *Prospects for Conservatives* (1956), and *Enemies of the Permanent Things* (1969); ERIK VON KUEHNELT-LEDDIHN, *Liberty or Equality* (1952); FRANK S. MEYER, *In Defense of Freedom* (1962) and *The Conservative Mainstream* (1969); THOMAS S. MOLNAR, *The Counter-Revolution* (1969); ENOCH POWELL, *Freedom and Reality* 1969); RONALD REAGAN, *The Creative Society* (1968).

Useful anthologies include PETER VIERECK (ed.), *Conservatism: From John Adams to Churchill* (paperback rev. ed. 1956); and PETER WITONSKI (ed.), *The Wisdom of Conservatism* (1971).

(Pe.V.)

Constable, John

John Constable, with J.M.W. Turner, dominated English landscape painting in the 19th century. But the works of these two contemporaries are widely different in aim and effect. Turner never lost the ambition to fuse landscape with literary, mythological, or historical overtones. He loved to represent the most dynamic forces of nature and in his later works evolved pictorial drama out of the sheer element of light and colour. In direct contrast, Constable's aim was to represent the English countryside with as much truth to natural appearances as he could achieve from his long study of the scenes familiar to him and the varying aspects of the sky. He was thus the last great exponent of the tradition of naturalistic landscape painting fostered by the 17th-century Dutch artists.

Constable was born in East Bergholt, Suffolk, on June 11, 1776. His birthplace was, and remains, a small village, standing on a ridge a short distance from the River Stour, which separates Suffolk from Essex. The Stour Valley in this region is rich in corn, pastureland, and fine trees and was known in the late 18th century for its efficient agriculture no less than its natural beauty. The men of Suffolk felt a jealous patriotism for their own county, and Constable remained at heart a Suffolk man, although



Constable, self-portrait, drawing in pencil and watercolour. In the National Portrait Gallery, London. 25.4 cm × 20.32 cm.
By courtesy of the National Portrait Gallery, London

he constantly crossed the bridge over the River Stour at Flatford into Essex.

The artist's father, Golding Constable, was a man of means who owned mills at Flatford and Dedham, on the Suffolk and Essex banks of the Stour, respectively. His business consisted of grinding corn raised in the local fields and shipping it round the coast of East Anglia to the London market. The Stour had been made into a canal, navigable beyond these mills, and the grain was transported on its waters in broad, flat-bottomed barges. The fact that Constable was born into the midst of the practical realities of country life has a direct bearing on his career and is reflected throughout his painting. He showed intellectual promise as a child and was brought up for the church; when this idea was abandoned he was trained to enter his father's business. But he had already conceived an enthusiasm for painting. This was fostered by his friendship with an amateur painter, John Dunthorne, the local plumber and glazier, and further encouraged by Sir George Beaumont. Constable's determination to make painting his profession was sealed by his acceptance as a probationer in the Royal Academy Schools in 1799, when he was 23.

Early works

Artistic development. At this time his performance did not reveal any marked promise; his execution was laboured and his drawing from the life weakly academic. But he already had a clear mental image of the type of pictures he wanted to paint and worked doggedly to overcome his technical defects. Seven or eight years after he had started his formal training, he discovered how to embody his idea of the English countryside in a manner both more realistic and more spirited than his predecessors. There were some modest successes to record in this period of self-training. He exhibited at the Royal Academy shows annually from 1802, with one single exception in 1804. He went on two of the sketching expeditions that it was then the practice for landscape painters to undertake, going to the Peak District, Derbyshire, in 1801 and the Lake District in 1806. He painted portraits of the Suffolk and Essex farmers and their wives and in 1805 attempted an altarpiece of "Christ Blessing the Children," in the manner of the American expatriate painter Benjamin West. But when he took stock of his progress after his return from the Lake District, he realized that he had been attempting too wide a range of subject and style, thus dissipating his energies. He then determined to concentrate on the scenes that had delighted him as a boy: the village lanes, the cornfields and meadows running down to the Stour, the slow progress of barges drawn by tow horses, the bustle of vessels passing the locks at Flatford or Dedham.

In the years 1809 to 1816 he established his mastery and evolved his individual manner; but these were years of personal stress. He was obliged to live much of each

year in London, where his professional associates were to be found and where he could participate in exhibitions. But he was uneasy at these enforced absences from the countryside, in which he felt most at home, and tried to pay yearly visits to Suffolk. The assiduity with which he studied the landscape on these visits is shown by two pocket sketchbooks, one of 1813 and one of 1814, which, still intact, are preserved in the Victoria and Albert Museum. These contain between them over 200 small sketches made in a limited area around his home village and reflect most aspects of the summer life of the fields and the river.

Years of personal stress

Deeper than the strain of exile from these scenes was the unhappy progress of his courtship of Maria Bicknell, with whom he had fallen in love in 1809 but whose grandfather, the elderly and tyrannical rector of East Bergholt, opposed her marriage to an impecunious artist. Nevertheless, Constable stuck to his purpose with a tenacity equal to that which he displayed in his art, and, in her unaggressive way, Maria was just as determined. A further anxiety for Constable came from the failing health of his parents; his mother died in 1815, and his father the following year. He was genuinely devoted to them and spent prolonged periods at home during their illnesses. But his father's death in 1816 provided a sufficient measure of economic independence for him to marry Maria Bicknell and to settle into the domestic life that was a prerequisite for his calm development and the full maturing of his art.

Once he had married, on October 2, 1816, and had established himself and his wife in a London home, Constable set to work to show what he could achieve in his art. He was 40 years old, and had painted a handful of accomplished pictures, which were original but on a small scale. These included "Dedham Vale: Morning," (1811; Sir Richard Proby Collection, Elton Hall, Huntingdonshire); "Boatbuilding near Flatford Mill," (1815; Victoria and Albert Museum, London); "The Stour Valley and Dedham Village," (1815; Museum of Fine Arts, Boston). But these were still products of the years of preparation. Most significant was the large number of small oil sketches and drawings that were to form the basis of his future and more ambitious painting. These sketches, of which he made a considerable number after 1808, were painted in the open air in front of the subject. They are most frequently in oils on paper about 12 inches wide, and they record the form of the landscape, the colours that predominate and also the more evanescent qualities of atmosphere and the reflection of light on particular details. Now they are recognized to be among Constable's most individual achievements and to have been unique at the time they were painted. But to the artist they were means to an end. His main ambition was to embody his concept of the Suffolk countryside in a series of larger canvases monumental enough to make an impression in the annual summer exhibitions of the Royal Academy. The first attempt was the "Flatford Mill on the River Stour" which he exhibited in 1817. It shows a reach of the river running up to the mill, in which Golding Constable had lived till within two years of Constable's birth, bordered by a meadow that has just been scythed.

Mature works. This was succeeded by a series of six paintings that are now among his best known and most highly regarded works. In order of exhibition they are "The White Horse"; "Stratford Mill"; "The Hay-Wain"; "View on the Stour near Dedham"; "The Lock"; "The Leaping Horse." These six canvases portray scenes on the River Stour that were easily within the compass of Constable's childhood walks; between the most easterly, "The Hay-Wain," and the most westerly, "Stratford Mill," there is hardly more than two miles distance in a direct line. To this unity of place is joined a unity of subject matter. With the exception of "The Hay-Wain," all show barges being manoeuvred along the canals. The appearance in these works of the fruits of Constable's deep, unprecedented study of the formation of clouds, the colour of meadows and trees, and the effect of light glistening on leaves and water enables them to communicate the concrete actuality of these everyday-life country scenes, as well as the feeling they evoked in him.

This series of Stour scenes was interrupted in 1823, when Constable's chief exhibit was a view of "Salisbury Cathedral from the Bishop's Grounds," in which the artist transmuted a commission from Bishop Fisher, intended mainly as a record of an architectural monument, into his own idiom, framing the spire between overarching trees, emphasizing the play of light and shade on the Gothic stonework, and setting the whole under a sky in which rain clouds are impending. This romantic treatment did not please the Bishop but was admired by the Bishop's nephew and Constable's old friend, Archdeacon John Fisher, who had already shown his faith in the artist by buying "The White Horse" at the exhibition of 1819.

A revealing correspondence between Constable and Fisher has been preserved. In it the painter gives his most intimate thoughts on his art without concealment or false modesty. There was much he could be satisfied with at this time. He was aware that he had achieved in his art a great deal of what he had set out to do. In addition, his work had deeply impressed the painters of the French Romantic school. Théodore Géricault had admired "The Hay-Wain" on its first exhibition in 1821; and when this work (along with the "View on the Stour near Dedham") was shown at the Paris Salon in 1824, it not only created a sensation but inspired Eugène Delacroix to repaint parts of his "Massacre at Chios." In England recognition was slower in coming. Although Constable had been made an Associate of the Royal Academy in 1819, full membership was delayed for ten years.

Effect of personal grief on his paintings

Meanwhile the presence, from 1819, of Hampstead scenes and, from 1824, of Brighton scenes among his repertoire of subjects indicates a deepening shadow over his domestic happiness. Mrs. Constable had long been delicate, and Constable took houses in these places in search of purer air. Her death from consumption in 1828, at the age of 41, was a loss from which he never fully recovered, though he bestirred himself into activity for the sake of his seven children, in whom he delighted. His financial situation had been eased by a large legacy from his father-in-law, and he became a full Academician in 1829. But from this time an increased restlessness is to be found in his paintings. "Hadleigh Castle" and "Salisbury Cathedral from the Meadows" show his growing recourse to broken accents of colour, sombre tones, and stormy skies. It was in 1829 also that he began his preparations for the publication of *English Landscape Scenery*, a selection of mezzotints executed by David Lucas from Constable's paintings and sketches in which the same dramatic qualities of light and shade are translated into a black-and-white medium. The admiration of his friend, the American-born artist C.R. Leslie, prompted the writing of the *Memoirs of the Life of John Constable, R.A.* This biography was first published in 1843 and still remains an indispensable source of information on Constable.

Late works

In the 1820s the use of colour by Constable's great contemporary and rival in landscape painting, J.M.W. Turner, was becoming bolder and even more uninhibited. This may have contributed to the greater readiness for change that we see in Constable's late works. His "Waterloo Bridge from Whitehall Stairs" is a monumental record of the opening ceremonial, painted in a high key of colour. His use of watercolour became more frequent, and in 1834, after he had been seriously ill, he sent no oils at all to the Royal Academy, depending for his principal exhibit on a large and remarkable watercolour, "Old Sarum" (Victoria and Albert Museum, London). A visit to Arundel in the same summer imbued him with enthusiasm for a new type of countryside dominated by steep wooded slopes.

In 1836 Constable sent "The Cenotaph at Coleorton" to the Royal Academy exhibition. It was the last painting he showed in his lifetime. He died suddenly during the night of March 31, 1837. The painting on which he had been working the day before, "Arundel Mill and Castle" (Toledo Museum of Art, Toledo, Ohio), was suf-

ficiently completed to be shown posthumously at the Academy. At his death his reputation was limited, but those who admired his work did so intensely. This admiration grew slowly throughout the 19th century, becoming more widespread as his sketches became available and their freshness and spontaneity were recognized. In 1843 his first biographer, C.R. Leslie, wrote that he was "the most genuine painter of English landscape," and that is a judgment now almost universally reaffirmed.

MAJOR WORKS

"Malvern Hall, Warwickshire" (1809?; Tate Gallery, London); "Boatbuilding near Flatford Mill" (1815; Victoria and Albert Museum, London); "Wivenhoe Park, Essex" (1816; National Gallery of Art, Washington, D.C.); "Flatford Mill on the River Stour" (1817; Tate Gallery); "The White Horse" (1819; Frick Collection, New York); "Dedham Lock and Mill" (1820; Victoria and Albert Museum); "Stratford Mill" (1820; Sir Reginald and Lady Macdonald-Buchanan Collection, Cottesbrooke Hall, Northamptonshire); "The Hay-Wain" (1821; National Gallery, London); "View on the Stour near Dedham" (1822; Huntington Library and Art Gallery, San Marino, California); "Salisbury Cathedral from the Bishop's Grounds" (1823; Victoria and Albert Museum); "The Lock" (1824; private collection, England); "The Leaping Horse" (1825; Royal Academy of Arts, London); "The Cornfield" (1826; National Gallery, London); "Chain Pier, Brighton" (1827; Tate Gallery); "Dedham Vale" (1828; National Gallery of Scotland, Edinburgh); "Hampstead Heath: Branch Hill Pond" (1828; Victoria and Albert Museum); "Hadleigh Castle" (1829; Mr. and Mrs. Paul Mellon Collection, Virginia); "Salisbury Cathedral from the Meadows" (1831; Lord Ashton of Hyde Collection, Moreton-in-Marsh, Gloucestershire); "The Grove (or Admiral's House) at Hampstead" (1832; National Gallery, London); "Waterloo Bridge from Whitehall Stairs" (1832; private collection, England); "A Cottage at East Bergholt" (*c.* 1835; Lady Lever Art Gallery, Port Sunlight, Cheshire); "The Cenotaph at Coleorton" (1836; Tate Gallery); "Stoke-by-Nayland" (after 1830; Art Institute, Chicago).

Pencil and oil sketches exist for many paintings, and the largest collection of these is in the Victoria and Albert Museum. This includes the full-scale sketches for "The Hay-Wain" (1821) and "The Leaping Horse" (1825); among the 92 oil sketches painted on a small scale in the open air two of the best known are "Barges on the Stour" and "Flatford Lock and Mill" (both *c.* 1810).

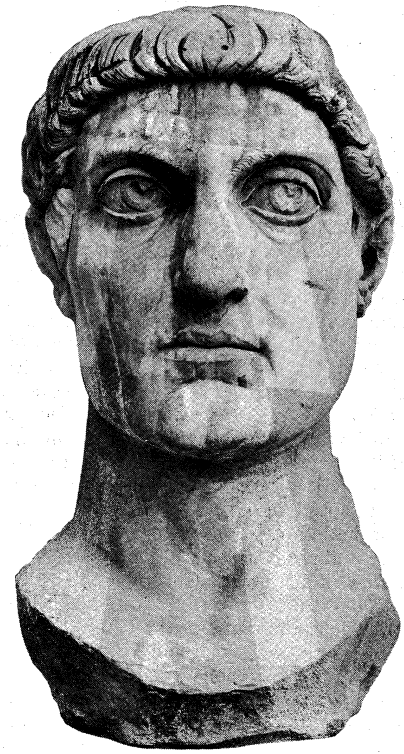
BIBLIOGRAPHY. C.R. LESLIE (ed.), *Memoirs of the Life of John Constable, Composed Chiefly of His Letters*, rev. ed. (1951), is written with affectionate understanding of its subject and with narrative power. R.B. BECKETT, *John Constable's Correspondence*, 6 vol. (1962–68), contains a transcription of all the known letters to and from Constable, with an explanatory account. These volumes, including the companion volume *John Constable's Discourses* (1970), are invaluable to the student. G. REYNOLDS, *Constable: The Natural Painter* (1965), is an up-to-date summary of the artist's life and work.

(G.R.)

Constantine the Great

Constantine I the Great, the first Roman emperor to profess Christianity, initiated not only the evolution of the empire into a Christian state but also provided the impulse for a distinctively Christian culture that prepared the way for the growth of Byzantine and Western medieval culture. He was born on February 27 of an unknown year, but probably in the later AD 280s, at Naissus (modern Niš in Yugoslavia) in the province of Upper Moesia, on the strategic road leading from Pannonia through Sirmium and Singidunum (Belgrade) to Byzantium. He was a typical product of the military governing class of the later 3rd century, the son of Flavius Valerius Constantius, an army officer, and his wife (or concubine) Helena. In AD 293, when Constantine was still a boy, his father was raised to the rank of Caesar, or deputy emperor, and was sent to serve under the Augustus (emperor) Maximian in the west. At the same time he had to separate from Helena in order to marry a stepdaughter of Maximian; and Constantine was brought up in the Eastern Empire, at the court of the senior emperor Diocletian at Nicomedia (modern İzmit in Turkey). Constantine was seen as a youth by his future panegyrist, Eusebius, bishop of Caesarea, passing with Diocletian through Palestine on the way to a war in Egypt; later, as a young officer, Constantine took part in a successful campaign on the lower Danube.

Background



Constantine the Great, colossal marble head, *c.* AD 330. In the Capitoline Museum, Rome.
By courtesy of the British School, Rome

Career and conversion. Constantine's experience as a member of the imperial court—a Latin-speaking institution—in the eastern provinces left a lasting imprint on him. Educated to less than the highest literary standards of the day, he was always more at home in Latin than in Greek: later in life he was in the habit of delivering edifying sermons, which he would compose in Latin and pronounce in Greek from professional translations. Christianity he encountered in court circles as well as in the cities of the east; while from 303, during the great persecution of the Christians that began at the court of Diocletian at Nicomedia and was enforced with particular intensity in the eastern parts of the empire, Christianity was a major issue of public policy. It is even possible that members of Constantine's family were Christians. Constantine himself was said to have converted his mother: his father's conduct in Britain during the persecution is uncertain; but the name of a half-sister of Constantine, Anastasia, has been thought to show Christian influence.

In 305 the two emperors, Diocletian and Maximian, resigned, to be succeeded by their respective deputy emperors, Galerius and Constantius. The latter were replaced by Maximinus Daia in the east and Severus in the west, Constantine being passed over. Constantius now requested his son's presence from Galerius: this was grudgingly conceded, and Constantine made his way through the territories of the hostile Severus to join his father at Boulogne. They then crossed together to Britain and fought a campaign in the north before Constantius' death at York in 306. Immediately acclaimed emperor by the army, Constantine now threw himself into a complex series of civil wars in which Maxentius, the son of the old Western emperor Maximian, rebelled at Rome, with his father's help suppressing Severus, the deputy emperor, who was replaced by Licinius. When Maximian was rejected by his son, he joined Constantine in Gaul, only to betray him and be forced to commit suicide (310). Constantine, who had in 307 married Maximian's daughter Fausta as his second wife, invaded Italy in 312 and after a lightning campaign defeated his brother-in-law Maxentius at the Milvian Bridge near Rome. He then

confirmed an alliance that he had already entered into with Licinius (Galerius having died in 311). Licinius, after defeating his rival Maximinus Daia, became Eastern emperor but lost territory in the Balkans to Constantine in 316. After a further period of tension, Constantine attacked Licinius in a second war of 324, routing him at Adrianople and Chrysopolis and becoming sole emperor until his death in 337.

Throughout his life, Constantine ascribed his success to his conversion to Christianity and the support of the Christian God. The triumphal arch erected in his honour at Rome after the defeat of Maxentius ascribed the victory to the "inspiration of the Divinity" as well as to Constantine's own genius. A statue set up at the same time showed Constantine himself, holding aloft a cross and the legend, "by this saving sign I have delivered your city from the tyrant and restored liberty to the Senate and people of Rome." After his victory over Licinius in 324, Constantine wrote that he had come from the farthest shores of Britain as God's chosen instrument for the suppression of impiety, and in a letter to the Persian king Shāpūr II he proclaimed that, aided by the divine power of God, he had come from the borders of the ocean to bring peace and prosperity to all lands.

Constantine's adherence to Christianity was closely associated with his rise to power. He fought the Battle of the Milvian Bridge in the name of the Christian God, having received instructions in a dream to paint the Christian monogram (☧) on his troops' shields. This is the account given by the Christian apologist Lactantius; a somewhat different version, offered by Eusebius, tells of a vision seen by Constantine during the campaign against Maxentius, in which the Christian sign appeared in the sky with the legend, "In this sign, conquer." Despite the Emperor's own authority for the account, given late in life to Eusebius, it contains anachronisms and is in general more problematic than the other: but a religious experience on the march from Gaul is suggested also by a pagan orator, who in a speech of 310 referred to a vision of Apollo received by Constantine at a shrine in Gaul.

"In this sign, conquer"

Yet to suggest that Constantine's conversion was "politically motivated" means little in an age in which every Greek or Roman expected that political success followed from religious piety. The civil war itself fostered religious competition, each side enlisting its divine support; and it would be thought in no way unusual that Constantine should have sought divine help for his claim for power and divine justification for his acquisition of it. What is far more remarkable is Constantine's subsequent development of his new religious allegiance to a quite extreme personal commitment.

Commitment to Christianity. After the defeat of Maxentius, Constantine met Licinius at Milan to confirm a number of political and dynastic arrangements. A product of this meeting was the so-called Edict of Milan, extending toleration to the Christians and the restoration after the persecution of their personal and corporate property. The extant copies of this decree are actually those posted by Licinius in the eastern parts of the empire. But Constantine went far beyond the joint policy agreed upon at Milan. By 313 he had already donated to the Bishop of Rome the imperial property of the Lateran, where a new cathedral, the Basilica Constantiniana (now S. Giovanni in Laterano) soon rose. The Church of St. Sebastian was also probably begun at this time: and it was in these early years of his reign that Constantine began issuing laws conveying upon the church and its clergy fiscal and legal privileges and immunities from civic burdens. As he said in a letter of 313 to the proconsul of Africa, the Christian clergy should not be distracted by secular offices from their religious duties ". . . for when they are free to render supreme service to the Divinity, it is evident that they confer great benefit upon the affairs of state." In another such letter, to the bishop of Carthage, Constantine mentioned the Spanish bishop Hosius, important later in the reign as his adviser and possibly—since he may well have been with Constantine in Gaul before the campaign against Maxentius—instrumental in the conversion of the Emperor.

Public and personal "theology"

Constantine's personal "theology" emerges with particular clarity from a remarkable series of letters, extending from 313 to the early 320s, concerning the Donatist schism in North Africa. The Donatists maintained that those priests and bishops who had once lapsed from the Christian faith could not be readmitted to the church. Constantine's chief concern was that a divided church would offend the Christian God and so bring divine vengeance upon the Roman Empire and Constantine himself. Schism, in Constantine's view, was "insane, futile madness," inspired by the Devil, the author of evil. Its partisans were acting in defiance of the clemency of Christ, for which they might expect eternal damnation at the Last Judgment (this was a Judgment whose rigours Constantine equally anticipated for himself). Meanwhile, it was for the righteous members of the Christian community to show patience and long-suffering. In so doing they would be imitating Christ, and their patience would be rewarded in lieu of martyrdom—for actual martyrdom was, as Constantine observed, no longer open to Christians in a time of peace for the church. Throughout, Constantine had no doubt whatever that to remove error and propagate the true religion was both his personal duty and a proper use of the imperial position.

Such pronouncements, expressed in letters to imperial officials and to Christian clergy, make untenable the view that Constantine's religious attitudes were even in these early years either veiled, confused, or compromised. Openly expressed, his attitudes show a clear commitment.

Constantine's second involvement in an ecclesiastical issue followed the defeat of Licinius as promptly as the involvement in Donatism followed that of Maxentius; but the Arian heresy, with its intricate explorations, couched in difficult Greek, of the precise nature of the Trinity, was as remote from Constantine's educational background as it was from his impatient, urgent temperament. The Council of Nicaea, which opened in May 325 with an address by the Emperor, had already been preceded by a letter to the chief protagonist, Arius of Alexandria, in which Constantine stated his opinion that the dispute was fostered only by excessive leisure and academic contention, that the point at issue was trivial and could be resolved without difficulty. His optimism was not justified: neither this letter nor, despite its subsequent authority, the Council of Nicaea itself, nor the second letter, in which Constantine urged acceptance of its conclusions, was adequate to solve a dispute in which the participants were as intransigent as the theological issues were subtle. Indeed, for 40 years after the death of Constantine, Arianism was actually the official orthodoxy of the Eastern Empire.

Arianism and Council of Nicaea

The Council of Nicaea coincided almost exactly with the celebrations of the 20th anniversary of the reign of Constantine, at which, returning the compliment paid by the Emperor's attendance at their council, the bishops were honoured participants. But Constantine's visit to the West in 326, to repeat the celebrations at Rome, brought the greatest political crisis of the reign. During his absence from the East, for reasons that remain obscure, Constantine had his eldest son, the deputy emperor Crispus, and his own wife Fausta, Crispus' stepmother, slain. Nor was the visit to Rome a success. Constantine's refusal to take part in a pagan procession offended the Romans; and when he left after a short visit, it was never to return.

These events set the course of the last phase of the reign of Constantine. Already after his defeat of Licinius he had renamed Byzantium as Constantinople: immediately upon his return from the West he began to rebuild the city on a greatly enlarged pattern, as his permanent capital and the "second Rome." The dedication of Constantinople, in May 330, effectively confirmed the divorce, which had been in the making for over a century, between the emperors and Rome, the traditional capital of the empire. Rome had for long been unsuited to the strategic needs of the empire: it was now to be left in splendid isolation, as an enormously wealthy and prestigious city —still, as time would show, the emotional focus of the empire, but of limited political importance.

Founding of Constantinople

It was perhaps in some sense to atone for the family catastrophe of 326 that Constantine's mother, Helena, embarked soon afterward on a pilgrimage to the Holy Land. Her journey was attended by almsgiving and pious works; above all, it was distinguished by her church foundations, on the Mount of Olives at Jerusalem and at the Cave of the Nativity at Bethlehem. By the initiative of another lady of the imperial house, Constantine's mother-in-law Eutropia, a church was also built at Mamre, where, according to an interpretation of Genesis shared by Constantine and Eusebius, Christ had first shown himself to men in God's appearance to Abraham; but the most famous of these foundations followed the sensational discovery of the Holy Sepulchre at Jerusalem. The discovery was taken up with enthusiasm by Constantine. In a letter to Macarius, bishop of Jerusalem, the Emperor instigated the building of a great new basilica at the spot, offering unlimited help with labour and materials and personal suggestions as to design and decoration.

Constantine's interest in church building was expressed also at Constantinople, particularly in churches of the Holy Wisdom (the original Hagia Sophia) and of the Apostles. At Rome, the great church of St. Peter was begun in the later 320s and lavishly endowed by Constantine with plate and property. Meanwhile, churches at Trier, Aquileia, Cirta in Numidia, Nicomedia, Antioch, Gaza, Alexandria, and elsewhere owed their development, directly or indirectly, to Constantine's interest.

The Emperor was always an earnest student of his religion and spent hours discussing it with bishops. Even before the defeat of Licinius he had summoned to Trier the aged theologian and polemicist Lactantius, to be the tutor of Crispus. In later years, he wrote to Eusebius to commission new copies of the Bible for the use of the growing congregations at Constantinople. He composed a special prayer for his troops and went on campaign equipped with a mobile chapel in a tent. He issued numerous laws relating to Christian practice and susceptibilities: for instance, abolishing the penalty of crucifixion and the practice of branding certain criminals, "so as not to disfigure the human face, which is formed in the image of divine beauty"; enjoining the observance of Sunday and saints' days; extending privileges to the clergy while suppressing at least some of the more offensive pagan practices.

Final years

Constantine had hoped to be baptized in the River Jordan, but perhaps because of the lack of opportunity to do so—together no doubt with the reflection that his office necessarily involved responsibility for actions hardly compatible with the baptized state—delayed the ceremony until the end of his life. It was while preparing for a campaign against Persia that he fell ill at Helenopolis. When treatment failed, he made to return to Constantinople but was forced to take to his bed near Nicomedia. There, Constantine received baptism, putting off the imperial purple for the white robes of a neophyte; and he died on May 22, 337. He was buried at Constantinople in his Church of the Apostles, whose memorials, six on each side, flanked his tomb. Yet this was less an expression of religious megalomania than of Constantine's literal conviction that he was, in a quite precise sense, the successor of the evangelists, having devoted his life and office to the spreading of Christianity.

Assessment. The reign of Constantine must be interpreted against the background of his clear and unambiguous personal commitment to Christianity. This is not to say that his public actions and policies were entirely without ambiguity. Roman opinion expected of its emperors not innovation or revolution but the preservation of traditional ways; Roman media of propaganda and political communication were conditioned, by statement, allusion, and symbol, to express these expectations. It is significant, for instance, not that the pagan gods and their legends survived for a few years on Constantine's coinage but that they disappeared so quickly: the last of them, the relatively inoffensive "Unconquered Sun" had been eliminated within little over a decade after the defeat of Maxentius.

Some of the ambiguities in Constantine's public policies were therefore exacted by the respect due to established practice and by the difficulties of expressing, as well as of making, total changes suddenly. The suppression of paganism, by law and by the sporadic destruction of pagan shrines, is balanced by particular acts of deference, such as the permission given in 326 to a pagan Athenian to use the imperial transport service to visit the Valley of the Tombs of the Kings in Egypt (a traditional centre of pagan "pilgrimage"). A town in Asia Minor mentioned the unanimous Christianity of its inhabitants in support of a petition to the Emperor; while, on the other hand, one in Italy was allowed to hold a local festival incorporating gladiatorial games and to found a shrine of the imperial dynasty—although direct religious observance there was firmly forbidden. In an early law of Constantine, priests and public soothsayers of Rome were prohibited entry to private houses; but another law, of 320 or 321, calls for their recital of prayer, "in the manner of ancient observance," if the imperial palace or any other public building were struck by lightning. Traditional country magic was tolerated by Constantine as salutary in object and inoffensive in practice. Classical culture and education, which were intimately linked with paganism, continued to enjoy enormous prestige and influence; provincial priesthoods, which were as intimately linked with civic life, long survived the reign of Constantine. Constantinople itself was predominantly a Christian city, its dedication celebrated by Christian services; yet its foundation was also attended by a famous pagan seer, Sopatros, who seems to have been requested to devise rites for the dedication of the new city.

Secular achievements

An objective assessment of Constantine's secular achievements is not easy—partly because of the predominantly religious significance with which the Emperor himself invested his reign, partly because the restlessly innovatory character that dissenting contemporaries saw in his religious policy was also applied by them to the interpretation of his secular achievement. Some of Constantine's contributions can, in fact, be argued to have been already implicit in the trends of the last half century. So may be judged the further development, taking place in his reign, of the administrative court hierarchy and an increasing reliance upon a mobile field army, to what was considered the detriment of frontier garrisons. The establishment by Constantine of a new gold coin, the *solidus*, which was to survive for centuries as the basic unit of Byzantine currency could hardly have been achieved without the work of his predecessors in restoring political and military stability after the anarchy of the 3rd century. Perhaps more directly linked with Constantine's own political and dynastic policies was the emergence of regional praetorian prefectures with supreme authority over civil financial administration but with no direct control over military affairs: this they yielded to new *magistri*, or "masters," of the cavalry and infantry forces. The reduction of the prefects' powers was seen by some as excessively innovatory; but the principle of the division of military and civil power had already been established by Diocletian. A real innovation, from which Constantine could expect little popularity, was his institution of a new tax, the *collatio lustralis*. It was levied every five years upon trade and business and seems to have become genuinely oppressive.

A lavish spender, Constantine was notoriously openhanded to his supporters and was accused of promoting beyond their deserts men of inferior social status. Yet he was not the first emperor to incur this criticism. More to the point is the accusation that his generosity was only made possible by his looting of the treasures of the pagan temples as well as by his confiscations and new taxes; and there is no doubt that some of his more prominent supporters owed their success, at least partly, to their timely adoption of the Emperor's religion.

The foundation of Constantinople, an act of crucial long-term importance, was very much Constantine's personal achievement. Yet it, too, had been foreshadowed; Diocletian had already enhanced Nicomedia to an extent that was considered to challenge Rome. The city itself

exemplified the "religious rapacity" of the Emperor, being filled with the artistic spoils of the Greek temples; while some of its public buildings and some of the mansions erected for Constantine's supporters soon showed signs of their hasty construction. Its Senate, too, created to match that of Rome, for long lacked the aristocratic pedigree and prestige of its counterpart.

In military policy, Constantine enjoyed unbroken success, with triumphs over the Franks, Sarmatians, and Goths to add to his victories in the civil wars; the latter, in particular, show a bold and imaginative mastery of strategy. Constantine was totally ruthless toward his political enemies, while his legislation, apart from its particular concessions to Christian sentiment, is mainly notable for a brutality that becomes only too characteristic of late Roman enforcement of law. Politically, Constantine's main contribution was perhaps that, in leaving the empire to his three sons, he re-established a dynastic succession; but it was secured only by a sequence of political murders immediately after his death.

Above all, Constantine's achievement was perhaps greatest in social and cultural history. It was the development, after his example, of a Christianized imperial governing class that, together with his dynastic success, most firmly entrenched the privileged position of Christianity; and it was this movement of fashion, rather than the enforcement of any program of legislation, which was the basis of the Christianization of the Roman Empire. Emerging from it in the course of the 4th century were two developments that contributed fundamentally to the nature of Byzantine and Western medieval culture: the growth of a specifically Christian, biblical culture that took its place beside the traditional Classical culture of the upper classes; and the extension of new forms of religious patronage, as initiated by Constantine, between the secular governing classes and bishops, Christian intellectuals and holy men. Constantine left much for his successors to do; but it was his personal choice made in 312 that determined the emergence of the Roman Empire as a Christian state. It is not hard to see why Eusebius regarded his reign as the fulfillment of divine providence—nor to concede the force of Constantine's assessment of his own role as that of the thirteenth Apostle. Constantine would not have objected to being told that he had changed the course of history.

BIBLIOGRAPHY. Two accounts remain classic: chapters 14–18 of EDWARD GIBBON, *The History of the Decline and Fall of the Roman Empire*, 6 vol. (1776–88); and JACOB BURCKHARDT, *Die Zeit Constantins der Grossen*, 2nd ed. (1880; Eng. trans., *The Age of Constantine the Great*, 1949). A.H.M. JONES, *The Later Roman Empire, 284–602*, 3 vol. (1964), should also be consulted. N.H. BAYNES, *Constantine the Great and the Christian Church* (1931; 2nd ed., with a preface by HENRY CHADWICK, 1972), has fundamentally influenced modern work by its emphasis on Constantine's own writings, of the genuineness of which there is now no question. See also A.H.M. JONES, *Constantine and the Conversion of Europe* (1948, reprinted 1972); ANDRAS ALFOLDI, *The Conversion of Constantine and Pagan Rome* (1948, reprinted 1969); and RAMSAY MACMULLEN, *Constantine* (1969). On Constantine's church building, RICHARD KRAUTHEIMER, *Early Christian and Byzantine Architecture* (1965), is reliable.

Sources: For Constantine's letters, see especially EUSEBIUS, *Ecclesiastical History*, Book X, and his *On the Life of Constantine* (both trans. in *A Select Library of Nicene and Post-Nicene Fathers of the Christian Church*, 2nd Series, vol. 1, 1961); OPTATUS, *Appendix* (ed. by C. ZIWSA, *Corpus Scriptorum Ecclesiasticorum Latinorum*, vol. 26, 1893; not translated); also LACTANTIUS, *De Mortibus Persecutorum* (trans. in *Fathers of the Church*, vol. 54, 1965). The ancient secular accounts are exiguous; the fullest, though tendentiously hostile, is ZOSIMUS, *Historia nova*, trans. by J.J. BUCHANAN and H.T. DAVIS (1967). The bulk of Constantine's surviving legislation is in the *Theodosian Code*, trans. by CLYDE PHARR (1952).

(J.F.Ma.)

Constantine VII Porphyrogenitus

Neither a great ruler nor an imaginative writer, but an unusually erudite, industrious, and well-meaning man, Emperor Constantine VII is regarded not only as one of the best sources of factual information on the Byzantine Empire and its neighbours but also as a typical exponent of the Byzantine upper class mentality, its ideals, and its foibles. His fame rests mainly on the books he wrote (no doubt with the assistance of a large secretarial staff) and on those he ordered to be written. For the spectacular recovery of the empire in his time from its previous weakness he is generally given no credit, on the ground that he was immersed in his studies while others took care of war and administration—a not altogether unwarranted judgment but one that should probably be qualified.



Constantine VII, coin. In the British Museum.
By courtesy of the trustees of the British Museum

Constantine's surname, Porphyrogenitus (that is, born in the Purple Chamber of the Imperial Palace in Constantinople, as befitted legitimate children of reigning emperors), pointedly answers the doubts expressed about the legitimacy of his birth in 905, which slowed down his career and contributed to his shyness. His mother was Zoë Carbonopsina, the mistress of his father, Leo VI, who married her shortly after, against the bitter opposition of the patriarch Nicholas Mysticus. It was Leo's fourth marriage, and the Greek Church normally forbade a widower to remarry more than once. As the infant was Leo's only male offspring, he had to be accepted and, in 911, was proclaimed co-emperor. But, on the death of his father, in 912, the succession fell to his uncle Alexander, whose death the next year cleared the way for seven-year-old Constantine. The Patriarch Nicholas, who became regent, found it expedient to appease the powerful tsar Symeon of Bulgaria—who had severely defeated the Byzantine armies and coveted the Byzantine imperial crown—by promising that the child emperor would marry Symeon's daughter. A palace revolt foiled the scheme, which looked like a betrayal of Byzantium to the Bulgarians. It was only after several years that a combination of diplomacy and successful defense of Constantinople succeeded in inducing Symeon to settle for recognition as emperor of the Bulgarians only. The strategist of this success, Adm. Romanus Lecapenus, rewarded himself by having Constantine marry his daughter (919) and crown him co-emperor (920). Gradually Constantine was made to play second fiddle not only to Lecapenus but to his sons as well.

Non-involvement in government

It is not surprising that the young emperor slipped into a pattern of noninvolvement in government. His mother had been relegated to a convent. His father-in-law relieved him of the burdensome tasks of politics and war and shouldered them masterfully but treated him with deference and left him a full share of the prestige and income belonging to the crown. From his father, Constantine had apparently inherited a passion for learning and writing; he worked full-time at it until he was almost 40, when he became sole emperor. Nor did he change tastes thereafter. *De thematibus*, probably his earliest book, is mainly a compilation of older sources on the origins and development of the provinces of the empire. An apologetic biography of his grandfather Basil I, which he appended to an anonymous chronicle known as *Theophanes Continuatus*, stressed the glory of the founder of his dynasty. *De administrando imperio*, a handbook of foreign

politics, is perhaps his most valuable work, a storehouse of information on Slavic and Turkic peoples about whom little else is known except through archaeology.

Yet, the longest book and the one that tells the most about the Byzantine mentality (and most particularly the mind of the writer) is *De ceremoniis aulae Byzantinae*, basically a minute description of the elaborate ceremonial and processions that made the emperor a hieratic symbol of the state and strove to impress foreigners with his grandeur. There is no doubt that it helped Byzantium in its relations with the northern "barbarians" and even with western Europe. A monument to Byzantine patriotism, the book bears traces of the spoken vernacular that crept into the stilted Greek of more academic writers. The more voluminous, encyclopaedic works compiled under Constantine's directions are not worth describing, but he exhibited notable zeal in recruiting teachers and students for the "university" of Constantinople, inviting them to court and preferring them for public offices. He signed pieces of legislation and is said to have dabbled in various fine and mechanical arts.

Decision to rule alone

Late in 944 the sons of Romanus Lecapenus, impatient to succeed their father, had him deported; but the populace of the capital, fearing only that the Porphyrogenitus emperor might be included in the purge accompanying the seizure of power, rioted until Constantine appeared at a window of the palace. This show of loyalty emboldened him to banish Romanus' sons January 945; he then ruled alone until his death in 959. He appointed to the highest army commands four members of the Phocas family, which had been in disgrace under Romanus Lecapenus, but took no further reprisals, except for an incidental remark, in *De ceremoniis*, that Romanus Lecapenus was neither an aristocrat nor a cultured man. That he did not depart from the Admiral's basic policy —at home, maintaining a delicate balance among civil and military officers, landed aristocrats, and peasant soldiers; abroad, friendship with the Russians, peace with the Bulgarians, a limited commitment in Italy, and a resolute offensive against the Muslims—may be ascribed to statesmanship as well as to timidity. The policy continued to be effective. Constantine's gently personal touch is apparent in his warm relations with his wife and children, his love for court ceremonial and spectacles, and his disarming kindness.

BIBLIOGRAPHY. The only book-size biography, A.N. RAMBAUD, *L'Empire grec au dixième siècle: Constantin Porphyrogénète* (1870), is partly outdated, but the essential elements may be found in ch. 4 of GEORGE OSTROGORSKY, *Geschichte des byzantinischen Staates* (1965; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968); and, still better, in *Cambridge Medieval History*, 2nd ed., vol. 4 (1966), ch. 4 by HENRI GREGOIRE. The latter two contain exhaustive bibliographic data on Constantine's life and works.

(R.S.L.)

Constants, Physical

Throughout all of the formulations of the basic theories of physics and their application to the real world, there appear again and again certain fundamental invariant quantities. These quantities, all of which are associated with specific and universally used symbols, called the fundamental physical constants, are of such importance that they must be known to as high an accuracy as is possible. They include the velocity of light in vacuum (c); the charge on the electron (e), which is the fundamental unit of electric charge; the mass of the electron (m_e); Planck's constant (h); and the fine-structure constant, symbolized by the Greek letter alpha (α). These will all be considered in detail below.

There are, of course, many other important quantities that can be measured with high accuracy—the density of a particular piece of silver, for example, or the lattice spacing (the distance between the planes of atoms) of a particular crystal of silicon, or the distance from the Earth to the Sun. These quantities, however, are generally not considered to be fundamental constants. First, they are not universal invariants because they are too specific, too closely associated with the particular properties of the material or system upon which the measurements are carried out. Second, such quantities lack universality because they do not consistently appear in the basic theoretical equations of physics upon which the entire science rests, nor are they properties of the fundamental particles of physics of which all matter is constituted.

It is important to know the numerical values of the fundamental constants with high accuracy for at least two reasons. First, the quantitative predictions of the basic theories of physics depend on the numerical values of the constants that appear in the theories. An accurate knowledge of their values is therefore essential if man hopes to achieve an accurate quantitative description of the physical universe. Second, and more important, the careful study of the numerical values of these constants, as determined from various experiments in the different fields of physics, can in turn test the overall consistency and correctness of the basic theories of physics themselves.

DEFINITION, IMPORTANCE, AND ACCURACY

The constants named above, five among many, were listed because they exemplify the different origins of fundamental constants. The velocity of light (c) and Planck's constant (h) are examples of quantities that occur naturally in the mathematical formulation of certain fundamental physical theories, the former in James Clerk Maxwell's theory of electric and magnetic fields and Albert Einstein's theories of relativity, and the latter in the theory of atomic particles, or quantum theory. For example, in Einstein's theories of relativity, mass and energy are equivalent, the energy (E) being directly proportional to the mass (m), with the constant of proportionality being the velocity of light squared (c^2)—*i.e.*, the famous equation $E = mc^2$. In this equation, E and m are variables and c is invariant, a constant of the equation. In quantum theory, the energy (E) and frequency, symbolized by the Greek letter nu (ν), of a photon (a single quantum unit of electromagnetic energy such as light or heat radiation) are related by $E = h\nu$. Here, Planck's constant (h) is the constant of proportionality.

The electron charge (e) and the electron mass are examples of constants that characterize the basic, or elementary, particles that constitute matter, such as the electron, alpha particle, proton, neutron, muon, and pion. Additionally, they are examples of constants that are used as standard units of measurement. The charge and mass of atomic and elementary particles are usually expressed in terms of the electron charge (e) and the electron mass (m_e); the charge of an alpha particle, the nucleus of the helium atom, is given as $2e$; whereas the mass of the muon is given as $206.77\ m_e$.

The fine-structure constant (α) is an example of a fundamental constant that can be expressed as a combination of other constants. The fine-structure constant is equal to a numerical constant times the velocity of light times the electron charge squared divided by twice Planck's constant, or $\mu_0 ce^2/2h$, μ_0 being the so-called permeability of free space, numerically equal to exactly $4\pi \times 10^{-7}$. (The system of measurement units used in this article is the Système Internationale d'Unités [International System of Units], or SI.) Because this particular combination of constants always appears in theoretical equations in exactly the same way, however, the fine-structure constant is really a fundamental constant in its own right. For example, the fine-structure constant is the fundamental constant of quantum electrodynamics, the quantum theory of the interaction (mutual influence) among electrons, muons, and photons. As such it is a measure of the strength of these interactions. Another quantity that is a combination of other constants is the Rydberg constant (symbolized R_∞), which is equal to the product $\mu_0^2 c^3 e^4 m_e / 8h^3$. It sets the scale (magnitude) of the various allowed electron energy states or levels in atoms such as hydrogen.

The accuracy with which many of the fundamental constants can be currently measured is a few parts in a million. By accuracy is meant the relative size of the

Uncertainty in measuring constants

uncertainty that must be assigned to the numerical value of any quantity to indicate how far from the true value it may be because of limitations in experiment or theory. This uncertainty is a quantitative estimate of the extent of the doubts associated with the value. The most commonly used uncertainty, the standard deviation, symbolized by the Greek letter sigma (σ), is such that there is about a 68 percent chance that the true value lies within plus or minus σ. Furthermore, there is a 95 percent chance that the true value lies between plus and minus two standard deviations, 2σ, and a 99.7 percent chance that it lies between plus and minus 3σ. (All uncertainties quoted in this article will be one standard deviation.)

In practice, an accuracy or uncertainty of one part per million (abbreviated ppm) is rather respectable. It corresponds to determining the length of a United States football field (100 yards, or about 91 metres) to within the thickness of two of these pages (one page is about 0.0022 inch thick). There are several quantities that have been measured with uncertainties approaching one part in 1,000,000,000,000 (one in 10^{12}); this uncertainty corresponds to determining the distance from New York to San Francisco to within one-tenth the thickness of this page.

Table 1 displays the best values for the six fundamental constants mentioned above and gives their uncertainties (in ppm) as accepted in the early 1970s. Here, $1/\alpha$, the inverse of the fine-structure constant, is given rather than α because it is a simpler number. The fine-structure constant is dimensionless; *i.e.*, it is a pure number and, therefore, has no units.

Table 1: Values of Some Selected Fundamental Constants

	symbol	units	value*	uncertainty (ppm)†
Velocity of light in vacuum	c	10^8 metres per second	2.9979250(10)	0.33
Electron charge	e	10^{-19} coulomb	1.6021917(70)	4.4
Planck's constant	h	10^{-34} joule-second	6.626196(50)	7.6
Electron rest mass	m_e	10^{-31} kilogram	9.109558(54)	6.0
Inverse fine-structure constant	$1/\alpha$	—	137.03602(21)	1.5
Rydberg constant	R_∞	10^7 per metre	1.09737312(11)	0.10

*The numbers in parentheses are the standard-deviation uncertainties in the last digits of the quoted values. †Parts per million.

HISTORICAL MEASUREMENTS

Millikan's oil-drop experiments

One of the earlier experiments to measure a fundamental constant to high accuracy, as well as an example of how the accurate determination of a fundamental constant using different methods can lead to an improved understanding of a particular physical phenomenon, was the measurement of the fundamental unit of charge (e) by Robert A. Millikan, a physicist in the United States. From about 1907 to 1917 he carried out his now-famous oil-drop experiment to determine e. In this method, the displacement of small, charged oil drops (the charge on the drop is usually just a few e) moving in air between two horizontal and parallel metal plates (with and without an applied known voltage) is followed as a function of time. The value of the fundamental constant e is then calculated from many observations on different drops and knowledge of other relevant quantities, especially the viscosity (resistance to flow) of the air. Millikan's final value, reported in 1917, was: $(4.774 \pm 0.002) \times 10^{-10}$ esu (esu being the electrostatic unit, one of the units of charge in the centimetre-gram-second [cgs] system of units; this cgs-esu system was in wide use before the general adoption of the SI system).

That this value was significantly in error became clear in the 1930s with the development of a new but indirect method for obtaining the value of e. The technique consisted of separately measuring N, the Avogadro constant (the number of atoms or molecules contained in a mole, which is defined as a mass in grams equal to the atomic or molecular weight of a substance), and F, the Faraday constant (the amount of charge that must pass through a solution to electrolytically deposit a mole of a singly charged, or monovalent, element contained in the solution). These two quantities are related by the simple equation that states that the Faraday constant is equal to the Avogadro constant times unit of charge, or $F = Ne$. It therefore follows that $e = F/N$; so that the constant e can readily be obtained if the two constants, Faraday and Avogadro, are known.

The Avogadro constant (N) was determined by measuring the density, molecular weight, and crystal lattice spacing of a particular crystal species such as rock salt, using X-ray techniques. The Faraday (F) was determined by measuring the mass of material (*e.g.*, silver) electrolytically deposited onto an electrode when a known current flowing for a known time was allowed to pass through a solution containing the material. The indirect value of the electron charge (e) deduced in this way was $(4.8021 \pm 0.0009) \times 10^{-10}$ esu, significantly different from the Millikan value. The major source of this disturbing discrepancy was traced in the latter part of the 1930s to the use by Millikan of an incorrect value for the viscosity of air. Millikan had taken a value that was almost entirely based on a measurement by one of his students; but it was later shown that the student had made a rather subtle experimental error. When Millikan's data were re-evaluated with a correctly determined value for the viscosity of air, the value of e obtained agreed with the indirect value calculated from the Faraday and the Avogadro constant.

Although this case is an example of the general fact that the experimentally determined value of a constant varies with each determination, it must be realized that it is just these variations from determination to determination in the measured numerical values of the constants that often furnish important clues to errors in experiment and theory.

The fundamental-constants field has advanced so rapidly since mid-20th century that nearly all of the measurements carried out before World War II may be considered historical (if not the method, at least the result). Indeed, few measurements of constants existed before about the turn of the 20th century, because not until then did the modern era of physics begin. Relativity, atomic physics, and quantum theory all emerged after 1900. Some of the more important historical measurements made before 1940, in addition to Millikan's oildrop measurement of e, the early iodine and silver measurements of the Faraday made about 1910 to 1915, and the early determinations of the Avogadro constant in the 1930s by physicists in the United States, Joyce A. Bearden and others using X-ray techniques, include:

Velocity of light in vacuum (c). To determine a velocity requires knowledge of both a distance and a time. Attempts to achieve measurements of the speed of light (c) date back to the 16th- and 17th-century Italian scientist Galileo Galilei, who supposedly tried unsuccessfully to determine c by having two men stand at a known distance from each other and alternately cover and uncover their hand-held lanterns as soon as they saw the light from the other man's lamp, thus seeking to determine the elapsed time for light to travel the known distance between the two men. A 17th-century Danish astronomer, Ole Rømer, calculated a value of c from the dependence of the period of revolution of a moon of Jupiter on the Earth's orbital position about the Sun. Similarly, in 1726, an English astronomer, James Bradley, determined c from the apparent change in position of a number of stars in the sky as the Earth moved about the Sun.

The problem of overcoming the short time interval associated with light travelling a readily measured distance on the Earth's surface was first solved by a French physicist, Armand-Hippolyte-Louis Fizeau, in the mid-19th century. He did this by having the light pass through a gap between the teeth of a toothed wheel rotating at a known rate, reflect off a fixed mirror a known distance away, and return to the wheel. A related method utilizing a rotating mirror was also employed by another French physicist, Jean Foucault, in 1862.

Michelson's determinations of *c*

The classic pre-World War II measurements of the constant *c* are associated with Albert A. Michelson, a physicist in the United States. From 1924 to 1926, Michelson measured *c* by reflecting light between a rotating mirror with a number of faces and a fixed mirror some 35 kilometres (22 miles) away. A second measurement using essentially the same method but in a 1.6-kilometre (one-mile) evacuated tube was carried out by Michelson and his associates over the period 1931 to 1935.

Ratio of the electron charge (e) to the electron mass (m_e), e/m_e. Numerous direct measurements of this quantity were carried out over the period 1897 to 1938. The experiments usually involved the deflection of beams of free electrons by electric and magnetic fields. Many of the experiments required the measurement of the velocity of the electrons combined with a simultaneous determination of the voltage used to initially impart kinetic energy (*i.e.*, velocity) to the electrons. Often the electron velocity was determined by a null deflection method in which the magnitudes of crossed electric and magnetic fields, through which the electron beam travelled, were so adjusted that the electric and magnetic deflecting forces just balanced each other. An English physicist, Joseph John Thomson, was the first to use this technique in 1897.

Ratio of Planck's constant (h) to the electron charge (e), h/e. The very first precision determination of the ratio h/e used the photoelectric effect: when light of a particular wavelength is allowed to impinge upon a metal surface, electrons are emitted from the surface. If a retarding voltage, or potential, is applied to the metal so that the electrons are just prevented from leaving the surface, then a unique relationship can be shown to exist between the wavelength of the light, the voltage, and the ratio h/e. Millikan, using sodium and lithium, first reported a result from this method in 1916.

A second method to determine the h/e ratio is the so-called short wavelength limit of the continuous X-ray spectrum. In this technique, a beam of electrons is accelerated through a known voltage and is allowed to strike a metal target. The maximum-energy X-ray (that is, the one having the highest frequency or shortest wavelength) is emitted when all of the electrical potential energy of an electron in the beam is converted to a single X-ray photon. By measuring the voltage and the wavelength of the emitted X-ray, the ratio h/e can be determined. The first precision measurement of this type was reported in 1921.

The Newtonian gravitational constant (G). The universal, or Newtonian (after Isaac Newton), gravitational constant (G) is the constant of proportionality in the equation relating the gravitational force between two separated bodies to their respective masses. There have been two different classes of experiment to measure the constant G. The first involves estimating the mass of the Earth and separately determining the radius of the Earth and the acceleration of an object falling toward the Earth because of gravity at its surface. Attempts to measure the mass of the Earth have been of two kinds. In 1775, a British astronomer, Nevil Maskelyne, used the deflection of a plumb line from the vertical when placed on either side of a mountain of known shape and density. In 1854 another astronomer, George Biddell Airy, of England, measured the gravitational constant by comparing the period of a pendulum's swing at the Earth's surface and at the bottom of a mine shaft of known depth.

Cavendish's measurement of the gravitational constant

The second general class of experiment for determining the gravitational constant, a significantly more accurate one, consists of measuring the gravitational force attracting two masses in the laboratory. In 1798 Henry Cavendish of England using a torsion balance designed a few years earlier, carried out the first such experiment. He suspended by a thin fibre a light, stiff rod with two solid five-centimetre (two-inch) diameter lead spheres attached at either end. He then brought two 30-centimetre diameter (12-inch) lead spheres near the smaller spheres. The gravitational attraction between them produced a torque, or turning force, that twisted or deflected the suspension fibre. The most accurate value of the gravitational constant currently available was obtained using the same general method during the period 1932–42.

MODERN MEASUREMENTS

The great progress made in determining the numerical values of the fundamental constants after World War II is the direct result of the advances made in the general fields of electronics, microwaves, and other technologies during the war. These advances have not only resulted in new and improved measurements of some of the constants listed earlier (for example, a 1957 velocity of light determination using a microwave interferometer) but for the first time permitted the direct measurement of a whole new group of constants and related quantities. The more important of these new measurements are as follows, the first two of which are concerned with the proton, an atomic particle having a mass approximately 1,836 times that of an electron and a charge identical to the electron charge, only positive:

Gyromagnetic ratio of the proton. This quantity, symbolized by the Greek letter gamma followed by a subscript p, γ_p, is a measure of how fast the axis of the proton's intrinsic rotational motion, or spin, precesses (swings) in a magnetic field much as does a child's top. It may be determined by first establishing a known magnetic field that has been produced by the passage of a known electric current through a precision coil, or solenoid, of known dimensions and then measuring by means of standard electronic techniques the precession frequency of the protons in a water sample within the solenoid. This procedure is the low-field method, in which the accurately determined magnetic field is only about 20 times the Earth's magnetic field. In the high-field method, the magnetic field, established by an electromagnet, may be 10,000 times larger. For this case, the field is determined by measuring the force it exerts on a small coil of known dimensions carrying a known current; this apparatus is often called a Cotton balance. The gyromagnetic ratio of the proton was first determined to high accuracy by the high-field method in 1950 and by the low-field method in 1957.

Magnetic moment of the proton. The magnetic moment of the proton in nuclear magnetons (μ_p/μ_n, in which μ is the Greek letter mu), is the ratio of the above-mentioned spin axis precession frequency of a proton in a magnetic field to the frequency of the proton's orbital, or circular, motion in the same field, called the cyclotron frequency. The first measurement of this ratio was reported in 1949 by physicists in the United States, John A. Hipple and his associates, using a small, metal, boxlike device called an omegatron. In this method, an adjustable radio-frequency electric field is applied to the omegatron at right angles to the direction of the magnetic field. When the frequency of this electric field is properly adjusted, protons in the omegatron are accelerated by it and spiral outward until they hit an internal collector and are detected. The frequency of the proton's orbital motion can then be determined from the adjusted frequency of the electric field.

Measuring the magnetic moment

Fine structure of atomic hydrogen. The term fine structure refers to the differences between certain states of energy or energy levels in atoms. The fine structure of atomic hydrogen was first measured with high accuracy by a United States physicist, Willis E. Lamb, Jr., and co-workers and reported in a series of classic papers over the period 1950 to 1953. In these experiments, changes in the energy state of particular atoms or, equivalently, transitions between energy levels in the atoms themselves were induced by irradiating a beam of the atoms with microwaves of a properly adjusted and known frequency as the beam was passed through a known magnetic field. The energy differences, or fine structure, could then be accurately calculated from the field and frequency.

Free electron g factor. Because the electron has an electric charge and an intrinsic rotational motion, or

spin, it behaves in some respects like a small bar magnet; that is, it is said to have a magnetic moment. Because the electron also has mass, it behaves in some respects like a spinning top; that is, it is said to have spin angular momentum. The g factor of the electron is defined as the ratio of its magnetic moment to its spin angular momentum. This factor is nominally 2 and was first measured with high accuracy during the period from 1961 to 1963. Using electric and magnetic fields, electrons were trapped with spins prealigned in a particular direction for a known length of time. The g factor was then obtained from the change in spin direction during the trapping period and the magnitude of the trapping magnetic field. Recent improvements in this basic method of measuring the g factor reduced the original 0.027 parts per million uncertainty obtained earlier to 0.003 parts per million.

Ground state hyperfine splitting (hfs) in atomic hydrogen. The ground state, or lowest energy state, hyperfine splitting in hydrogen is basically equal to the energy difference between a hydrogen atom in which the spin of the orbital electron is in the same direction as that of the spin of the central proton and a hydrogen atom in which the spins of the electron and proton are in opposite directions. Polykarp Kusch, a physicist working in the United States, reported the first high-accuracy measurement of this quantity in 1955, using a microwave-excitation method not too unlike that used for fine-structure measurements.

Use of a maser to measure hyperfine splitting

A much more accurate value was obtained in 1963 by a U.S. physicist, Norman F. Ramsey, and co-workers using the so-called hydrogen maser, which is a type of microwave amplifier. In this device, excited hydrogen atoms are focussed by a magnetic field onto an aperture in a Teflon-coated quartz bulb that is located in a radio-frequency-resonant metallic cavity. When the cavity frequency is tuned to the frequency corresponding to the hyperfine-splitting-energy difference, maser-like oscillations are produced. The measurable oscillation frequency then equals the hyperfine splitting.

Table 2: Values of the Auxiliary Constants Used in the 1969 Adjustment

quantity	symbol	value*	units	uncertainty (ppm)†
Velocity of light	c	299,792.50(10)	km/sec	0.33
Ratio of absolute ohm to National Bureau of Standards ohm	$\Omega_{ABS}/\Omega_{NBS}$	1.00000036(70)		0.70
Velocity of light times above ratio	$c\Omega_{ABS}/\Omega_{NBS}$	299,792.61(12)	km/sec	0.39
Acceleration of gravity (U.S. National gravity base, Washington, D.C.)	g(CB)	980,104.23(10)	mgal	0.10
Acceleration of gravity (National Physical Laboratory, Teddington, Eng.)	g(BFS)	981,181.86(10)	mgal	0.10
Acceleration of gravity (All Union Institute of Metrology, Leningrad)	g(VNIIM)	981,917.00(1.00)	mgal	1.0
Electron magnetic moment in Bohr magnetons	μ_e/μ_B	1.001159639(3)		0.003
Proton magnetic moment in Bohr magnetons	μ_p/μ_B	0.00152103264(46)		0.30
Proton magnetic moment in Bohr magnetons	μ_p'/μ_B	0.00152099312(10)		0.066
Ratio of proton to electron magnetic moments	μ_p/μ	0.00151927083(46)		0.30
Ratio of proton to electron magnetic moments	μ_p'/μ_e	0.00151923136(10)		0.066
Diamagnetic shielding constant	$\sigma(p')$	26.0(3)	ppm	0.3
Proton atomic mass	M_p*	1.00727661(8)	amu	0.08
Unity plus ratio of free electron mass to free proton mass	$1 + m_e/M_p$	1.000544630(10)		0.01
Unity plus ratio of free electron mass to free deuteron mass	$1 + m_e/M_d$	1.000272450(10)		0.01
Unity plus ratio of free electron mass to free alpha particle mass	$1 + m_e/M_\alpha$	1.000137097(10)		0.01
Rydberg constant for infinite mass	R_∞	10,973,731.2(1.1)	m^{-1}	0.10

*The numbers in parentheses are the uncertainties in the last digits of the quoted values. †Parts per million. Abbreviations are as follows: mgal = milligal (10^{-3} cm/sec^2); amu = atomic mass unit ($^{12}C = 12$); p' = "for protons in a spherical sample of water."

NEW DETERMINATIONS AND INTERRELATIONSHIPS AMONG THE CONSTANTS

To these important post-World War II measurements must be added the determination of the ratio of twice the electron charge to Planck's constant ($2e/h$). First reported in 1967 by the United States physicists William H. Parker, Barry N. Taylor, and Donald N. Langenberg, this low-temperature physics experiment is probably the best example in recent years of the important consequences that can follow from a high-accuracy determination of a fundamental constant. Furthermore, its discussion naturally leads to the exploration of other ideas such as the interdependency of the constants, standards of measurement, conversion factors, and least squares adjustments of the constants.

The $2e/h$ measurement in question is based on a remarkable phenomenon in superconductors (metals that lose their electrical resistance at extremely low temperatures) known as the alternating current (ac) Josephson effect. In 1962, a physicist, Brian Josephson of Cambridge University, England, showed theoretically that, if two superconductors are weakly connected together, then an alternating resistanceless current, or ac supercurrent, will flow between them if they are maintained at a finite voltage difference (V). He also showed that the frequency, symbolized by the Greek letter nu (ν), of the ac supercurrent is proportional to the voltage difference, the constant of proportionality being simply $2e/h$. Thus, $\nu = (2e/h)V$. Because the Josephson frequency–voltage relation is believed to be exact and independent of a wide variety of experimental variables, the determination of the ratio $2e/h$ is rather straightforward when compared with most other fundamental-constants experiments. It is necessary to measure only the voltage difference between the two superconductors and the frequency of the oscillating supercurrent. Indeed, the final result of these measurements of the ratio $2e/h$ had an uncertainty of only 2.4 parts per million, some 20 times less than the best previous value that had been determined from an X-ray experiment. More recent Josephson-effect measurements of $2e/h$ by several different workers have reduced the uncertainty to a few tenths of a part per million.

Use of the Josephson effect

The Josephson-effect determination of the constant $2e/h$ has had its greatest impact in the field of quantum electrodynamics (QED). QED is the quantum theory of interacting electrons, muons, and photons, and as previously noted, the fine-structure constant (α) determines the strength of these interactions. It is one of the most important of the modern theories of physics and one of the few that are capable of making highly accurate numerical predictions. Such predictions, however, are only possible if an accurate value of α is available, because of the fact that the theoretical expressions derivable from QED that describe the various physical quantities of interest are generally in the form of mathematical expressions involving α.

Heretofore, the most accurate values of the fine-structure constant ($\alpha = \mu_0 ce^2/2h$) were obtained from experiment with the aid of theoretical equations containing significant contributions from quantum electrodynamics itself. This situation made it difficult to compare QED theory and experiment unambiguously. Now, however, by combining the value of $2e/h$ obtained from the alternating current Josephson effect with the measured values of certain other constants, a highly accurate indirect value of α can be obtained without any essential use of quantum electrodynamic theory. As a result, definitive comparisons can be made between QED theory and experiment.

In practice, there are several ways of obtaining an indirect value of the fine-structure constant from the value of $2e/h$. One method involves combining it with some accurately known constants, such as the Rydberg con-

stant (R_∞) and the velocity of light (c), and a value for the proton gyromagnetic ratio (γ_p). Another involves combining the value of $2e/h$ with an experimental value for the Faraday (F), again some accurately known constants, and a value for the magnetic moment of the proton in nuclear magnetons (μ_p/μ_n).

In principle, there are also other potential sources of information on the fine-structure constant that do not require the use of either quantum electrodynamic theory or the quantity $2e/h$. For example, a value of α can be obtained from a measurement of the electron Compton wavelength, which is symbolized by the Greek letter lambda (λ) with a subscript C, λ_C. This value is the wavelength of the radiation emitted by an electron at rest when it annihilates with a positive electron, or positron, at rest.

The implied equations in the last two paragraphs above not only illustrate the complex relationships that exist among the constants but, more importantly, the fact that a particular constant may be determined by a direct measurement or indirectly by an appropriate combination of several other directly measured constants. Indeed, the indirect value is often so much more accurate than the direct value that the latter is discarded.

This situation is true in the case of the electron charge (e); the oil-drop determination of the electron charge has such a low accuracy compared with indirect methods that it is never used. If the direct and indirect values do in fact have comparable accuracy, then both must be taken into account in order to arrive at a best value for that quantity. (By "best value" is meant that numerical value for the quantity that is believed to be closest to the true but unknown value.) Clearly, because of the interrelationships existing among the constants and the concomitant existence of indirect values, a new determination of one constant will generally affect significantly the best values of others.

Meaning of best value

Unfortunately, the problem of determining a best value for the fine-structure constant that is independent of quantum electrodynamic theory or otherwise, as well as similar values for other fundamental constants, is made still more difficult by the existence of conversion factors relating absolute units to as-maintained units (see below for definitions).

STANDARDS OF MEASUREMENT, CONVERSION FACTORS, AND RELATED CONSIDERATIONS

Conversion factors. The measurement of any quantity must be carried out in terms of certain units. The dominant system of units currently in use throughout the world is called the Système Internationale d'Unités, or SI (see WEIGHTS AND MEASURES). It is based in part on the kilogram, metre, second, and ampere. In practice, everyday working standards of mass, length, and time can be constructed that are directly traceable to their fundamental SI definitions and that have an accuracy comparable with that implicit in the definition. This situation is not true, however, of the ampere; because a current is a flow of charge, it is not easy to construct a storable working current standard. Instead, standards of voltage and resistance must be separately maintained, usually by the national standards laboratories, and a standard of current must be derived by the application of Ohm's law—*i.e.*, current equals voltage divided by resistance.

In terms of such as-maintained amperes, currents can usually be measured to an accuracy of about 0.1 parts per million. No experiment has yet been devised, however, that will allow a given current to be determined in SI units to anything like the 0.01 ppm accuracy inherent in the definition of the SI ampere; present current balances (those that measure the force between current carrying coils) can at best give a result accurate to between 5 and 10 ppm. This means that the uncertainties of the conversion factors relating the different as-maintained amperes to the defined SI ampere are 5 to 10 ppm. (This uncertainty applies to as-maintained volt to SI volt conversion factors as well, because as-maintained ohm to SI ohm conversion factors can be determined to an accuracy of a few tenths of a ppm using the calculable capacitor method.) Because many fundamental-constants experiments require the accurate measurement of a current, or a voltage, the rather large uncertainty in the current, or voltage, conversion factor is of great significance. Indeed, the conversion factor may be considered equal in importance to the associated fundamental constants, and separate experiments must be undertaken for its determination.

Another example of a conversion factor comes from the

Table 3: QED Independent Input Data Used in the 1969 Adjustment

publication date and author	quantity	symbol	method	value*	uncertainty (ppm)†
1968, W.H. Parker, D.N. Langenberg, A. Denenstein, B.N. Taylor	ratio of twice the electron charge to Planck's constant	$2e/h$	Josephson effect	$4.835976(12) \times 10^{14}\text{Hz}/V_{NBS}$	2.4
1968, R.L. Driscoll, P.T. Olsen	conversion factor relating U.S. as-maintained ampere to absolute ampere	K	Pellat electrodynamometer	1.0000102(97)	9.7
1958, R.L. Driscoll, R.D. Cutkosky	conversion factor relating U.S. as-maintained ampere to absolute ampere	K	NBS current balance	1.0000092(77)	7.7
1965, P. Vigoureux	conversion factor relating U.S. as-maintained ampere to absolute ampere	K	NPL current balance	1.0000080(60)	6.0
1960, D.N. Craig, J.I. Hoffman, C.A. Law, W.J. Hamer	Faraday constant	F	silver-perchloric acid coulometer	$9.648570(66) \times 10^{7} A_{NBS}$ kmole^{-1}	6.8
1958–68, R.L. Driscoll, P.T. Olsen, P.L. Bender	proton gyromagnetic ratio	γ_p'	low field	$2.6751525(99) \times 10^{8}\text{Hz}/T_{NBS}$	3.7
1962, P. Vigoureux	proton gyromagnetic ratio	γ_p'	low field	$2.675144(16) \times 10^{8}\text{Hz}/T_{NBS}$	5.8
1962–66, G.K. Yagola, V.I. Zingerman, V.N. Sepetyi	proton gyromagnetic ratio	γ_p'	high field	$2.675105(20) \times 10^{8}\text{Hz}/T_{NBS}$	7.4
1949–51, H. Sommer, H.A. Thomas, J.A. Hipple	proton magnetic moment in nuclear magnetons	μ_p'/μ_n	omegatron	2.792690(30)	11
1957–63, J.H. Sanders, K.C. Tuberfield, D.J. Collington, A.N. Dellis	proton magnetic moment in nuclear magnetons	μ_p'/μ_n	inverse cyclotron	2.792701(73)	26
1961, H.S. Boyne, P.A. Franken	proton magnetic moment in nuclear magnetons	μ_p'/μ_n	cyclotron	2.792832(55)	20
1965, B.A. Mamyrin, A.A. Frantsuzov	proton magnetic moment in nuclear magnetons	μ_p'/μ_n	mass spectrometer	2.792794(17)	6.2
1967, B.W. Petley, K. Morris	proton magnetic moment in nuclear magnetons	μ_p'/μ_n	omegatron	2.792746(52)	19

*The numbers in parentheses are the standard-deviation uncertainties in the last digits of the quoted values. †Parts per million.

Table 4: Physical Constant Values*

quantity	symbol	value†	uncertainty (ppm)‡	units SI	units cgs
Velocity of light	c	2.9979250(10)	0.33	10^{8} m sec^{-1}	10^{10} cm sec^{-1}
Fine-structure constant, $[\mu_0 c^2/4\pi](e^2/\hbar c)$	α α^{-1}	7.297351(11) 137.03602(21)	1.5 1.5	10^{-3}	10^{-3}
Electron charge	e	1.6021917(70) 4.803250(21)	4.4 4.4	10^{-19} C	10^{-20} emu 10^{-10} esu
Planck's constant	h $\hbar = h/2\pi$	6.626196(50) 1.0545919(80)	7.6 7.6	10^{-34} J · sec 10^{-34} J · sec	10^{-27} erg · sec 10^{-27} erg · sec
Avogadro constant	N	6.022169(40)	6.6	10^{26} kmole^{-1}	10^{23} mole^{-1}
Atomic mass unit	amu	1.660531(11)	6.6	10^{-27} kg	10^{-24} g
Electron rest mass	m_e m_e*	9.109558(54) 5.485930(34)	6.0 6.2	10^{-31} kg 10^{-4} amu	10^{-28} g 10^{-4} amu
Proton rest mass	M_p M_p*	1.672614(11) 1.00727661(8)	6.6 0.08	10^{-27} kg amu	10^{-24} g amu
Neutron rest mass	M_n M_n*	1.674920(11) 1.00866520(10)	6.6 0.10	10^{-27} kg amu	10^{-24} g amu
Ratio of proton mass to electron mass	M_p/m_e	1,836.109(11)	6.2		
Electron charge to mass ratio	e/m_e	1.7588028(54) 5.272759(16)	3.1 3.1	10^{11} C kg^{-1}	10^{7} emu g^{-1} 10^{17} esu g^{-1}
Magnetic flux quantum, $[c]^{-1}(hc/2e)$	Φ_0 h/e	2.0678538(69) 4.135708(14) 1.3795234(46)	3.3 3.3 3.3	10^{-15} T · m^2 10^{-15} J · sec C^{-1}	10^{-7} G · cm^2 10^{-7} erg · sec emu^{-1} 10^{-17} erg · sec esu^{-1}
Quantum of circulation	$h/2m_e$ h/m_e	3.636947(11) 7.273894(22)	3.1 3.1	10^{-4} J · sec kg^{-1} 10^{-4} J · sec kg^{-1}	erg · sec g^{-1} erg · sec g^{-1}
Faraday constant, Ne	F	9.648670(54) 2.892599(16)	5.5 5.5	10^{7} C kmole^{-1}	10^{3} emu mole^{-1} 10^{14} esu mole^{-1}
Rydberg constant, $[\mu_0 c^2/4\pi]^2(m_e e^4/4\pi\hbar^3 c)$	R_∞	1.09737312(11)	0.10	10^{7} m^{-1}	10^{5} cm^{-1}
Bohr radius, $[\mu_0 c^2/4\pi]^{-1}\hbar^2/m_e e^2) = \alpha/4\pi R_\infty$	a_0	5.2917715(81)	1.5	10^{-11} m	10^{-9} cm
Classical electron radius $[\mu_0 c^2/4\pi](e^2/m_e c^2) = \alpha^3/4\pi R_\infty$	r_0	2.817939(13)	4.6	10^{-15} m	10^{-13} cm
Electron magnetic moment in Bohr magnetons	μ_e/μ_B	1.0011596389(31)	0.0031		
Bohr magneton, $[c]$ $(e\hbar/2m_e c)$	μ_B	9.274096(65)	7.0	10^{-24} J T^{-1}	10^{-21} erg G^{-1}
Electron magnetic moment	μ_e	9.284851(65)	7.0	10^{-24} J T^{-1}	10^{-21} erg G^{-1}
Gyromagnetic ratio of protons in H_2O	γ_p' $\gamma_p'/2\pi$	2.6751270(82) 4.257597(13)	3.1 3.1	10^{8} rad sec^{-1} · T^{-1} 10^{7} Hz T^{-1}	10^{4} rad sec^{-1} · G^{-1} 10^{3} Hz G^{-1}
γ_p' corrected for diamagnetism of H_2O	γ_p $\gamma_p/2\pi$	2.6751965(82) 4.257707(13)	3.1 3.1	10^{8} rad sec^{-1} · T^{-1} 10^{7} Hz T^{-1}	10^{4} rad sec^{-1} · G^{-1} 10^{3} Hz G^{-1}
Magnetic moment of protons in H_2O in Bohr magnetons	μ_p'/μ_B	1.52099312(10)	0.066	10^{-3}	10^{-3}

field of X-rays. Wavelengths of X-rays and lattice spacings of crystals are most conveniently measured in terms of a standard wavelength or a standard crystal lattice spacing. To convert the measurements to metres, the SI unit of distance, requires the use of an appropriate conversion factor that must be separately determined.

Summarizing, then, the main points of the last two sections, it may be stated that:

1. The determination of $2e/h$ using the ac Josephson effect is of great importance because it enables a value for the fine-structure constant (α) to be determined without essential use of quantum electrodynamic theory. QED theory and experiment can then be compared unequivocally.

2. There are several ways of obtaining a value of α from the value $2e/h$, and, in fact, information bearing on the value of any specific constant can come from several different fundamental-constants experiments. A series of complex relationships exist among the constants, and both direct and indirect values must be taken into account when determining best values.

3. The units in terms of which constants are measured can play an important role. Conversion factors often enter the equations relating the different fundamental constants, and these conversion factors must be considered constants in their own right.

Now in general, each of the many different routes that can be followed, both direct and indirect, in order to obtain a value for a particular constant will give a slightly different answer—that is, value. A situation such as this may best be handled by a mathematical technique known as least squares.

Table 4: Physical Constant Values* (continued)

quantity	symbol	value†	uncertainty (ppm)‡	units	
				SI	cgs
Proton magnetic moment in Bohr magnetons	μ_p/μ_B	1.52103264(46)	0.30	10^{-3}	10^{-3}
Proton magnetic moment	μ_p	1.4106203(99)	7.0	10^{-26} J T^{-1}	10^{-23} erg G^{-1}
Magnetic moment of protons in H_2O in nuclear magnetons	μ_p'/μ_n	2.792709(17)	6.2		
μ_p'/μ_n corrected for diamagnetism of H_2O	μ_p/μ_n	2.792782(17)	6.2		
Nuclear magneton, $[c](e\hbar/2M_pc)$	μ_n	5.050951(50)	10	10^{-27} J T^{-1}	10^{-24} erg G^{-1}
Compton wavelength of the electron, h/m_ec	λC $\lambda C/2\pi$	2.4263096(74) 3.861592(12)	3.1 3.1	10^{-12} m 10^{-13} m	10^{-10} cm 10^{-11} cm
Compton wavelength of the proton, h/M_pc	$\lambda C,_p$ $\lambda C,_p/2\pi$	1.3214409(90) 2.103139(14)	6.8 6.8	10^{-15} m 10^{-16} m	10^{-13} cm 10^{-14} cm
Compton wavelength of the neutron, h/M_nc	$\lambda C,_n$ $\lambda C,_n/2\pi$	1.3196217(90) 2.100243(14)	6.8 6.8	10^{-15} m 10^{-16} m	10^{-13} cm 10^{-14} cm
Gas constant	R_0	8.31434(35)	42	10^3 J $kmole^{-1} \cdot K^{-1}$	10^7 erg $mole^{-1} \cdot K^{-1}$
Boltzmann's constant, R_0/N	k	1.380622(59)	43	10^{-23} J K^{-1}	10^{-16} erg K^{-1}
Stefan-Boltzmann constant, $\pi^2k^4/60\hbar^3c^2$	σ	5.66961(96)	170	10^{-8} W m^{-2} K^{-4}	10^{-5} erg $sec^{-1} \cdot cm^{-2} \cdot K^{-4}$
First radiation constant, $2\pi hc^2$	c_1	3.741844(28)	7.6	10^{-16} W · m^2	10^{-5} erg cm^2 sec^{-1}
Second radiation constant, hc/k	c_2	1.438833(61)	43	10^{-2} m · K	cm · K
Gravitational constant	G	6.6732(31)	460	10^{-11} N · m^2 kg^{-2}	10^{-8} dyn · cm^2 g^{-2}
kx-unit-to-angstrom conversion factor, $\Lambda = \lambda(Å)/\lambda(kxu)$; $\lambda(CuK\alpha_1) \equiv 1.537400$ kxu	Λ	1.0020764(53)	5.3		
Å*-to-angstrom conversion factor, $\Lambda = \lambda(Å)/\lambda(Å)^*$; $\lambda(WK\alpha_1) \equiv 0.2090100$ Å*	Λ^*	1.0000197(56)	5.6		

*The unified atomic mass scale $^{12}C = 12$ has been used throughout; amu = atomic mass unit, C = coulomb, G = gauss, Hz = hertz = cycles per second, J = joule, K = kelvin, T = tesla (10^4 G), V = volt, and W = watt. Where formulas for constants are given, the relations are written as the product of two factors. The second factor, in parentheses, is the expression to be used when all quantities are expressed in cgs units, with the electron charge in electrostatic units. The first factor, in brackets, is to be included only if all quantities are expressed in SI units. With the exception of the auxiliary constants which have been taken to be exact, the uncertainties of these constants are correlated, and therefore the general law of error propagation must be used in calculating additional quantities requiring two or more of these constants. †The numbers in parentheses are the standard deviation uncertainties in the last digits of the quoted value, computed on the basis of internal consistency. ‡Parts per million.

Least squares adjustments of the constants. The least squares method provides a self-consistent procedure for calculating best compromise values of the constants from all of the available measurements. For a given set of data, it automatically takes into account all the possible routes for obtaining values of each of the constants being calculated. It then determines a single final value for each constant by automatically weighting the values of the constant obtained from the various routes according to their relative reliability or uncertainty. The uncertainty for each route is determined from the uncertainties of the individual measurements comprising the original set of data. A United States physicist, Raymond T. Birge, first pioneered least squares studies of the constants in the late 1920s and continued them into the mid-1940s. Similar studies have been continued to the present by a number of physicists, including Jesse W.M. DuMond and E. Richard Cohen, and Bearden and colleagues. The latest and most comprehensive critical analysis and least squares adjustment of the constants was made by Barry N. Taylor and co-workers in 1969.

A determination of uncertainties

A least squares adjustment of the constants is generally carried out by first dividing the available measurements into two groups. One group, known as the auxiliary constants, contains quantities that have uncertainties sufficiently small so that they can be considered as exactly known (see Table 2). An example is the Rydberg constant (R_∞), which has an uncertainty of 0.1 part per million. The other group contains the more imprecise, or stochastic, input data. An example of these is the proton gyromagnetic ratio (γ_p) with its approximate 4 ppm uncertainty. Next, a subset of constants is chosen in terms of which all of the stochastic input data can be individually expressed, if necessary, with the aid of the auxiliary constants. It is actually the constants comprising this subset that are directly subject to adjustment and that are termed the adjustable constants. In the adjustment carried out by Taylor and colleagues in 1969, the adjustable constants were taken to be the fine-structure constant (α), the electron charge (e), the conversion factor (K), equal to the ratio of the ampere as maintained by the United States National Bureau of Standards to the SI ampere, and the Avogadro constant (N). With just these four quantities and the aid of the auxiliary constants, a series of equations, known in general as observational equations, were formed for all the stochastic input data. The actual data considered for use by Taylor and his co-workers are summarized in Table 3. Table 3 summarizes the 13 pieces of stochastic input data considered for use in the least squares adjustment to obtain best values of the constants independent of quantum electrodynamic theory. The symbols and their meanings

Table 5: Energy Conversion Factors

quantity	symbol	value*	uncertainty (ppm)†	unit
1 kg		5.609538(24)	4.4	10^{29} MeV
1 amu		931.4812(52)	5.5	MeV
Electron mass		0.5110041(16)	3.1	MeV
Proton mass		938.2592(52)	5.5	MeV
Neutron mass		939.5527(52)	5.5	MeV
1 electron volt	eV	1.6021917(70)	4.4	10^{-19} J
				10^{-12} erg
		2.4179659(81)	3.3	10^{14} Hz
		8.065465(27)	3.3	10^{5} m^{-1}
				10^{3} cm^{-1}
		1.160485(49)	42	10^{4} K
Energy–wavelength conversion		1.2398541(41)	3.3	10^{-6} eV · m
				10^{-4} eV · cm
Rydberg constant	R_∞	2.179914(17)	7.6	10^{-18} J
				10^{-11} erg
		13.605826(45)	3.3	eV
		3.2898423(11)	0.35	10^{15} Hz
		1.578936(67)	43	10^{5} K
Bohr magneton	μ_B	5.788381(18)	3.1	10^{-5} eV T^{-1}
		1.3996108(43)	3.1	10^{10} Hz T^{-1}
		46.68598(14)	3.1	$m^{-1} \cdot T^{-1}$
				10^{-2} $cm^{-1} \cdot T^{-1}$
		0.671733(29)	43	K T^{-1}
Nuclear magneton	μ_n	3.152526(21)	6.8	10^{-8} eV T^{-1}
		7.622700(42)	5.5	10^{6} Hz T^{-1}
		2.542659(14)	5.5	10^{-2} $m^{-1} \cdot T^{-1}$
				10^{-4} $cm^{-1} \cdot T^{-1}$
		3.65846(16)	44	10^{-4} K T^{-1}
Gas constant	R_0	8.20562(35)	42	10^{-2} $m^3 \cdot$ atm $kmole^{-1} \cdot K^{-1}$
Standard volume of ideal gas	V_0	22.4136	42	m^3 $kmole^{-1}$

*The numbers in parentheses are the standard deviation uncertainties in the last digits of the quoted value, computed on the basis of internal consistency. †Parts per million.

are as follows: T = tesla (an mks unit of magnetic induction, weber per square metre); V = volt; A = ampere; Hz = hertz (cycles per second); and s = second. The subscript "NBS" means "in units as maintained by the U.S. National Bureau of Standards." The different methods used for measuring the proton magnetic moment in nuclear magnetons, $\mu_{p'}/\mu_n$ are all similar in that radiofrequency fields are used to alter the motion of protons (or other ions) which are moving in a fixed magnetic field.

Once the actual numbers are substituted for the auxiliary constants and stochastic data, it is a rather straightforward procedure, with the aid of a computer, to solve the observational equations for the least squares adjusted values of the adjustable constants, in the case under consideration, α, e, K, and N. Although numerical values (with uncertainties) for the adjustable constants are the sole result of an adjustment, optimum values for all of the constants are actually obtained, because any constant not chosen for direct adjustment can be calculated from appropriate combinations of those subject to the adjustment—that is (for the present case), from α, e, K, and N. For example, the mass of the electron (m_e) may be obtained from the equation $m_e = \mu_0 R_\infty e^2/\alpha^3$.

One of the main problems in carrying out a least squares adjustment of the constants is making the critical analysis of the input data and deciding what uncertainty should be assigned to each measurement. Correct uncertainty assignment is of the utmost importance, because the weight any particular experiment carries in an adjustment is proportional to the reciprocal of the square of its uncertainty—if one measurement of a particular quantity has half the uncertainty of another, it carries four times as much weight. One reason for the uncertainty problem is that, in most experiments, sufficient data is taken to reduce the so-called random, or statistical, uncertainty to negligible amounts, and the final uncertainty assigned the measurement is determined from estimates of the systematic uncertainties. Systematic uncertainties arise from effects that the experimenter knows little about; their estimation is somewhat subjective and is usually obtained from what can only be called educated guesses.

Discrepancy in data

Another difficult task associated with the adjusting of constants is deciding what to do with discrepant data—that is, measurements for which the assigned uncertainty seems to be correct but that differ from each other "more than they should." By this phrase is meant that the difference between two values of the same quantity is large compared with the standard deviation of the difference, as obtained by taking the square root of the sum of the squares or root-sum-square of the standard deviations of the individual measurements. If two measurements differ by between two and three times the standard deviation of their difference, they are inconsistent and one or the other is highly likely to be incorrect. Such data are not to be included in an adjustment uncritically because the inconsistencies imply either fallacious error estimates or the presence of unknown systematic uncertainties.

When faced with inconsistent data, the constants adjuster has two major choices: (1) to use as input data all apparently reliable measurements even though they may be inconsistent, but to expand (increase) the assigned uncertainties sufficiently so that they are compatible; and (2) to decide, on as sound an experimental and theoretical basis as is possible, which of the inconsistent data are least reliable and expurgate them, but expand no errors. Clearly these two possibilities indicate some of the arbitrariness or subjectiveness in the fundamental-constants field and the fact that different adjusters might treat the same data differently, thereby arriving at a slightly different set of best values.

Implications of the ac Josephson effect value of α. As intimated above, the quantum electrodynamics or QED independent adjusted value of the fine-structure constant

Table 6: Comparison of 1963 and 1969 Adjustments

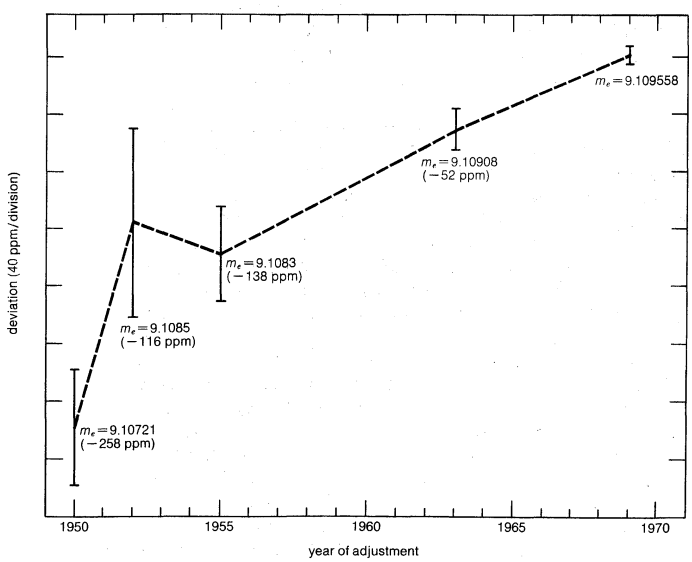
constant	symbol	units	value 1969 adjustment*	uncertainty (ppm)†	value 1963 adjustment*	uncertainty (ppm)†	change (ppm)†
Inverse fine-structure constant	$1/\alpha$	—	137.03602(21)	1.5	137.0388(6)	4.4	−20
Electron charge	e	10^{-19} coulomb	1.6021917(70)	4.4	1.60210(2)	12	+57
Planck's constant	h	10^{-34} joule-second	6.626196(50)	7.6	6.62559(16)	24	+91
Electron rest mass	m_e	10^{-31} kg	9.109558(54)	6.0	9.10908(13)	14	+52
Avogadro constant	N	10^{26} per kilomole	6.022169(40)	6.6	6.02252(9)	15	−58

*The numbers in parentheses are the standard-deviation uncertainties in the last digits of the quoted values.
†Parts per million.

(α) resulting from the data of Table 3, and therefore from the ac Josephson effect value of $2e/h$, has had its greatest impact in the field of quantum electrodynamics. The adjustment result is $1/\alpha = 137.03608$, with an uncertainty of 1.9 parts per million (ppm). Before the ac Josephson effect measurement of $2e/h$, the accepted value was 137.0388, with an uncertainty of 4.4 ppm. The two values are in clear disagreement; their difference exceeds the standard deviation of their difference by more than four times. The probability for this to occur by chance is less than one in 15,000.

This previously accepted value of α led during the mid-1960s to what was termed one of the major unsolved problems of quantum electrodynamics. It concerned an apparent discrepancy between the theoretically calculated and experimentally measured values for the ground state hyperfine splitting (hfs) in hydrogen. The theoretical equation for the hydrogen hfs obtained from QED involves only accurately known auxiliary constants and α, but it is limited to an accuracy of a few parts per million because of the difficulty in calculating from theory some of the terms in the equation. One such term is called the proton polarizability correction, symbolized by the Greek letter delta with subscript N, δ_N. It arises from the fact that the proton in the hydrogen atom cannot be regarded as simply a charged spinning ball. The proton has an internal structure of its own that affects the magnitude of the hyperfine splitting. To date, however, all calculations of δ_N show it to be rather small—1 or 2 ppm at most.

The small size of the theoretical value for the correction δ_N is in marked contrast to what is implied by the old value of the fine-structure constant. If this value is used to compare theory and experiment, then it is found that $\delta_N = 43$ ppm with an uncertainty of 9 ppm. This means that the probability for δ_N to be as small as predicted by theory—that is, one or two ppm—is only one in 20,000. The discrepancy and resulting challenge to quantum electrodynamics is clear. On the other hand, if one uses the value of α implied by the ac Josephson effect measurement of $2e/h$, it is found that $\delta_N = 2.5$ ppm with an uncertainty of 4.0 ppm, quite consistent with what is predicted theoretically. Thus, the Josephson effect value of α removes the discrepancy and resulting challenge to QED. This case is, therefore, an excellent example of how fundamental-constants experiments carried out in one field of physics can have significant implications for other fields and how accurate measurements of the fundamental constants can illuminate apparent inconsistencies in the physical description of nature.



Variation of the accepted value of the rest mass of the electron (m_e). Each value resulted from a least squares adjustment of the constants in the year indicated. The actual quantity plotted is the parts per million (ppm) deviation of the different values of m_e from the 1969 adjusted value. (The deviations themselves are the numbers in parentheses.) Also given (in units of 10^{-31} kilograms) are the actual values of m_e resulting from the different adjustments. The vertical bars correspond to plus and minus the assigned standard deviation uncertainty.

Recommended set of fundamental constants. To obtain a final best or recommended set of fundamental constants, Taylor and colleagues considered all of the available data, not just measurements that did not require quantum electrodynamic theory for their analysis. Such a set may be readily found by combining the most reliable QED data with the non-QED data previously used (Table 3) and carrying out a new least squares adjustment. This procedure was, in fact, followed, and some of the resulting values of this final adjustment are given in Table 4. In Table 4 the stochastic input data include items 1 through 13 (with the exception of 11 and 12) of Table 3, plus a value of the fine-structure constant derived from the theoretical expression for, and the experimental measurement of, the ground state hyperfine splitting (hfs) in atomic hydrogen. It should be noted that not all of these constants result from the least squares adjustment; for example, the auxiliary constants such as c, R_∞, M_p, and G. Table 5 is a list of energy conversion factors.

Combining results

In Table 6 the values for some of the more important constants are compared with the corresponding previously accepted best values resulting from the 1963 adjustment of Cohen and DuMond. It is evident from Table 6 that the values have changed by several times the uncertainties assigned the old values and that the uncertainties resulting from the 1969 adjustment are significantly less than the corresponding 1963 adjustment uncertainties. The changes serve as a good example of the intimate relationships that exist among the fundamental constants —a significant shift in the value of one constant will generally cause important shifts in the values of others.

It is also of interest to consider how knowledge of the fundamental constants has changed over the years. As a typical example, the Figure shows the variation of the accepted value of the electron mass (m_e), resulting from several different adjustments carried out since 1950. It is clear from this graph that significant changes in the electron mass from one adjustment to another are common. This figure graphically illustrates the fact that no set of fundamental constants should be taken as final.

Official adoption of values for the fundamental constants. Until the 1960s, there were no national or international bodies established for the express purpose of of-

ficially adopting a particular set of recommended values for the fundamental constants. In the past, various international organizations adopted values for individual constants that related directly to their areas of interest. In 1961, the Advisory Committee on Fundamental Constants under the National Academy of Sciences-National Academy of Engineering-National Research Council (United States) was established as a study group for the purpose of arriving at best values for the more important constants. The work of this new committee led to the 1963 set of adjusted constants of Cohen and DuMond. These were adopted officially throughout the United States, and more or less throughout the world, but not officially because the academy committee was only national in scope and membership.

About 1968, the academy committee was to a large extent supplanted by the creation of an equivalent international one, the Task Group on Fundamental Constants, under the auspices of the International Council of Scientific Unions, Committee on Data for Science and Technology (Codata). Although this committee has yet to generate or adopt a new set of constants, it is anticipated that it will do so sometimes in the 1970s.

BIBLIOGRAPHY. E. RICHARD COHEN, KENNETH M. CROWE, and JESSE W.M. DUMOND, *The Fundamental Constants of Physics* (1957), a good historical review of the fundamental constants field from its earliest days up to about 1955; B.N. TAYLOR, W.H. PARKER, and D.N. LANGENBERG, *The Fundamental Constants and Quantum Electrodynamics* (1969), the most recent critical review and least-squares adjustment of the constants currently available; K.D. FROOME and L. ESSEN, *The Velocity of Light and Radio Waves* (1969), a detailed, historical account of the various measurements of the velocity of light carried out from the 16th to the mid-20th centuries; D.N. LANGENBERG and B.N. TAYLOR (eds.), *Proceedings of the International Conference on Precision Measurement and Fundamental Constants* (1971), includes 90 or so papers given at a conference held at the United States National Bureau of Standards that brought together theoretical, experimental, and applied scientists for the purpose of discussing modern techniques of precision physical measurement and their application; J.A. BEARDEN and J.S. THOMSEN, "A Survey of Atomic Constants," *Nuovo Cim.*, Series 10, vol. 5 suppl. (1957), a long article reviewing the status of the fundamental constants field as of about 1957.

(B.N.T.)

Constitutional Law

Constitutional law is the body of doctrines and practices that govern the operation of the political state. All states may be said to have constitutions, whether or not they are written in documentary form. Modern constitutional law is an offspring of nationalism. As states have multiplied in number, so have constitutions and with them the body of constitutional law.

The article is divided into the following sections:

I. Constitutions and constitutional law
  The nature of constitutional law
  Characteristics of constitutions
II. Unitary and federal systems
  The distinction between unitary and federal states
  Quasi-federal systems
  Unitary systems
III. Executives and legislatures
  The constitution and the executive
  Unicameral and bicameral legislatures
IV. Judicial review: the unique role of the United States judiciary
  The Supreme Court and the separation of federal powers
  The Supreme Court and the federal system
  The Supreme Court and individual rights

I. Constitutions and constitutional law

THE NATURE OF CONSTITUTIONAL LAW

In the broadest sense, a constitution is a body of rules governing the affairs of an organized group. A parliament, a church congregation, a social club, or a trade union may operate under the terms of a formal written document labelled constitution. This does not mean that all of the rules of the organization are in the constitution, for usually there are many other rules such as by-laws and customs. Invariably, by definition, the rules spelled out in the constitution are considered to be basic, in the sense that all other rules must conform with its provisions. Thus the presiding officer of a club is obliged to rule that a proposal is out of order if it is contrary to a provision of its constitution. Implicit in the concept of a constitution is that of a higher law that is operative, a law that takes precedence.

Every national state has a constitution, at least in the sense that it operates its important institutions according to some fundamental body of rules. In this sense of the term, the only conceivable alternative to a constitution is a condition of anarchy. Constitutions may be written or unwritten; they may be complex or simple; they may provide for vastly different patterns of governance. Even if the only rule that matters much is the supreme authority of an absolute dictator, that may be said to be the constitution.

Constitutional law describes the whole body of interpretations that apply the constitution to the myriad institutions, functions, and problems of the state. In some countries, jurists look to judicial decisions for much of their constitutional law. But not all constitutional questions are justiciable, such as those relating to the conduct of foreign affairs by the American president, and for such matters constitutional law is developed by branches of government other than the judicial. In a larger sense, in all states, the meaning of the constitution derives from the interpretations of all who are involved in processes of government. When a legislative body adopts a statute, it may well establish or expand thereby the meaning of a constitutional provision, and acts of the executive often have the same impact. The citizen also, by his behaviour, helps to establish the content of constitutional principles. The scope of free speech often depends as much upon the limits of public tolerance as upon how judges define it in law cases. Constitutional law, therefore, embraces the whole body of doctrines and practices that derive from the constitution among all who live under it and who are subject to its provisions.

That the political community is governed by law and that its highest law is rooted in natural law concepts has been a central element of political theory and jurisprudence in the Western world since the time of the Greek city-states of antiquity. Ancient Greece recognized the existence of a higher law derived through reason from nature and regarded this law as being universal, divine, and immutable. Plato taught that human law is at best only an imperfect image of an ideal law of nature.

Aristotle (384–322 BC) recognized various kinds of constitutions. He classified them as good or perverted depending upon the ability of any particular system to achieve justice through rule in the common interest by persons chosen for their ability. Monarchy, defined as rule by a virtuous man, was characterized as best, whereas rule by a selfish man was a tyranny, the worst of the perverted systems. Second best was aristocracy, or rule by a virtuous few, and third best was a polity, or moderate democracy. Extreme democracy, described as mob rule, was regarded as the best of the perverted forms of government, and oligarchy, or rule by a selfish, wealthy minority, was considered the perverted form of aristocracy. Ancient Greek thinkers devoted a great deal of attention to the analysis of internal changes within the state, which is to say that they were deeply concerned with constitutional problems.

Aristotle did not share Plato's belief in the possibility of discovering and training philosopher-kings to rule the state. Since Aristotle did not believe it possible to find perfect men to rule, he emphasized constitutionalism and the rule of law. To him constitutionalism meant that governments must be responsible to the governed, that citizens should be involved in the process of making the laws and that they are obliged to obey, and that, however else men may be unequal, they are equal under the law.

The development of law was Rome's greatest contribution to Western civilization. While it was accepted that the emperor was the chief source of law, under his

authority scholars and jurists developed the body of Roman law, culminating in the 6th-century Justinian Code. There was a gradual development of the concept of law as legislation and of the concept of the need for popular consent to law. Political and jurisprudential writers agreed that the highest law in the community was the law of nature. Cicero held that the law of nature was the constitution of the world. In *De re publica* ("On the Republic") he declared that "true law is right reason in agreement with nature; it is of universal application, unchanging and everlasting. . . ." It is impossible to annul this law, he wrote, and people can never be absolved from the obligation to obey it. It is derived from God, and he who disobeys it denies thereby his own true nature. The best state, Cicero taught, was a mixed one that combines elements of monarchy, aristocracy, and democracy, and one that seeks to promote the common good and achieve justice.

The early church fathers taught that government is of divine origin and is both necessary and sacred. It followed that the individual has a duty to obey the law, and that there is no right of resistance; bad rule was God's way of punishing sin. The state, in cooperation with the church, keeps the peace and maintains order.

Although political obligations were regarded as contractual in the feudal system, in which obedience was traded for protection, there was a strong element of consent in the medieval concept of fundamental law. The dominant thought of the Middle Ages was that the law was the people's and applied even to the ruler.

The medieval schoolmen brought God and reason together. They believed that natural law is divine in origin and that man is a rational creature. The influential St. Thomas Aquinas divided law into four categories: (1) external law, derived from God; (2) natural law, ascertained by reason; (3) divine law, spelled out in the Bible; and (4) human law, consisting of codes and customs.

This general theme continued into the late Middle Ages. Marsilius of Padua, rector of the University of Paris, in *Defensor pacis* (1324) taught that good government rests upon popular sovereignty, that the king is elected by the people and is responsible to them, and that he is bound by the laws. Indeed, William of Ockham argued that the power of the secular ruler was based on the consent of the people and is limited by law.

The tradition that the ruler's authority is derived from the people and that a ruler who violates the law may be resisted and removed was strengthened during the 16th century in the writings of a group known as the monarchomachs, who resisted the divine-right claims of the rising national monarchs. Such was the position of the French Huguenots, one of whose spokesmen, François Hotman, wrote in *Franco-Gallia* (1573) that the king derived his power from the people, not from God, that as a matter of history, the people are sovereign and have a right, through their representatives in the estates-general, to remove a tyrant from office. Similarly, Theodore Beza wrote that the people have natural rights that are protected by a contract entered into between the ruler and his subjects and that, if the king violates the contract, the people have a right to resist.

So powerful was this body of doctrine that even such a stout defender of the sovereignty of the ruler as Jean Bodin pointed out in his *Six Livres de la République* (1576) that there are limitations upon the king. He must secure his office in a lawful way; he may not deprive the family of its natural right to property; he may not impose taxation without the consent of the taxed; he is subject to the law of nature; he is limited by fundamental or constitutional principles, such as that which forbids alienation of the domain or any change in the succession to the throne; and he is limited by his oaths. This marks a long stride in the direction of modern concepts.

During the period of the Puritan Commonwealth in England in the mid-17th century, considerable efforts were made to state basic principles of constitutional law in specific documentary form. Repeated attempts in the Cromwell era to create a formal written constitution foundered on the inability of those engaged in the effort to reconcile the authority of the head of the state with the authority of the people as represented in Parliament. It was precisely in this Commonwealth period that James Harrington in his widely read *Commonwealth of Oceana* (1656) advocated the written constitution as an essential element of the ideal state. Harrington was well known in the American colonies, and contributed to the American taste for written constitutions.

Modern constitution making

Much of modern thinking about constitutionalism derives from the writings of the English common-sense philosopher John Locke (1632–1704). In his *Two Treatises of Government* (1690), he provided a classic analysis of the concept of the state that involves natural law and the social compact. He maintained that man originally lived according to the law of nature in what he described as a "state of perfect freedom" and equality. Although Locke did not believe, as Thomas Hobbes (1588–1679) had argued before him in *Leviathan* (1651), that the state of nature was intolerably bad, he did believe that it was not a very convenient condition because of the lack of "established, settled, known law"; the lack of "known and indifferent (impartial) judges"; and the absence of an executive to enforce judgments. Accordingly, by unanimous consent, men abandoned the state of nature and entered into a social compact that created a government designed to protect their natural rights. The only right surrendered in the process was the right to enforce the law. According to Locke, government rests upon consent of the governed and is bound to observe the terms of the compact that created it; if government violates its trust, men have a right to resort to revolution. This became the rational justification for the Glorious Revolution of 1688 in England, and it supplied the basic line of argument in defense of the revolution in the American colonies a century later. Thus, in the opening of the Declaration of Independence (1776) Thomas Jefferson affirmed that all men are created equal in that they are endowed by their creator with certain unalienable rights that government is instituted to secure; when government fails in accomplishing this basic purpose, the people have a right to abolish it and in its place institute a new government.

When independence from England was declared in 1776, the 13 former colonies promptly began to write constitutions spelling out the basic terms upon which the new state governments would function. In 1777, Congress prepared a written constitution for the whole country, the Articles of Confederation, which after ratification by the states was put into effect in March 1781. Since this constitution soon proved inadequate, a new constitution was drafted by a convention of delegates from the states in 1787 and became effective in 1789. This, the oldest surviving written national constitution, inaugurated a period of constitution writing that has continued to the present day. Most, though not all, national governmental entities now have written constitutions. In a broad sense, the idea of constitutionalism may be described as universal, though the content and effectiveness of the concept varies a great deal from country to country.

CHARACTERISTICS OF CONSTITUTIONS

Every state has a constitution, since every state functions on the basis of certain rules and principles. It has often been asserted that the United States has a written constitution but that the constitution of Great Britain is unwritten. This is true only in the sense that in the United States there is a formal document called the Constitution, whereas there is no such document in Great Britain. In fact, however, many parts of the British constitution exist in written form, whereas important aspects of the American constitution are wholly unwritten. The British constitution includes the Bill of Rights (1689), the Act of Settlement (1700–01), the Parliament Act of 1911, the successive Representation of the People acts (which extended the suffrage), the statutes dealing with the structure of the courts, the various local government acts, and many others. These are not ordinary statutes, even though they were adopted in the ordinary legislative way, and they are not codified within the structure of a single

orderly document. On the other hand, such institutions in the United States as the presidential cabinet and the system of political parties, though not even mentioned in the written constitution, are most certainly of constitutional significance. The presence or absence of a formal written document makes a difference, of course, but only one of degree. A single-document constitution has such advantages as greater precision, simplicity, and consistency. In a newly developing state such as Israel, on the other hand, the balance of advantage has been found to lie with an uncodified constitution evolving through the growth of custom and the medium of statutes. Experience suggests that some codified constitutions are much too detailed. An overlong constitution invites disputes and litigation, is rarely read or understood by the ordinary citizen, and injects too much rigidity in cases in which flexibility is often preferable. Since a very long constitution says too many things on too many subjects, it must be amended often, and this makes it still longer. The United States Constitution of 7,000 words is a model of brevity, whereas many of that country's state constitutions are much too long—the longest being that of the state of Louisiana, whose constitution now has about 255,000 words. The very new, modern constitutions of the recently admitted states of Alaska and Hawaii and of the Commonwealth of Puerto Rico have, significantly, very concise constitutions ranging from 9,000 to 15,000 words. The 1949 constitution of India, with 395 articles, is the wordiest of all national constitutions. In contrast, some of the world's new constitutions, such as those of Japan and Indonesia, are very short indeed.

Normative and nominal constitutions

Some constitutions are buttressed by powerful institutions such as an independent judiciary, whereas others, though committed to lofty principles, are not supported by governmental institutions endowed with the authority to defend these principles in concrete situations. Accordingly, many juristic writers distinguish between "normative" and "nominal" constitutions. A normative constitution is one that not only has the status of supreme law but is also fully activated and effective; it is habitually obeyed in the actual life of the state. A nominal constitution may express high aspirations, but it does not, in fact, reflect the political realities of the state. Article 125 of the 1936 constitution of the Soviet Union and article 87 of the 1954 constitution of the People's Republic of China both purport to guarantee freedom of speech, but in those countries even mild expressions of dissent are likely to be swiftly and sternly repressed. Where the written constitution is only nominal, behind the verbal facade will be found the real constitution containing the basic principles according to which power is exercised in actual fact. Thus in the Soviet Union the rules of the Communist Party describing its organs and functioning are more truly the constitution of that country than are the grand phrases of the 1936 Stalin constitution. Every state, in short, has a constitution, but in some a real constitution operates behind the facade of a nominal constitution.

Many juristic writers, notably Lord Bryce (1838–1922), have classified constitutions as being either "rigid" or "flexible" depending upon the methods by which they are changed. The British constitution was described as flexible because it can be amended by Parliament in the ordinary legislative way. Since the U.S. Constitution can be amended only by an extraordinary (two-thirds) majority of both houses of Congress plus ratification by three-fourths of the states, it has been characterized as a rigid constitution. This distinction may not be very significant; indeed, it may serve to obscure other more important aspects of constitutions. It is quite true that the British Parliament could, by the seemingly simple device of the enactment of a statute in the ordinary way, abolish the writ of habeas corpus or eliminate all aspects of local self-government, but on such fundamental institutions as habeas corpus and local self-government neither the Parliament nor the people of Britain are in actual fact very flexible.

In the long history of the state constitutions of the United States, one finds little, if any, correlation between the ease of amendment and the actual number of amendments adopted. States having identical methods of constitutional change often diverge considerably in the frequency with which the amending process has been utilized. States having what seem to be simple methods of change often have a history of fewer changes than other states that have sought to make the amending process more difficult. The decisive factors seem to be the generality or detail of the constitutional document, the tempo of social and economic change in the community, the class structure, and the nature of the struggle for power, which are quite unrelated to the formal processes of constitutional amendment.

Methods of constitutional amendment

Constitutions can be amended in many different ways, and these differences may be significant. The constitutions of Great Britain, Colombia, and Israel may be amended by the national legislature acting through an ordinary parliamentary majority. In some countries, including Argentina, Cambodia, and Chile, the constitution may be amended by the national lawmaking body but only through a special parliamentary majority. Where such majorities are needed there may be additional requirements, such as joint sessions of bicameral bodies, more than one debate, or the lapse of a certain amount of time between proposal and debate. Many constitutions cannot be amended by the legislature alone but require action by other bodies as well. Thus the United States Constitution can be amended only by (1) a two-thirds majority vote in each house of Congress or (2) a convention called by two-thirds of the states—with subsequent ratification, in either case, by the legislatures or especially elected conventions of three-fourths of the states. Still other constitutions can be amended only through some sort of popular voting: an amendment to the Australian constitution must secure, after an absolute majority in parliament, a majority vote in a popular referendum; the Belgian constitution may be amended by a two-thirds vote of both houses of parliament but only after a dissolution of that body and an intervening election; the Japanese constitution requires a two-thirds vote of the parliament followed by a popular referendum; after the Danish parliament has adopted a constitutional amendment it is dissolved, an election is held, and the new parliament must then vote in favour of the same amendment before it goes into effect; the Irish constitution may be amended by a simple majority of both houses of parliament subject to approval by a popular referendum.

The 50 states of the United States, all of which have written constitutions, exemplify in their history a vast amount of experience with problems and techniques of constitutional change. Generally speaking, state constitutional amendments are initiated by the legislature, although in 14 states amendments may also be initiated by popular petition. Where amendments are proposed by the legislative bodies, the size of the required vote is a simple majority (18 states) or a two-thirds vote (18 states) or a three-fifths vote (9 states). A few states have rather special provisions. In Connecticut, an amendment may be initiated by a majority vote in each house of the legislature in two sessions or a three-fourths vote in each house in one session; whereas Vermont requires a two-thirds vote in the Senate and a majority vote in the House on first passage and a majority in both houses on the second. In 13 states a constitutional amendment must be approved by two sessions of the legislature, and this is an alternative method in three other states. In all states but Delaware the amendment must be ratified by the electorate, usually by a simple majority of those voting.

Other differences among constitutions

Some constitutions may be characterized as indigenous, in the sense that they are rooted in the historical experiences of the states that produced them. Others are essentially imitative. Clearly the constitution of France's Fifth Republic (1958) is an indigenous document rooted in French political history. On the other hand, the Japanese constitution of 1947 is an imitative document imposed upon a nation defeated in war and controlled at the time by an occupying army.

The constitutions of the English-speaking countries have proved to be very stable. Many Latin American constitutions, on the other hand, have been exceedingly fragile. Among the new nations that emerged from the liquidation of the British Empire, some constitutions have proved to be quite stable while others have not. Thus the constitutions of India, Kenya, and Malaysia have functioned effectively with relatively few changes, whereas the constitutions of others, such as Pakistan, Ghana, and Nigeria, have not.

Some writers have contrasted "organizational" documents with "constitutionalist" documents. The organizational, such as those of Sri Lanka and Indonesia, are concerned mainly with describing the principal political institutions of the state. The constitutionalist documents, such as that of India, emphasize general principles, spell out the rights of the citizen, and are concerned with how these principles are to be enforced. Constitutions also differ fundamentally in how they deal with that most important institution, the political party. The constitutions of the Western liberal democracies do not mention parties at all. On the other hand, the constitutions of the Communist countries—*e.g.*, the Soviet Union, China, Albania, Romania, and Yugoslavia—have much to say about political parties. So do the constitutions of many of the new African states, such as Tanzania and Malawi. Where the central fact of the nation's political life is the domination of a single party endowed with a monopoly of power, it is to be anticipated that the constitution will not ignore the subject.

II. Unitary and federal systems

THE DISTINCTION BETWEEN UNITARY AND FEDERAL STATES

States may be described as unitary or federal depending upon the distribution of powers between central and local governments. Modern states, even rather small ones such as Denmark and Costa Rica, cannot be governed altogether from a central point. Almost all modern states find it necessary to have local as well as central governments. A unitary system differs from a federal system in the way this division of powers is made. In a unitary system the central government defines and by hypothesis may redefine the powers of local government. In a federal system the powers of local government derive from a constitution that is not subject to change by the central government itself. In other words, the powers of local government in a federal system are not at the mercy of the legislative body of the central government.

This does not mean that local governments in unitary states are without power or influence. Nor does it matter what the local governments are called. The principal local governments of the United States and Australia are called states, and the systems as a whole are federal in character. The principal local governments of India and West Germany are also called states, but the systems are at best quasi-federal in character. Italy, France, and Japan have unitary systems, but no constitutional significance attaches to the fact that the principal local governments in Italy are called provinces, in France *départements*, and in Japan prefectures.

The British and American systems

A comparison of the British and American systems of government will facilitate an explanation of the basic difference between unitary and federal systems. Since the British system is unitary, the powers of local government are defined by statutes enacted by Parliament, the legislative organ of the central government. This does not mean that local government is unimportant in Britain. On the contrary, the English counties and boroughs are significant units of government endowed with very real powers. It should be added that the unitary character of the British system does not warrant the inference that the people of the country do not have a high regard for their local institutions. It means only that as a matter of constitutional law the powers of local government are defined by acts of Parliament and that Parliament is free to expand or reduce or rearrange these powers. The system of the United States is federal because, as a matter of constitutional law, the powers of the states cannot be altered by the national government acting alone. In the Constitution the powers of the national government are enumerated, whereas the powers of the states are not. The states have what is left over, the so-called residual powers. But there is no special significance in this pattern of power distribution. In the Canadian system, which is also federal, the powers of the provinces are enumerated and the national government possesses the residual powers. Either way, the system is federal so long as local powers do not derive from the central government.

These distinctions are to some degree formal, as can be seen from a closer examination of actual experience in the life of the United States Constitution. Although it is quite true that the distribution of powers spelled out in that Constitution and confirmed by the explicit command of the Tenth Amendment cannot be altered formally by the central government acting alone, it is equally significant that the central government is the final judge of the scope of its delegated powers. If Congress enacts a statute, and the president signs it, and the Supreme Court finds, in the course of appropriate litigation, that it is constitutional, there is nothing, as a matter of law, that the states can do about it. Article VI of the federal Constitution specifically declares that all statutes enacted by Congress shall constitute the supreme law of the land, anything in the laws of the states to the contrary notwithstanding. To be sure, the Supreme Court has on occasion ruled a federal statute unconstitutional on the ground that it invaded the reserved powers of the states, but this has not happened very often. It must be remembered that the Supreme Court, as umpire of the federal system, is a national institution in no way subject to state control; in the long sweep of American legal history the court has on the whole taken a nationalist rather than a state point of view. Nor is this development unique to the United States. In all federal systems in modern times there has been a drift in the direction of stronger and more pervasive central government, often at the expense of the powers of local government.

Swiss federalism

The best examples of federalism in the contemporary world are the governmental systems of the United States, Canada, Australia, and Switzerland. The Swiss system illustrates many important facets of the federal idea. A federation of the 22 cantons was the only feasible way of creating national unity without eliminating traditional local autonomy. The federal system also facilitated the unification of the country by permitting regional adjustment to religious and linguistic differences. In all countries where federalism has been adopted, the main thrust has been to seek an equilibrium between unity and diversity, to achieve centralized authority for the nation as a whole without sacrificing local loyalties, cultures, and institutions. The 1848 constitution of Switzerland, as extensively revised in 1874, delegates certain powers to the national government, the cantons retaining all residual powers. The powers delegated to the central government are greater than those delegated to the national government by the United States Constitution, particularly with respect to governmental monopolies, the definition of criminal and commercial law, conservation, and social welfare. Furthermore, there is no supreme court in Switzerland with the power to declare a statute of the central government unconstitutional, and federal law always prevails over any contrary cantonal law. Even so, the cantons exercise important powers with regard to the administration of law and order, health and sanitation, public works, and education. In addition, the federal laws are, for the most part, administered by the cantons; for example, the national criminal code is enforced in the cantonal courts. The federal character of the Swiss constitution is also reflected in the nature of the amending process, for all constitutional amendments must be approved by a majority vote in a national referendum, with a favourable decision in a majority of the cantons. Amendments may be initiated by a popular petition signed by 50,000 voters or by the national legislature, and it is to be noted that the cantons are represented equally in the upper house, the Council of States.

QUASI-FEDERAL SYSTEMS

Germany and India

Several governmental systems that are usually called federal are, in fact, only quasi-federal. That the Federal Republic of Germany is at best only partly federal is reflected in the magnitude of the powers vested in the central government and in the fact that if there is any conflict between the laws of that government and those of the states, or *Länder*, the laws of the national government must always prevail. The constitution gives the central government exclusive legislative powers in such areas as foreign affairs, the postal system, citizenship, railroads, and the currency, and concurrent powers with the *Länder* over many significant areas such as the definition of civil and criminal law, control of the economy, labour, agriculture, public welfare, and shipping on the high seas. The ten *Länder* are given primary responsibility over education and cultural affairs, and they administer most federal laws. The constitution may be amended by a two-thirds vote in each house of parliament, and there is a special federal constitutional court endowed with the power of judicial review. This is a federal system that leans strongly upon the national government. To a very considerable extent it is a system of centralized legislative power tempered by local administration.

The pattern of government in India is often described as federal, for it is a union of 21 states and 8 union territories; the constitution provides, in elaborate detail, for a territorial division of powers. The various subjects of legislation are grouped in three lists: a union list of 97 subjects (including defense, foreign affairs, currency, banking, communications, and customs), over which the union parliament has exclusive jurisdiction; a state list of 66 subjects (including police and public order, agriculture, education, public health, and local government), over which the state legislatures have exclusive authority; and a concurrent list of 47 subjects (including economic and social planning, labour, and price controls), over which both the central and state governments may legislate. But this division of powers is seriously qualified by the fact that the national Parliament may legislate on any subject in an emergency or if it deems a subject to have national importance. The governor of each state is appointed by the union president, and Parliament has the power to abolish the legislative council of any state and create a new one. In the light of these facts, the government of India cannot be characterized as a truly federal system. The same is true of the Soviet Union, whose constitution purports to provide for a union of various socialist republics but where power is so highly centralized as to warrant the conclusion that the system as a whole is only partly federal. The governmental systems of Argentina, Austria, Brazil, Burma, Mexico, Libya, South Africa, Venezuela, and Yugoslavia are all described in their constitutions as federal in nature, but they are doubtfully so. In some of them there is a considerable amount of local administration of national laws, and in others the central governments have devolved significant powers upon the local governments, but these are quite compatible with unitary government.

UNITARY SYSTEMS

Italy, France, and Britain

The essential characteristics of unitary systems are suggested by an analysis of the Italian constitution of 1948. Though the Italian system of government has been unitary ever since the establishment of the modern state in 1870, this has never meant that local government in Italy is unimportant. There are 20 regions, 94 provinces, and 8,056 communes (in 1971), each having an elective council. Although each region is said to be autonomous, the central government exercises many controls; a commissioner selected by the national government represents it in the region; he must approve of regional legislation, but if he disapproves his decision may be appealed to the constitutional court or to Parliament. The regional government has no independent taxing power, all of its financial authority being spelled out in national statutes. Furthermore, the president of the republic has the power to dissolve any regional council. There is even greater central control over the provinces and the communes. Most of the powers exercised by them are set out in national statutes, and they depend very heavily upon financial subventions from the national state. The chief representative of the national state in the provinces is a prefect, who is chosen by and is responsible to the government of the nation. He has vast powers in regard to the maintenance of law and order and may veto any act or financial decision of the provincial and communal governments. The prefect also has the authority to suspend any local official or employee and to dissolve the provincial or communal councils. Thus, although there is a great deal of important activity in Italy below the national level, the relationship between the central and local governments is, as a matter of law, one of dominance and subservience—the hallmark of the unitary pattern.

The French and British systems of government have always been unitary, even though local units in England have more power than those in France. In both countries the powers of local government are spelled out in national statutes. The same is true in New Zealand, where the local governments (counties, boroughs, and town districts) derive their authority from statutes enacted by Parliament. Although the Republic of South Africa is often described as a union of provinces, each of which has its own elected council, the system as a whole is unitary; the administrators of the provinces are appointed by the president in council, and all ordinances of the provincial councils are subject to veto by the president in council.

III. Executives and legislatures

THE CONSTITUTION AND THE EXECUTIVE

Monarchical and republican executives

States may be classified as monarchical or republican. From another point of view, they may be described as having presidential or parliamentary executives (see below).

Though the institution of monarchy is as old as the recorded history of man, the modern age has been moving steadily in the direction of republican government. Depending on how they are counted, there are today about 28 or 30 monarchies. Many monarchs, as in Great Britain, Japan, the Scandinavian countries, and the Low Countries, are best described as constitutional monarchs. This means that they are mainly titular heads of state and do not in fact possess important powers of government. Most of the executive powers are in the hands of ministers, headed by a prime minister, who are politically responsible to the parliament and not to the king. The executive powers of government in Great Britain, for example, are exercised by ministers who hold their offices by virtue of the fact that they command the support of a majority in the popularly elected House of Commons. The monarch can act only on the advice of his ministers; he cannot exercise an independent will of his own. The position of the monarchs in Scandinavia and the Low Countries is precisely the same as that of the British: they reign but they do not rule. In countries where no political party has a majority of its own in the parliament, the king may exercise some discretion in deciding whom to invite to serve as prime minister and to form a government. Even so, since the king must first consult with the various party leaders, he is not likely to have much discretion. In a country with a stable two-party system, all the king can do is offer the prime ministership to the leader of the majority party. A constitutional monarch is the head of the state, not of the government. Standing above party and the active political controversies of the day, he is a focus of national loyalty and a useful symbol of the nation's unity and its historical past.

In a few monarchies, however—for example, those of Ethiopia, Jordan, Iran, and Saudi Arabia—the king exercises real powers of government. The ministers are chosen by and are responsible only to the king, rather than to some elective parliamentary body. The constitution of Ethiopia declares that "supreme power rests in

the hands of the emperor," that his person is "sacred," and that his power is "indisputable." Hereditary rulers with this sort of personal power were quite common in the 18th century, but they are rare today.

Presidential and parliamentary systems

Far more significant than the distinction between monarchy and republicanism is the contrast between presidential and parliamentary executives. Since the United States has for long been the world's leading exponent of presidential government and Great Britain the oldest and most successful practitioner of parliamentary government, their systems may be taken as models with which the systems of other countries can be compared.

The American system is based upon the concept of separation of powers: the executive, legislative, and judicial powers of government are vested by the Constitution in three separate branches. The president is not selected by Congress, nor is he a member of Congress. He has a fixed term of office of four years, and he holds it no matter how his legislative program fares in Congress and whether or not his political party controls either or both houses of Congress. The members of the Cabinet are chosen by the president and are politically responsible to him. The Constitution does not permit them to be members of Congress; it provides that "no Person holding any Office under the United States, shall be a member of either House during his Continuance in Office."

The parliamentary executive system proceeds upon radically different assumptions. In Great Britain, whose system many countries have chosen to emulate, the executive officers of the state are not separated from the legislative branch. On the contrary, the British Cabinet may be described as the leading committee of Parliament. Although the prime minister, the head of the government, could at one time hold a seat in either the House of Lords or the House of Commons, the contemporary convention is that he be a member of the House of Commons. The other ministers who make up the cabinet must be members of one or the other house of Parliament. If the prime minister wants someone who is not in Parliament to serve in his Cabinet, he must either appoint him to the peerage or find a vacancy in the House of Commons to which he can be elected.

Whereas in the American system, with its separation of powers, the chief executive and his Cabinet officers are institutionally apart from the legislature, in Great Britain the ministers of the crown hold their powerful executive positions only so long as they enjoy the support of a majority of the House of Commons. A Cabinet that loses that support must either dissolve the House and call a new election, thus in effect putting the issue to the voters, or resign and permit others to form a government. In the 20th century most changes in power have occurred as a result of the outcome of a general election.

It follows that in the British Parliament the prime minister and his Cabinet colleagues are fully in charge. They are responsible, as the guiding committee of Parliament, for the preparation and enactment of most legislation and of the budget. There can be no permanent or serious conflict between the House of Commons and the Cabinet, for responsibility means that the government of the day must either prevail or give way to another government. Thus the deadlocks between the chief executive and the Congress that are a frequent occurrence in the United States cannot occur under the parliamentary system.

A system may appear to be parliamentary or presidential without actually being either. In Latin American and African states, many presidential systems have been converted into military dictatorships. The Communist governments of central and eastern Europe are parliamentary in form, but power is effectively in the hands of a party leader. The fact that the head of state is called a president does not mean that the system is presidential. The head of the Italian state, for example, is a president elected for a seven-year term by a joint session of Parliament, but the executive power is in fact exercised by a prime minister and other Cabinet ministers who are responsible to Parliament. In former monarchies that now have parliamentary systems, the abolition of monarchy invariably led to the substitution of a president for the hereditary ruler.

Hybrid forms

There are some hybrid forms of government that combine features of both presidential and parliamentary systems. France's Fifth Republic (1958) is a good example. According to the terms of a constitutional amendment adopted in 1962, the president of the republic is elected by direct vote of the people for a seven-year term. This gives him an enormous moral power derived from the fact that he is a product of universal suffrage. Although in the exercise of some of his powers the president needs the signature of the prime minister or of some other minister, he has great substantive powers of his own: he appoints the prime minister; he dominates the management of foreign relations; he may dissolve the National Assembly, though not more often than once a year; and he possesses vast emergency powers. The cabinet, called the Council of Ministers, is presided over by the president. Members of the council cannot be members of Parliament, but they have access to both chambers; and they may speak there, though they do not vote. The council is responsible to the National Assembly and can be defeated by censure motion. Thus the French system of government is neither presidential nor parliamentary in form; it combines elements of both in a unique fashion.

The governmental system of the Federal Republic of Germany is mainly parliamentary but with some interesting variations. The head of state is the president, elected for a five-year term by a body known as the Federal Assembly, which consists of members of the lower house of Parliament (the Bundestag), plus a Bundesrat chosen by the legislatures of the *Länder*. The head of government is the Chancellor, who must have majority support in the Bundestag; but only he, and not the Cabinet, is responsible. The Bundestag can express its lack of confidence in the chancellor only by electing a successor by majority vote. In this fashion it was hoped to eliminate some of the weaknesses of the Weimar constitution of the post-World War I period. The chancellor may seek a vote of confidence, and if he fails he may ask the president to dissolve the Bundestag and call a new election.

The governmental system of India is more typically parliamentary in form. The president is elected for a five-year term by an electoral college consisting of all members of the national Parliament and of the state legislative assemblies. He can act only on the advice of ministers who are responsible to Parliament. The lower House of the People is subject to dissolution. Similarly, the president of the Republic of South Africa is elected for a seven-year term by an electoral college consisting of a joint meeting of the two houses of Parliament. The executive power is in fact exercised by the prime minister and ministerial colleagues whom he selects, and they are responsible to Parliament, both houses of which are subject to dissolution.

The Swiss executive is quite unique. The national Cabinet consists of seven persons elected for four-year terms by the parliament (the Federal Assembly). They are elected as individuals, and they are never forced to resign; in fact, they are almost always re-elected, some serving for as long as 25 to 30 years. A disagreement with the Federal Assembly leads neither to resignation of the Cabinet nor to a dissolution of the parliament; the ministers simply adjust their positions to conform with the wishes of the parliamentary majority. This does not mean that the Cabinet is not an important body; as a group it originates most new legislation, and its members, as individuals, head up the great departments of government. Each year the parliament elects a member of the cabinet to serve as president of the confederation. The president is chairman of the Cabinet and titular head of state. Although this system has not been adopted in other countries, it has worked very well for the Swiss.

UNICAMERAL AND BICAMERAL LEGISLATURES

A central feature of any constitution is the legislature. It may be a unicameral body with one chamber or a bicameral body with two chambers.

Unicameral legislatures are to be found in small states with unitary systems of government, among them Denmark, Finland, Israel, and New Zealand, or in very tiny states such as Andorra, Luxembourg, and Liechtenstein. They are also found in states undergoing very rapid social and political change, such as the Socialist states of eastern Europe and newly freed and developing states such as Kenya and Indonesia. Federal and quasi-federal states, whether large or small, tend to have bicameral legislatures, one house of which represents the main territorial subdivisions. The classic example is the Congress of the United States, which consists of a House of Representatives of 435 members elected for two-year terms from single-member districts of approximately equal population, and a Senate consisting of two persons of each state elected by the voters of the state. The fact that all states are represented equally in the Senate, regardless of their size, stems from the federalistic character of the American union.

Bicameralism and federalism

The federal character of the Swiss constitution is likewise reflected in the makeup of the nation's central legislature, which is bicameral. One house, the National Council, consists of 200 members apportioned among the cantons according to population; the other house, the Council of States, consists of 44 members elected by direct ballot—two from each canton. The Canadian Parliament is also bicameral, but both houses are apportioned on the basis of population.

Bicameralism is also characteristic of governmental systems that are best described as quasi-federal. Here, too, bicameralism is expressive of the territorial subdivisions that are joined together to form the national state. The parliament of the Federal Republic of Germany includes a Bundestag elected by general suffrage and a Bundesrat appointed by the ten *Länder*, or state governments. Since most members of the Bundesrat are ministers in the *Land* governments, the chamber is a significant link between the *Länder* and the central government. The Parliament of India includes a Council of States and a House of the People, the former elected by the state legislative assemblies and the latter elected directly from territorial constituencies. The Parliament of the Republic of South Africa is also bicameral, but both houses are apportioned more or less on the basis of population.

Bicameralism and unitary systems

A unitary governmental system does not imply unicameralism in the legislature. Most legislatures of unitary states are, in fact, bicameral, though one chamber is usually more powerful than the other. This is true for the world's oldest and most successful parliament, that of Great Britain, which consists of a House of Lords and House of Commons. The House of Commons has become by far the more powerful of the two chambers, and the Cabinet is politically responsible only to it. The House of Lords has no control over finances and, with respect to other legislation, only a modest suspensory veto, which can be easily overcome in the House of Commons by a second vote at an early date. The parliaments of Italy, Japan, and France are also bicameral. In the United States, all of the 50 states except Nebraska have bicameral legislatures, even though their governmental systems are unitary. In all states the two houses have equal legislative authority, but the so-called upper houses, usually called senates, have the special function of confirming the governors' appointments.

IV. Judicial review: the unique role of the United States judiciary

The power of United States courts to rule on the constitutionality of legislation and to refuse to enforce legislation that in their judgment violates the Constitution has come to be known as judicial review. Few courts in the world have this extraordinary power. Thus a statute passed by the British Parliament may indeed violate the constitution of the realm—for example, a gross statutory invasion of free speech—but no British court has the right to refuse to enforce it on constitutional grounds. Parliament is sovereign, but in the United States the Constitution, as construed by the courts and as the embodiment of the will of the whole people, is sovereign. Therefore, it is necessary here to offer a separate survey of the almost unique position of the U.S. Supreme Court in constitutional law.

Judicial review is not explicitly mentioned in the Constitution and is itself a product of judicial construction. Chief Justice John Marshall in *Marbury* v. *Madison* (1803) reasoned that since the Constitution is the supreme law of the land (it says as much in Article VI) and since it is the province of the Supreme Court to uphold the law, it follows that when a statute is inconsistent with a provision of the Constitution the latter must be preferred to the former because it declares the law of superior obligation. Without judicial review, he argued, a written constitution would be quite futile as a means of limiting the abuses of governmental power.

Grounds for review

Although the Supreme Court possesses the ultimate power of declaring federal and state legislation invalid on constitutional grounds, it has exercised it with great restraint. Only about 85 federal statutes had by the early 1970s been so declared, though the number of state statutes held invalid was much larger. The court exercises this power only if necessary to decide cases and controversies and does not give mere advice to Congress or the executive in the form of advisory opinions. Judicial review, therefore, always grows out of concrete litigation in which the rights of adversary parties are actually involved. Furthermore, the Supreme Court imposes a number of its own restraints upon the exercise of judicial review. It always begins with the assumption that the legislative body did not intend to violate the Constitution when it adopted the challenged statute. Therefore the burden of proof rests upon the party who questions the validity of the statute. If the court can possibly decide the case without ruling on the constitutionality of the statute, it will prefer to do so; it comes to constitutional questions last, not first, and seeks to dispose of cases on minimal grounds. If a statute is susceptible of any reasonable interpretation that will save it, the court will adopt that interpretation. In cases of reasonable doubt, the constitutionality of legislation is upheld.

The court takes a strict view of the question of who is eligible to raise constitutional questions to litigation: to be able to sue in court, a party must have a direct and substantial interest at stake. Thus it was held that a mere citizen and lawyer has insufficient interest to question in court the qualifications of a particular justice for his office (*Ex parte Levitt*, 1937). A mere taxpayer qua taxpayer does not have sufficient interest in the United States Treasury to attack in court the validity of a federal appropriation act (*Frothingham* v. *Mellon*, 1923). This rule, which bars federal taxpayer suits, closed the door to a possible flood of litigation that would have been most embarrassing to the government. The court opened the door a little bit, however, in 1968 in *Flast* v. *Cohen*, in which it was held that a federal income taxpayer had standing to challenge an expenditure of federal funds designed to finance instruction in religious schools on the ground that the First Amendment, which forbids an "established" or national church, operated as a specific constitutional limitation upon the exercise by Congress of its taxing and spending powers. But a statute may be assailed only by one who relies upon an alleged invasion of his own constitutional rights; one does not have standing to sue in behalf of others.

Finally, the thrust of judicial review is limited by the doctrine of the political, nonjusticiable question. Certain questions arising under the Constitution are regarded by the court as being political in character and therefore outside the scope of judicial action. For example, the Constitution (IV, 4) declares that "The United States shall guarantee to every State in this Union a Republican Form of Government." But it was early decided (*Luther* v. *Borden*, 1849) that whether a particular state government is republican in form is for the political branches of the government—that is, Congress and the president—to decide, not the courts. Similarly, the president's

duty to see to the faithful execution of the laws is political and therefore not subject to judicial process (*Mississippi* v. *Johnson*, 1867).

Many military questions, such as the necessity for calling out the militia, and foreign policy questions, such as the government's power to recognize a particular foreign government or international boundary or treaty, are considered political in character and therefore nonjusticiable.

THE SUPREME COURT AND THE SEPARATION OF FEDERAL POWERS

The framers of the Constitution accepted as an unchallengeable maxim that the only way to avoid governmental tyranny is to put the legislative, executive, and judicial powers in separate departments. This separation of powers is not only a political theory about the proper organization of government but also a doctrine of constitutional law involving the Supreme Court. It is in accordance with this principle that the federal courts decline to perform nonjudicial functions and that the exercise by Congress of nonlegislative functions in connection with legislative investigations tends to be judicially censured.

Since the legislative power is vested in Congress, the lawmaking power as a whole cannot be delegated to the president. It was on this theory that the court in 1935 invalidated the National Industrial Recovery Act (NIRA), holding that the code-making authority given the president vested in him a virtually unfettered discretion to make law (*Schechter Poultry Corp.* v. *United States*). This was an unusual holding, since government without large-scale delegations is quite impossible, and as Chief Justice Hughes said in the case, the court had "repeatedly recognized the necessity of adapting legislation to complex conditions involving a host of details with which the national legislature cannot deal directly." This is well illustrated in the great regulatory commissions, such as the Interstate Commerce Commission (ICC) and the Federal Communications Commission (FCC), established by Congress to deal with exceedingly complex and rapidly changing sectors of the economy. All of these agencies operate under the authority of broad delegations of power by Congress and necessarily exercise legislative and judicial, as well as administrative, functions.

The steady growth in the power and prestige of the office of president became a leading feature of U.S. constitutional government in the 20th century. Contrary to 18th-century assumptions about the separation of powers, the president takes the initiative in proposing legislation; and, indeed, his record as chief executive is more likely to be judged by the success of his legislative program than by any other event. The president possesses vast delegated legislative powers given him in a multitude of statutes. Nevertheless, there are limits to what the president can do, dramatically illustrated in the 1952 holding of the Supreme Court that President Harry S. Truman's seizure of the steel mills without statutory authorization to avoid a strike during the Korean War was illegal (*Youngstown Sheet & Tube Co.* v. *Sawyer*). The court ruled that since Congress could have authorized the seizure under its various powers, in acting without permission of a statute the President usurped legislative power in violation of the separation of powers principle. It rejected the contention that the president could seize private property under his authority as commander in chief and as chief executive, his power being only to enforce the laws, not to make them. The Constitution clearly vests the lawmaking function in Congress, and while Congress undoubtedly has the right to authorize the taking of private property for public use under its power of eminent domain, "the Constitution did not subject this lawmaking power of Congress to presidential or military supervision or control."

THE SUPREME COURT AND THE FEDERAL SYSTEM

A central fact of the U.S. Constitution is that it creates a federal system under which the powers of government are divided between the national government and the states. As noted earlier, the national government has those constitutional powers that are delegated to it; the states, unless they are otherwise restricted, possess all of the remaining powers of government, often referred to as the residual powers.

Defining the national powers

Although the national government is limited to its enumerated powers, several important facts about these powers must be noted: (1) The national Constitution, and the statutes and treaties adopted pursuant to it, constitute, in the language of Article VI, "the supreme Law of the Land; . . . any Thing in the Constitution or Laws of any State to the Contrary notwithstanding." This means that any state law, otherwise valid, must yield if contrary to a valid federal law, since the latter is the supreme law of the land. (2) Although the national government is limited to the exercise of enumerated powers, these powers are spelled out in broad terms, and the "elastic" clause (I, 8) states that Congress shall have the authority "To make all Laws which shall be necessary and proper for carrying into Execution" the various powers vested in the national government by the Constitution. It follows that in addition to the specified powers, Congress possesses implied powers, a proposition definitively established by Chief Justice John Marshall in the celebrated case of *McCulloch* v. *Maryland* (1819). Here the Supreme Court held that Congress had the power to incorporate a national bank, in spite of the fact that the Constitution is silent on both the creation of corporations and the chartering of banks. It was concluded that since a national bank would facilitate the accomplishment of purposes confided to the national government, such as the collection of taxes and the maintenance of armed forces, Congress had a choice of means to achieve these proper ends. "Let the end be legitimate," said the chief justice, "let it be within the scope of the constitution, and all means which are appropriate, which are plainly adapted to that end, which are not prohibited, but consist with the letter and spirit of the constitution, are constitutional." This doctrine of implied powers became a powerful force in supplying the constitutional basis for the steady growth, over the years, of national power. (3) In the field of foreign affairs the doctrine of enumerated powers breaks down almost entirely, since the Supreme Court takes the view that whether enumerated or not, the national government, from the very nature of things, has full responsibility over foreign affairs and has the inherent power to do all things that sovereign nations customarily do in this area. Thus the court early took the position that the national government has unlimited power to exclude aliens from the country and very wide authority to deport them on any terms it desires, though the Constitution says not a word on this subject. Similarly, in 1936, in a decision upholding an arms embargo, the court went so far as to say that in contrast with the enumerated powers exercised internally, the federal powers of external sovereignty do not even depend upon affirmative grants of the Constitution.

Whether the national government or the states have overstepped the boundaries of their constitutional authority is ultimately a legal question for the Supreme Court to decide, if the issues can be brought to it in proper litigation. In this sense the Supreme Court may be said to be the umpire of the federal system. It has on occasion declared acts of Congress invalid as invading the reserved powers of the states, and it has often held state laws unconstitutional as being contrary to the supreme law of the land. Nevertheless, though the Supreme Court is the umpire of the federal system, it is chosen by only one side; it is a national, and in no sense at all a state, institution, and on the whole its leanings are toward the national position and an ever-expanding conception of national power.

In still another sense the Supreme Court is the umpire of the federal system, for it is given original jurisdiction by the Constitution over disputes between the states. The last recourse a state has in such circumstances is to sue

in the Supreme Court, and by the late 1950s every state in the union but one had been at least once a plaintiff or defendant in such lawsuits, involving such difficult questions as boundaries, debts, water rights, and water pollution. The court performs an important function in successfully adjudicating disputes, which might lead to the use of force in the case of independent nations similarly involved.

The competing concepts of federal supremacy and states' rights were brought early in U.S. history into sharp focus in the field of commercial regulation. The commerce clause (I, 8) of the Constitution simply authorizes Congress "To regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes." In the formative years of the republic the Supreme Court took a broad view of national power under this clause, as in the seminal case of *Gibbons* v. *Ogden* (1824), which held that the power of Congress over commerce includes navigation. As new methods of interstate transportation and communication came into use, ranging from the stagecoach, the railroad, and the telegraph line to the ticker tape, trucking, radio, airlines, and television, the court interpreted the national power to keep pace with the times.

Balancing national and state regulatory powers

The court nevertheless has not regarded the commerce power as being exclusively national, the states being up to a point conceded the right also to regulate interstate commerce. The problem was to locate the point. After considerable fumbling the court hit upon a formula (*Cooley* v. *Board of Wardens of the Port of Philadelphia*, 1851). Upholding state regulation of pilotage, the court ruled that the states may regulate those aspects of interstate commerce that Congress has not attempted to regulate and that do not "imperatively [demand] a single uniform rule" for the whole country. In areas in which Congress has been silent and diversity of treatment is desirable, the states enjoy the power of concurrent regulation; but where uniformity of treatment is imperative, the power of congress is "exclusive." In balancing competing local and national needs in commerce, the Supreme Court has since exercised a wide discretion in considering the equities of each case.

THE SUPREME COURT AND INDIVIDUAL RIGHTS

The national government is obliged by many provisions of the Constitution to respect the basic rights of man. Some civil rights limitations were imposed on this government in the original document, notably in the provisions guaranteeing the writ of habeas corpus and forbidding bills of attainder (legislative trials) and ex post facto laws and guaranteeing trial by jury in criminal cases. But the most significant limitations of this character upon the national government were added in 1791 with the ratification of the first ten amendments, the Bill of Rights. Here are guaranteed basic rights of conscience, such as freedom of religion, speech, press, and petition; fundamental guarantees of fair procedure for persons accused of crime, such as protection against unreasonable searches and seizures, compulsory self-incrimination and excessive bail, a speedy and public trial by a local, impartial jury before an impartial judge, and representation by counsel; and rights of private property, as in the provision that private property shall not be taken for public use without just compensation.

Applying the Fourteenth Amendment

For the protection of such rights as these against state, as opposed to federal, action, the citizen originally had to look to the state constitution, each state constitution possessing a more or less elaborate bill of rights enforceable by the state courts. But an important new federal limitation on the states was added to the national constitution with the ratification, in 1868, of the Fourteenth Amendment, by which the states are forbidden to deny any person life, liberty, or property without due process of law or to deny to any person the equal protection of the laws. By a process of interpretation at the hands of the Supreme Court these two clauses took on increased meaning.

The liberty that the due process clause of the Fourteenth Amendment guarantees against state action has been held to include the liberties of religion, speech, and press that the First Amendment protects against violation by the national government. Similarly, some of the guarantees of a fair trial in other parts of the federal Bill of Rights, such as the defendant's right to an impartial judge and the assistance of counsel, have also been judicially absorbed into the Fourteenth Amendment. Indeed, a minority of Supreme Court justices argued that the Fourteenth Amendment was intended to incorporate as limitations upon the states all of the provisions of the federal Bill of Rights. But a majority of the court clung to the position that not all of the guarantees of the federal Bill of Rights are so fundamental to justice that they are required by due process of law. Thus, to cite an early example, the Fifth Amendment guarantee of indictment by grand jury was held inapplicable to the states (*Hurtado* v. *California*, 1884), and in fact most states abandoned this type of indictment in favour of allowing the public prosecutor to bring criminal charges. A number of the purely procedural guarantees of the federal Bill of Rights have been held inapplicable to the states, although they of course bind the agencies of the national government.

The other great limitation on the states in the Fourteenth Amendment, that no state shall deny any person within its jurisdiction the equal protection of the laws, has a long and important history in U.S. constitutional law. Thus over the years the Supreme Court has set aside many criminal convictions of Negroes on the ground that the state courts that convicted them arbitrarily and systematically excluded Negroes from their juries, this being construed as a denial of equal justice to members of the excluded group. By all odds the most spectacular and controversial application of the equal protection clause came on May 17, 1954, when the court ruled that states that practiced racial segregation in the public schools violated the Constitution (*Brown* v. *Board of Education*). In so ruling the court set aside the 1896 ruling (*Plessy* v. *Ferguson*) that "separate but equal" facilities satisfied the command of the Fourteenth Amendment, holding by unanimous vote that separate educational facilities, by creating feelings of inferiority, are inherently unequal.

It remains to be noted that none of the great constitutional rights of conscience, however vital to a free society, is absolute in character. Thus, while the constitutional guarantee of freedom of religion goes a long way, it does not serve to protect acts judged to be morally licentious, such as polygamous marriages. Children cannot be required to execute a flag salute forbidden by religious belief (*West Virginia State Board of Education* v. *Barnette*, 1943), but parents are not free to ignore child labour laws on the ground of religious practice (*Prince* v. *Massachusetts*, 1944). Similarly, freedom of speech, often defended by the courts, does not extend to the seditious utterances of a conspiracy that, in the considered opinion of Congress, poses a clear and present danger to the safety of the republic (*Dennis* v. *United States*, 1951). But the court has emphasized that the act of Congress on this subject, the Smith Act, does not forbid mere advocacy of abstract doctrine but only incitement to action designed to accomplish the illegal purpose of overthrowing the government (*Yates* v. *United States*, 1957). The state is not free to license the privilege of giving a speech (*Thomas* v. *Collins*, 1945), but it may punish utterance of "fighting words," which may lead to breaches of the peace (*Chaplinsky* v. *New Hampshire*, 1942), or the publication of obscene matter (*Roth* v. *United States*, 1957).

There was much concern at mid-20th century with the rights of persons accused of crime. The right of an indigent defendant to representation by counsel has been judicially underscored, and the courts appeared increasingly meticulous to safeguard the right to reasonable bail and the right to a fair trial, including a public, speedy trial before an unbiassed judge and a jury free of mob domination. In 1966 a Cleveland osteopath, Samuel H. Sheppard, won a reversal by the U.S. Supreme Court of his conviction on the ground that his trial was prejudiced by excessive newspaper publicity. The uses of the writ of

habeas corpus were steadily expanded. Some rights, such as the privilege against self-incrimination, came under heavy public fire, but the proposition that no one should be compelled to testify himself into jail retained its vitality in U.S. courts. Everywhere, in criminal cases, the burden of proof remained on the prosecution, and the accused carried with him the presumption of innocence.

BIBLIOGRAPHY For the philosophical and historical origins of the concept of constitutionalism, useful works are F.D. WORMUTH, *The Origins of Modern Constitutionalism* (1949); C.J. FRIEDRICH, *Constitutional Government and Democracy: Theory and Practice in Europe and America* (1950); J. BRYCE, *Studies in History and Jurisprudence*, vol. 1 (1901), see esp. the essay, "Flexible and Rigid Constitutions." General collections of constitutions include: A.J. PEASLEE, *Constitutions of Nations*, 3rd ed., 3 vol. (1968); W.G. ANDREWS (ed.), *Constitutions and Constitutionalism* (1961); and L. WOLF-PHILLIPS (ed.), *Constitutions of Modern States* (1968). On the issue of constitutional stability, see B. AKZIN, "On the Stability and Reality of Constitutions," in R. BACHI (ed.), *Studies in Economic and Social Sciences* (1956). For a comparative study of representative institutions, see M. AMELLER (ed.), *Parliaments* (1966). A highly perceptive analysis of federalism as a constitutional form is W.S. LIVINGSTON, *Federalism and Constitutional Change* (1956). On the separation of powers, see M.J.C. VILE, *Constitutionalism and the Separation of Powers* (1967).

For surveys of modern constitutional trends, see J.A. HAWGOOD, *Modern Constitutions Since 1787* (1939); H.J. SPIRO, *Government by Constitution* (1959); C.F. STRONG, *Modern Political Constitutions*, 6th rev. ed. (1963); K.C. WHEARE, *Modern Constitutions*, 2nd rev. ed. (1966); and A.J. ZURCHER (ed.), *Constitutions and Constitutional Trends since World War II*, 2nd ed. (1955).

For studies of the constitutions of particular groups of countries, see S.A. DE SMITH, *The New Commonwealth and its Constitutions* (1964); M.C. NEEDLER (ed.), *Political Systems of Latin America* (1964); W.F. ABBOUSHI, *Political Systems of the Middle East in the 20th Century* (1970); and H.G. SKILLING, *The Governments of Communist East Europe* (1966). Surveys of individual countries include: G.A. CODDING, JR., *The Federal Government of Switzerland* (1961); N. KOGAN, *The Government of Italy* (1962); W. PICKLES, *The French Constitution of October 4, 1958* (1960); E. PLISCHKE, *The Contemporary Government of Germany* (1961); A. GLEDHILL, *The Republic of India*, 2nd ed. (1964); J.D. LEGGE, *Indonesia* (1965); A.W. BURKS, *The Government of Japan* (1961); M. FAINSOD, *How Russia Is Ruled*, rev. ed. (1963); E.C.S. WADE and G.G. PHILLIPS, *Constitutional Law*, 7th ed. (1965); and A.E. SUTHERLAND, *Constitutionalism in America* (1965).

General works on the United States Constitution include: E.S. CORWIN, *The Constitution and What It Means Today*, 12th ed. (1958); A.N. HOLCOMBE, *Our More Perfect Union* (1950); and C.H. PRITCHETT, *The American Constitution* (1959). For its history, see M. FARRAND, *The Framing of the Constitution of the United States* (1913); C.B. SWISHER, *American Constitutional Development*, 2nd ed. (1954); and A.H. KELLY and W.A. HARBISON, *The American Constitution: Its Origins and Development* (1948).

(D.Fe.)

Constitution and Constitutional Government

The general idea of a constitution and of constitutionalism originated with the ancient Greeks and especially in the systematic, theoretical, normative, and descriptive writings of Aristotle. In his *Politics*, *Nicomachean Ethics*, *Constitution of Athens*, and other works, Aristotle used the Greek word for constitution (*politeia*) in several different senses. The simplest and most neutral of these was "the arrangement of the offices in a *polis*" (state). In this sense of the word, every state has a constitution, no matter how badly or erratically governed it may be.

THEORIES ABOUT CONSTITUTIONS

Classical conceptions. Aristotle's famous and influential classification of the "forms of government" was intended by him as a classification of constitutions, both good and bad. Under good constitutions—monarchy, aristocracy, and the mixed kind to which Aristotle applied the same term *politeia*—one man, a few men, or the many rule in the interest of the whole *polis*. Under the bad constitutions—tyranny, oligarchy, and democracy—the tyrant, the rich oligarchs, or the poor *dēmos*, or people, rule in their own interest alone.

Aristotle regarded the mixed constitution as the best practicable arrangement of offices in the *polis*. Such a *politeia*, or mixed constitution, would contain monarchic, aristocratic, and democratic elements. Its citizens, after they had learned to obey, were to be given opportunities to participate in ruling. This was a privilege only of citizens, however, since neither noncitizens nor slaves would have been admitted to any political office by Aristotle or by his contemporaries in the Greek city-states. Aristotle regarded some human beings as natural slaves, a point on which later Roman philosophers, especially the Stoics and jurists, disagreed with him. Although slavery was at least as widespread in Rome as in Greece, Roman law generally recognized a basic equality among all human beings. This was because, the Stoics argued, all human beings are endowed by nature with a spark of reason by means of which they can perceive a universal natural law that governs all the world and can bring their behaviour into harmony with it.

Influence of Roman law

Roman law thus added to Aristotelian notions of constitutionalism the concepts of a generalized equality, a universal regularity, and a hierarchy of types of laws. Aristotle had already drawn a distinction between the constitution (*politeia*), the laws (*nomoi*), and something more ephemeral that corresponds to what could be described as day-to-day policies (*psēphismata*). The latter might be based upon the votes cast by the citizens in their assembly and be subject to frequent changes, but *nomoi*, or laws, were meant to last longer. The Romans conceived of the all-encompassing rational law of nature as the eternal framework to which constitutions, laws, and policies *ought* to conform. The law of nature was the constitution of the universe.

Influence of the church. Christianity endowed this universal constitution with a clearly monarchical cast. The Christian God, it came to be argued, is the sole ruler of the universe. Unlike his Old Testament precursor, he is not arbitrary. His laws are passed down to men and are to be obeyed. Christian men were under an obligation to try to constitute their earthly cities on the model of the City of God, though in their ignorance and sinfulness they would never be able to succeed.

Both the church and the secular authorities with whom the church came into conflict in the course of the Middle Ages needed clearly defined arrangements of offices, functions, and jurisdictions. Medieval constitutions, whether of church or state, were considered legitimate because they were believed to be ordained of God or tradition or both. Confirmation by officers of the Christian Church was regarded as a prerequisite of the legitimacy of secular rulers. Coronation ceremonies were incomplete without a bishop's participation. The Holy Roman emperor travelled to Rome in order to receive his crown from the pope. Oaths, including the coronation oaths of rulers, could be sworn only in the presence of the clergy because oaths constituted promises to God and invoked divine punishment for violations. Even in the event of an obvious imposition of a new constitutional order, novelty could always be legitimized by reference to an alleged return to a more or less fictitious "ancient constitution." It was only in Italy during the Renaissance and in England after the Reformation that the "great modern fallacy" (as the Swiss historian Jacob Burckhardt called it) established itself, according to which men could rationally, deliberately, give themselves a new constitution to meet their current needs.

Today there is no state that is a member of the "family of nations"—whether Western or Eastern, "democratic" or "communist," advanced or traditionalist, developed or backward, monarchical, dictatorial, civilian-ruled, or militarist—that does not claim to have a constitution.

The social contract. The theoretical foundations of modern constitutionalism were laid down in the great works on the social contract, especially those of the English philosophers Thomas Hobbes and John Locke in the 17th century and the French philosopher Jean-Jacques Rousseau in the 18th. They were writing in the aftermath of the Protestant Reformation, which had dis-

turbed the fundamental constitution of Western Christendom.

Impact of the Reformation

As a result of the Reformation, the basis of divinely sanctioned contractual relations was broken up. Catholic bishops encouraged subjects of Protestant princes to break their pledges of fealty in order that they might keep faith with God and his Catholic Church. The Holy Roman Empire was torn apart by the wars of the Reformation. Henry VIII made the Church of England independent of Rome. In these circumstances, it became necessary to search for a new basis of order and stability, loyalty and obedience. In their search, political theorists —and especially the Protestants among them—turned to the old biblical concept of a covenant or contract, such as the one between God and Abraham and the Israelites of the Old Testament. The assumption was that God had constituted the political unit by choosing his partners in an eternal covenant.

In a sense, the secular theorists of the social contract almost reversed the process of choice. Instead of God choosing his people, a people through its representatives was now looked upon as choosing its governors, or its mode of governance, under God, by means of a social contract or constitution. According to modern theories of the social contract, the political unit is nevertheless established as in the biblical model by means of a promise or promises. Thomas Hobbes's state, or "Leviathan," comes into being when its individual members renounce their powers to execute the laws of nature, each for himself, and promise to turn these powers over to the sovereign—which is created as a result of this act—and to obey thenceforth the laws made by this sovereign. These laws enjoy authority because individual members of society are in effect their co-authors. According to Locke, individuals promise to agree to accept the judgments of a common judge (the legislature) when they accede to the compact that establishes civil society. After this (in one interpretation of Locke's *Second Treatise on Civil Government*), another set of promises is made—between the members of the civil society, on the one hand, and the government, on the other. The government promises to execute its trust faithfully, leaving to the people the right to rebel in case the government breaks the terms of the contract, or, in other words, violates the constitution. Subsequent generations accept the terms of the compact by accepting the inheritance of private property that is created and protected by the compact. Anyone who rejects the constitution must leave the territory of the political unit and go *in vacuis locis*, or "empty places"—America, in Locke's time. In his *Letters on Toleration*, Locke characteristically excluded atheists from religious toleration because they could be expected either not to take the original contractual oath or not to be bound by the divine sanctions invoked for its violation.

For Rousseau, too, the willingness to subject oneself to the "general will" to which only the popular sovereign can give expression is the essential ingredient of the social contract. In taking this position, Rousseau may have been influenced by the experience of his native Geneva. The Swiss Confederation is still referred to officially, in German, as an *Eidgenossenschaft*, a term best translated as "fellowship of the oath."

Hobbes's view on sovereignty

Hobbes's main contribution to constitutionalism lies in his radical rationalism. Individuals, according to Hobbes, come together out of the state of nature, which is a state of disorder and war, because their reason tells them that they can best ensure their self-preservation by giving all power to a sovereign. The sovereign may consist of a single person, an assembly, or the whole body of citizens; but regardless of its form, all the powers of sovereignty have to be combined and concentrated in it. Hobbes held that any division of these powers destroyed the sovereign and thereby returned the members of the commonwealth to the state of nature, in which the condition of man is " . . . solitary, poore, nasty, brutish, and short." Hobbes therefore preferred the singular sovereign since he was less likely than an assembly or than the whole body of citizens to become internally or functionally divided. Powers over war and peace, taxation, and the judiciary, along with all other governmental powers, should be concentrated in the sovereign. The individual should retain only his natural rights, which he cannot surrender into the common pool of sovereign powers. These rights include the right against self-incrimination, the right to purchase a substitute for compulsory military service, and the right to act freely in instances in which the laws are silent.

Locke attempted to provide firm assurance of the individual's natural rights, partly by assigning separate though coordinated powers to the monarch and Parliament and partly by reserving the right of revolution against a government that had become unconstitutionally oppressive. Locke does not use the word sovereignty. In this as in other respects, he remained within the English constitutional tradition, which had eschewed the concentration of all powers in a single organ of government. The closest that English constitutionalists came to identifying the centre of sovereign power was in the phrase, used frequently from the 16th century onward, the king (or queen) in Parliament.

Rousseau's theory of the general will

Whereas Hobbes created his unitary sovereign through the mechanism of individual and unilateral promises and whereas Locke prevented excessive concentration of power by requiring the cooperation of different organs of government for the accomplishment of different purposes, Rousseau merged all individual citizens into an all-powerful sovereign whose main purpose was the expression of the general will. By definition, the general will can never be wrong; for when something contrary to the general interest is expressed, it is defined as the mere "will of all" and cannot have emanated from the sovereign. In order to guarantee the legitimacy of government and laws, Rousseau would have enforced universal participation in order to "force men to be free," as he paradoxically phrased it. In common with Hobbes and Locke, Rousseau required the assent of all to the original social contract. He required smaller majorities for the adoption of laws of lesser importance than the constitution itself. His main concern was to provide for legitimacy through universal participation in legislation, whereas Locke and Hobbes were more concerned to provide constitutional stability through consent. As a result, Rousseau's thought appears to be more democratic than that of his English predecessors. He has even been accused of laying the philosophical foundations of "totalitarian democracy," for the state he describes in *The Social Contract* would be subject, at the dictates of its universal and unanimous sovereign, to sudden changes, or even transformations, of its constitution.

Influence of social contract theory

In the political thought of Hobbes, Locke, and Rousseau may be found theoretical consideration of the practical issues that were to confront the authors of the American and French constitutions. The influence of theories of the social contract, especially as they relate to the issues of natural rights, and the proper functions of government, pervades the constitution making of the revolutionary era that began with the U.S. War of Independence and is indeed enshrined in the great political manifestos of the time, the American Declaration of Independence and Bill of Rights, and the French Declaration of the Rights of Man and the Citizen.

The constitutional experience of these two countries, and, of course, of England, had great influence on liberal thought in Europe and other parts of the world during the 19th century and found expression in the constitutions that were demanded of the European monarchies. The extent to which the ideal of constitutional democracy has become entwined with the practice of constitutional government will be apparent from the examination in the following section of the main features of constitutional government.

FEATURES OF CONSTITUTIONAL GOVERNMENT

Virtually all contemporary governments have constitutions, but possession and publication of a constitution does not make a government constitutional. Constitu-

tional government is in fact comprised of the following elements.

Procedural stability. Certain fundamental procedures must not be subject to frequent or arbitrary change. Citizens must know the basic rules according to which politics are conducted. An action that is considered legal and constitutional today must not be condemned as illegal tomorrow. Stable procedures of government and politics provide citizens with adequate foreknowledge of the probable consequences of their actions. By contrast, under many nonconstitutional regimes, such as Hitler's in Germany and Stalin's in the Soviet Union, individuals, including high government officials, could never know from one day to the next whether the whim of the dictator's will would not turn today's hero into tomorrow's public enemy. The people did not even know the procedures by means of which the dictator's will would be made known to the general public.

Accountability. Under constitutional government, those who govern are regularly accountable to at least a portion of the governed. In a constitutional democracy, this accountability is owed to the electorate by all persons in government. Accountability can be enforced through a great variety of regular procedures, including elections, systems of promotion and discipline, fiscal accounting, recall, and referendum. In constitutional democracies, the accountability of government officials *to* the citizenry makes possible the citizens' responsibility *for* the acts of government. The most obvious example of this two-directional flow of responsibility and accountability is the electoral process. A member of the legislature or the head of government is elected by adult citizens and is thereby invested with authority and power in order that he may try to achieve those goals to which he committed himself in his program. At the end of his term of office, the electorate has the opportunity to judge his performance and to re-elect him or dismiss him from office. The official has thus rendered his account and has been held accountable.

Representation. Those in office must conduct themselves as the representatives of their constituents. To represent means to be present on behalf of someone else who is absent. Elections, of course, are not the only means of securing representation or of ensuring the representativeness of a government. Hereditary medieval kings considered themselves, and were generally considered by their subjects, to be representatives of their societies. Hobbes termed his sovereign *the* representative of the commonwealth. Of the social contract theorists, only Rousseau denied the feasibility of representation for purposes of legislation. The elected status of officeholders is sometimes considered no guarantee that they will be "existentially representative" of their constituents, unless they share with the latter certain other vital characteristics such as race, religion, sex, or age. The problems of representation are in fact more closely related to democratic than to constitutionalist criteria of government: a regime that would be considered quite unrepresentative by modern standards could still be regarded as constitutional so long as it provided procedural stability and the accountability of office holders to some but not all of the governed and so long as the governors were representative of the best or the most important elements in the body politic.

Division of power. Constitutional government requires a division of power among several organs of the body politic. Preconstitutionalist governments, such as the absolute monarchies of Europe in the 18th century, frequently concentrated all power in the hands of a single person. All powers of law making, enforcement, and interpretation were concentrated in the ruler himself so that the individual subject in conflict with his government faced the same single power or its agents no matter which way he turned. The same has been true in modern dictatorships such as Hitler's in Germany. Constitutionalism, on the other hand, by dividing power—between, for example, local and central government and between the legislature, executive, and judiciary—ensures the presence of restraints and "checks and balances" in the political system. Citizens are thus able to influence policy by resort to any of several branches of government.

Openness and disclosure. Democracy rests upon popular participation in government, constitutionalism upon disclosure of and openness about the affairs of government. In this sense, constitutionalism is a prerequisite (and has, in fact, been the historical precursor) of successful democracy since the people cannot participate rationally in government unless they are adequately informed of its workings. Originally, because they were concerned with secrets of state, bureaucracies surrounded their activities with a dark veil of secrecy. The ruler himself always retained full access to administrative secrets and often to the private affairs of his subjects, into which bureaucrats such as tax collectors and the police could legally pry. But when both administrators and rulers were subjected to constitutional restraints, it became necessary that they disclose the content of their official activities to the public to which they owed accountability. This explains the provision contained in most constitutions obliging the legislature to publish a record of its debates.

As a result of the modern proliferation of bureaucracy, the Scandinavian countries established the office of ombudsman. The ombudsman can protect the constitutional rights of citizens against bureaucratic violations by forcing administrative offices to open to him—and through him to the public—the records of otherwise secret bureaucratic files and proceedings. The office of ombudsman has recently been copied by many constitutional governments. This office and the general constitutional requirements for openness and disclosure of official records contrast sharply with the secrecy surrounding the much more pervasive activities of those closed societies that are usually labelled totalitarian.

Constitutionality. Written constitutions normally provide the standard by which the legitimacy of governmental actions is judged. In the United States, the practice of the judicial review of congressional legislation for its constitutionality—that is, for its conformity with the U.S. Constitution—though not explicitly provided for by the Constitution, developed in the early years of the republic. More recently, other written constitutions, including the Basic Law of the Federal Republic of Germany and Italy's republican constitution, provided explicitly for judicial review of the constitutionality of parliamentary legislation. This does not necessarily mean that a constitution is regarded as being prior and superior to all law. Although several European countries, including France, Italy, and West Germany, adopted new constitutions after World War II, they kept in force their codes of civil law, which had been legislated in the 19th century; and the U.S. Constitution guarantees citizens certain substantive and procedural rights to which they deemed themselves entitled as subjects of the British crown under the ancient English common law. Despite the greater antiquity of law codes, however, portions of them have been revised from time to time in order to eliminate conflicts between the law and certain constitutional norms that are regarded as superior. Parts of German family law and of the criminal code, for example, were revised in order to bring them into conformity with the constitutional provisions regarding the equality of persons irrespective of sex and with the individual's constitutionally guaranteed right to the free development of his personality.

Constitutional interpretation

Conflicting interests or parties are, of course, likely to place different interpretations on particular provisions of a constitution, and means, therefore, have to be provided for the resolution of such conflicts. The constitution itself may establish an institution the task of which is to interpret and clarify the terms of that constitution. In the American system, the Supreme Court is generally regarded as the authoritative interpreter of the Constitution and, therefore, of the intentions of the founders and of later contributors to the total constitutional fabric. But the Supreme Court cannot be regarded as the "final" interpreter or arbiter of the meaning of the Constitution

for a number of reasons. The court can always reverse itself, as it has done. The president can gradually change the interpretative outlook of the court through the nomination of new justices, and the Congress can exert a more negative influence by refusing to confirm presidential nominations of justices.

Provision was made in the constitution of the Fifth French Republic for the interpretation of certain constitutional matters by a Constitutional Council. Soon after the French electorate, in a referendum in 1958, had voted to accept the Constitution, a controversy erupted in France over the question of whether the president of the republic could submit to popular referendum issues not involving constitutional amendments but on which parliament had taken a position at odds with the president's. The Constitution itself seemed to provide that the Constitutional Council could rule definitively on this question, but Pres. Charles de Gaulle chose to ignore its ruling, which was unfavourable to himself. As a result, the Constitutional Council lost authority as the final interpreter of the meaning of the Constitution of the Fifth Republic.

It may thus be seen that because of the inherent difficulties in assessing the intentions of the authors of a constitution and because of the possibility that the executive or legislative branch of government may be able to ignore, override, or influence its findings, it is difficult to ensure constitutional government merely by setting up an institution whose purpose is constitutional interpretation.

Constitutional change. Written constitutions are not only likely to give rise to greater problems of interpretation than unwritten ones; they are also harder to change. Unwritten constitutions tend to change gradually, continually, and often imperceptibly, in response to changing needs. But when a constitution lays down exact procedures for the election of the president, for relations between the executive and legislative branches, or for defining whether a particular governmental function is to be performed by the federal government or a member state, then the only *constitutional* way to change these procedures is by means of the procedure provided by the constitution itself for its own amendment. Any attempt to effect change by means of judicial review or interpretation is unconstitutional, unless, of course, the constitution provides that a body (such as the U.S. Supreme Court) may change, rather than interpret, the constitution.

Many constitutional documents make no clear distinction between that which is to be regarded as constitutional, fundamental, and organic, on the one hand, and that which is merely legislative, circumstantial, and more or less transitory, on the other. The constitution of the German Weimar Republic could be amended by as little as four-ninths of the membership of the Reichstag, without any requirement for subsequent ratification by the states, by constitutional conventions, or by referendum. Although Hitler never explicitly abrogated the Weimar Constitution, he was able to replace the procedural and institutional stability that it had sought to establish with a condition of almost total procedural and institutional flux.

A similar situation prevailed in the Soviet Union under the rule of Stalin. But Stalin took great trouble and some pride in having a constitution bearing his name adopted in 1936. Despite occasional efforts to replace the Stalin constitution with a more up-to-date document, it continues, together with the Rules of the Communist Party of the Soviet Union, to serve as the formal framework of government. The procedures established by both of these documents, however, have not been able to provide Soviet citizens and politicians with reliable foreknowledge of the rules of the political process from one year to the next or with guidance as to which institutions and practices they were to consider fundamental or virtually sacrosanct and which they could safely afford to criticize. As a result, changes in the personnel and policies of the Soviet Union and of similar Communist regimes have rarely been brought about smoothly and have frequently required the use of violence.

Constitutional stability. If one distinguishes between stability and stagnation on the one hand and between flexibility and flux on the other, then one can consider those constitutional systems most successful that combine procedural stability with substantive flexibility—that is, that preserve the same general rules of political procedure from one generation to the next while at the same time facilitating adaptation to changing circumstances of internal and external economic, social, cultural, and power relationships. By reference to such criteria, those written constitutions have achieved the greatest success that are comparatively short; that confine themselves in the main to matters of procedure (including their own amendment) rather than matters of substance; that, to the extent that they contain substantive provisions at all, keep these rather vague and generalized; and that contain procedures that are congruent with popular political experience and know-how. These general characteristics appear to be more important in making for stability than such particular arrangements as the relations between various organs and levels of government or the powers, functions, and terms of tenure of different officers of state.

Characteristics of stable constitutions

There is little evidence to support the thesis that a high level of citizen participation necessarily contributes to the stability of constitutional government. On the contrary, the English political economist Walter Bagehot, who in 1867 wrote a classic analysis of the English constitution (*The English Constitution*), stressed the "deferential" character of the English people, who were quite happy to leave government in the hands of the governing class.

Much more important than formal citizen behaviour, such as electoral participation, are informal attitudes and practices and the extent to which they are congruent with the formal prescriptions and proscriptions of the constitution itself. Constitutional government cannot survive effectively in situations in which the constitution prescribes a pattern of behaviour or of conducting affairs that is alien to the customs and way of thinking of the people. Whereas in many developing countries in the decades after World War II a new and alien kind of constitutional democracy is imposed or adopted, a gap may soon develop between constitutionally prescribed and actual governmental practice. This in turn renders the government susceptible to attack by opposition groups. Such attack is especially easy to mount in situations in which a constitution has a heavy and detailed substantive content, when, for example, it guarantees the right to gainful employment or the right to a university education for all qualified candidates. In the event of the government being unable to fulfill its commitment, the opposition is able to call the constitution a mere scrap of paper and to demand its improvement or even its complete replacement. Such tactics often have succeeded, but they ignore the dual strategic function of the constitution. It is meant not only to arrange the offices of the state, in Aristotle's sense, but also to state the goals toward which the authors and ratifiers of the constitution want their community to move.

THE PRACTICE OF CONSTITUTIONAL GOVERNMENT

Britain. It is accepted constitutional theory that Parliament (the House of Commons and the House of Lords acting with the assent of the monarch) can do anything it wants to, including abolish itself. The interesting aspect of British government is that, despite the absence of restraints such as judicial review, acts that would be considered unconstitutional in the presence of a written constitution are attempted very rarely, certainly less often than in the United States.

The English constitution and the English common law grew up together, very gradually, more as the result of the accretion of custom than through deliberate, rational legislation by some "sovereign" lawgiver. Parliament grew out of the Curia Regis, the King's Council, in which

the monarch originally consulted with the great magnates of the realm and later with commoners who represented the boroughs and the shires. Parliament was, and is, a place in which to debate specific issues of disagreement between, initially, the crown, on the one hand, and the Lords and Commons, on the other. The conflicts were settled in Parliament so that its original main function was that of a court—it was in fact known as "the High Court of Parliament" as late as the 16th century.

The locus of power in the English constitution shifted gradually as a result of changes in the groups whose consent the government required in order to be effective. In feudal times, the consent of the great landowning noblemen was needed. Later, the cooperation of those commoners who were willing to grant revenue to the crown —that is, to pay taxes—was sought. The crown itself, meanwhile, was increasingly institutionalized, and the distinction was drawn ever more clearly between the private and public capacities of the king. During the course of the 18th century, effective government passed more and more into the hands of the king's first minister and his cabinet, all of them members of one of the two houses of Parliament. Before this development, the king's ministers depended upon their royal master's confidence to continue in office. Henceforward they depended upon the confidence of the House of Lords and especially the House of Commons, which had to vote the money without which the king's government could not be carried on. In this way the parlay that was originally between the monarch and the houses of Parliament was now struck between the ministry and its supporters, on the one hand, and opposing members of Parliament ("His Majesty's Loyal Opposition"), on the other. Parliamentary factions were slowly consolidated into parliamentary parties, and these parties reached out for support into the population at large by means of the franchise, which was repeatedly enlarged in the course of the 19th century and eventually extended to women and then to 18-year-olds in the 20th.

When a prime minister loses a vote of confidence in the House of Commons, he can either resign to let the leader of the Opposition form a new government or ask the monarch to dissolve Parliament and call for new elections. As a result of the strong party discipline that developed in the 20th century, prime ministers generally do not lose votes of confidence any more, and they call for new elections at the politically most favourable moment. According to an act of Parliament, elections must be held at least every five years—but another act of Parliament can change or suspend this apparently "constitutional" provision, as was done during World War II, when the life of the incumbent House of Commons was extended until the defeat of Germany. Similarly, relations between, and the relative powers of, the House of Lords and the House of Commons have been repeatedly redefined to the disadvantage of the House of Lords by acts of Parliament, to such an extent that the Lords retain only a weak suspensory veto. All such fundamental constitutional changes have occurred either informally and without any kind of legislation at all or as a result of the same legislative procedures employed to pass any other ordinary circumstantial bill.

United States. The U.S. Constitution is replete with phrases taken from the British constitutional vocabulary, such as the "advice and consent" required of the Senate for the ratification of treaties and the confirmation of presidential appointments, which echoes the advice and consent of the Lords and Commons to enactments of "the King's Most Excellent Majesty" in acts of Parliament. In several respects, the U.S. Constitution represents a codification of its authors' understanding of the English constitution, to which they added ingenious federalist inventions and the formal amending procedure itself. Despite the availability of this procedure, however, many if not most of the fundamental changes in American constitutional practice have not been effected by formal amendments. The Constitution still does not mention political parties or the president's cabinet. Nor was the Constitution changed in order to bring about or to sanction the fundamentally altered relations between the executive and the Congress, between the Senate and the House, and between the judiciary, the legislature, and the executive. The presence of a constitutional document, however, has made American politics more consciously "constitutionalist," at least in the sense that politicians in the United States take more frequent recourse than their British counterparts to legalistic argumentation and to actual constitutional litigation. The United States, moreover, is denied the kind of flexibility illustrated by the postponement of British parliamentary elections during World War II since the Constitution explicitly provides the dates for congressional and presidential elections. It is one of the remarkable facts of American constitutional history that the constitutional timetable for elections has always been observed, even during external war and the Civil War of the 19th century.

Europe. France, Germany, and Italy, as well as most non-European countries influenced by continental concepts of constitutionalism, have no record of unbroken constitutional fidelity similar to that found in Britain and the U.S.

Constitutional government in France

Because of the highly substantive and ideological content of most French constitutions, the best way to change them has been to replace them altogether with a new, ideologically different document. Only the constitution of the Third Republic (established in 1870) was exceptional in this respect, since it consisted of very short, highly procedural organic laws, which served France well for 70 years, until the German invasion of 1940.

The main political problem attributed to the constitution of the Third Republic was the instability of cabinets. The negative majorities that voted "no confidence" in a cabinet usually could not stay together for the positive purpose of confirming a new cabinet. The constitution of the Fourth Republic (1946–1958) made the overthrow of governments by the National Assembly more difficult. In fact, however, the life of the average cabinet in the Fourth Republic was even shorter than in the Third, and French government became virtually paralyzed when it had to deal with the problems raised by the Algerian independence movement. To avert a military takeover, General de Gaulle was given wide discretion in 1958 in the formulation of a new constitution, which was overwhelmingly accepted in a referendum. The constitution of the Fifth French Republic gives the president of the Republic the power to dissolve Parliament and the means of circumventing a hostile National Assembly through the referendum. Since 1958, French cabinets have been very stable indeed, and the constitution proved resilient during the "revolution of 1968."

Germany, which was unified as a national state only in 1871, established its first democratic constitution in 1919, after its defeat in World War I. Although some of the greatest German jurists and social scientists of the time participated in writing the Weimar Constitution, it has been adjudged a failure, largely because it was followed by Hitler's brutal dictatorship. Political parties became highly fragmented, a phenomenon that was explained partly by an extremely democratic electoral law (not a part of the constitution) providing for proportional representation. Some of the parties of the right, such as Hitler's Nazis, and of the left, such as the Communists, were opposed to the constitutional order and used violence in their efforts to overthrow the Republic. To deal with these threats, the President used his constitutional emergency powers under which he could suspend civil rights in member states of the federal system. Several chancellors (the German equivalent of a prime minister) stayed in office after the President had dissolved a Parliament in which the chancellor lacked a supporting majority. They continued to govern with the help of presidential emergency powers and by legislating on the basis of powers previously delegated to them by Parliament.

When a new constitution was being drafted for the Western zones of occupation after World War II, every effort was made to correct these constitutional errors, to which, in retrospect, the failure of the Weimar Republic

was attributed. Under the Basic Law of Bonn, Parliament cannot delegate its legislative function to the chancellor, and civil rights cannot be suspended without continuous parliamentary surveillance. The president has been turned into a figurehead on the model of the French presidents of the Third and Fourth Republics, and Parliament cannot overthrow a chancellor and his cabinet unless it first elects a successor with the vote of a majority of its members. Negative majorities cannot paralyze government unless they can agree on alternative policies and personnel. The extreme form of proportional representation used before Hitler came to power was replaced by a mixed electoral system under which half the members of the Bundestag (the lower house of the legislature) are elected from party lists by proportional representation, while the other half are elected in single member constituencies. In order to benefit from proportional representation, a party must obtain at least 5 percent of the votes cast in the Federal Republic. As a result, the number of parties steadily contracted during the first two decades of the Federal Republic and extremist parties were kept out of Parliament. Cabinets have been very stable, and each Parliament has served out its full four-year term. The provision for the "constructive vote of no confidence" has not been invoked.

Latin America, Africa, and Asia. The experience of constitutional government in continental Europe exerted great influence on the newly independent former colonies of Europe in the Middle East, Asia, and Africa. In the early years of their independence from Spain, most Latin American countries adopted constitutions similar to that of the United States. But since they lacked the background that had produced the American Constitution, including the English common law, most of their efforts at constitutional engineering were unsuccessful. In Asia and Africa and in the Caribbean, former colonies of Great Britain, such as India, Nigeria, Zambia, and Jamaica, have generally been comparatively more successful in the operation of constitutional government than former colonies of the continental European countries (*e.g.*, Indonesia, the Congo, and Haiti). The British usually left a modified and simplified version of their own constitution upon granting independence to their former subjects, some of whom they had previously trained in the complicated operating procedures of the British constitution. British parliamentary procedure proved sufficiently adaptable to remain in use for some time after the departure of the British themselves. France's former colonies in Africa, because they achieved independence after the founding of the Fifth Republic, modelled their new constitutions upon General de Gaulle's, partly because this enhanced the power of the leaders under whom independence had been achieved.

BIBLIOGRAPHY. The most comprehensive contemporary book on this subject is C.J. FRIEDRICH, *Constitutional Government and Democracy: Theory and Practice in Europe and America*, 4th ed. (1968). See also H.J. SPIRO, *Government by Constitution: The Political Systems of Democracy* (1959). The best study of constitutional change is W.S. LIVINGSTON, *Federalism and Constitutional Change* (1956). For a compendium of current constitutions, see A.J. PEASLEE (ed.), *Constitutions of Nations*, 3 vol. (1956). Aristotle's classic *Politics* is available in translation by SIR ERNEST BARKER (1962). Sir Ernest Barker has also edited *Social Contract: Essays by Locke, Hume, Rousseau* (1947). The best history of the origins of English constitutionalism is C.H. MCILWAIN, *The High Court of Parliament and Its Supremacy: An Historical Essay on the Boundaries Between Legislation and Adjudication in England* (1910). See also F.D. WORMUTH, *The Origins of Modern Constitutionalism* (1949). On the related problems of accountability and responsibility, see H.J. SPIRO, *Responsibility in Government: Theory and Practice* (1969). For an understanding of the origins of American constitutionalism, *The Federalist* by ALEXANDER HAMILTON, JAMES MADISON, and JOHN JAY is indispensable. The classic exposition of *The English Constitution* was provided by WALTER BAGEHOT (1867); for a recent analysis, see S.H. BEER and A.B. ULAM (eds.), *Patterns of Government: The Major Political Systems of Europe* (1962). On French constitutionalism, see also STANLEY HOFFMANN *et al.*, *In Search of France: The Economy, Society, and Political System in the Twentieth Century* (1963). RALF DAHRENDORF, *Society and Democracy in Germany* (1967), is a sociological account, complemented by A.J. HEIDENHEIMER, *The Governments of Germany*, 2nd ed. (1966). For excerpts of constitutional documents of the major European governments, see J.J. WUEST and M.C. VERNON, *New Source Book in Major European Governments* (1966). The concept of congruence between political and social patterns of authority is elaborated by H.H. ECKSTEIN, *Division and Cohesion in Democracy: A Study of Norway* (1966). For studies of constitutional development in newly independent countries, see H.J. SPIRO (ed.), *Patterns of African Development: Five Comparisons* (1967).

(H.J.Sp.)

Consumer Credit

Consumer credit consists of short- and intermediate-term loans used to finance the purchase of commodities or services for personal consumption or to refinance debts incurred for such purposes. Consumer loans facilitate the purchase of automobiles, refrigerators, and other durable goods by enabling buyers to draw upon future income. The loans may be supplied by lenders in the form of cash loans or by sellers in the form of sales credit.

The volume of consumer credit available and used in industrialized countries has grown rapidly as more and more people have come to receive regular income in the form of dependably fixed wages and salaries and as mass markets for durable consumer goods have become feasible. Its first growth was in the United States after World War I, but it subsequently developed on a large scale in other industrialized countries, particularly in Great Britain and western Europe. The Soviet Union and other Communist countries began to introduce retail installment sales in the 1960s.

TYPES OF CONSUMER CREDIT

Installment credit. Consumer loans may be classified into two broad categories: installment loans, to be repaid in two or more payments, and noninstallment loans, to be repaid in a lump sum. Installment loans, which are by far the most important, include (1) automobile loans, (2) loans for other consumer goods, (3) home repair and modernization loans, and (4) personal loans. These are extended by many lenders, including commercial banks, sales finance and consumer finance companies, credit unions, savings and loan associations, mutual savings banks, pawn brokers, and other financial intermediaries; installment loans in the form of sales credit are supplied by department stores and other retail establishments.

Types of lenders

Automobile loans. Installment loans to finance the purchase of automobiles constitute the largest single segment of consumer credit. Approximately two-thirds of the new cars purchased in the United States between 1955 and 1970 were financed in this way. The terms of an automobile installment loan—the amount of the down payment, the term until maturity, and the interest rate—determine the size of the monthly payment. The maturity on automobile loans for new cars has generally become standardized at 36 months, the down payment at approximately 20 percent, and the finance charge at a rate between 9 and 12 percent. When the terms on automobile loans are tightened, sales of automobiles decline because consumers become less willing to borrow.

Other consumer goods loans. Installment loans are also used for the purchase of household appliances, boats, jewelry, furnishings, and, in recent years, mobile homes. They may be either cash loans or sales credit. This category also includes various types of revolving-credit and bank credit-card plans; these establish explicit credit limits for consumers that may be drawn upon at any time and are not necessarily restricted to the financing of particular durables.

Repair and modernization loans. Installment loans to finance home repairs and improvements usually run for longer terms than do other types of installment credit. These loans typically have maturities between five and seven years.

Personal loans. A wide variety of unsecured loans may be extended on a personal basis rather than for the purchase of durable goods. They may be used to pay taxes, consolidate previous debt, or to finance education and travel expenditures, auto repairs, and medical or funeral expenses.

Noninstallment credit. Some consumer loans are designed to be repaid in a lump sum. The most common of these are single-payment loans by financial institutions, charge accounts of retail stores, and the service credit extended by doctors, hospitals, and utility companies.

THE HISTORICAL DEVELOPMENT OF CONSUMER CREDIT IN INDUSTRIALIZED COUNTRIES

The tremendous growth in consumer lending after World War I was of fundamental importance in the development of mass markets for consumer durables and luxury goods. In many countries it required a change in the traditional attitudes toward moneylending. In medieval Europe, consumer loans were regarded as nonproductive, and charging interest on such loans was considered usury. Laws established maximum rates of interest and regulated the conditions under which loans could be made. The rise of capitalism brought about a rethinking, led by such men as the French economist A.R.J. Turgot and the British philosopher Jeremy Bentham, who attacked the basic idea of legal control of interest rates. By the latter part of the 19th century their teachings had led to the abandonment of all control over interest rates in England and in most of the countries of Europe.

The interest rate question. The repeal of the statutory maximum interest rate in England left unsolved the question whether a particular rate was to be considered fair or unfair. The courts groped their way through innumerable cases of abuse and extortion, particularly in personal loans, to a method of providing relief when the transaction was harsh or unconscionable. A decision in 1880 established the power of chancery to reopen any interest transaction to relieve a borrower of excessive interest and in other ways to soften the terms of an offensive contract. This power was subsequently incorporated in English statutes. Protection for installment purchasers was provided by the Hire Purchase Act of 1938 and later. In 1958 British banks began to offer their depositors personal loans for any reasonable purpose without security, in amounts from 50 to 500 pounds, at an effective rate of about 10 percent. West German banks began to do likewise in 1959 with loans of DM. 300 to DM. 2,000.

Usury laws

In the United States the so-called usury laws setting maximum interest rates remained on the books until well into the 20th century in most states. The maximum rate was often as low as 6 percent, with the result that established financial institutions could not earn a reasonable profit on consumer loans, and therefore left the field to illegal lenders, so-called loan sharks, who charged extremely high rates of interest and used strong-arm collection techniques. These abuses stirred various efforts to find other ways of meeting the legitimate needs of lower income borrowers.

One method followed by the Russell Sage Foundation, in its pioneering draft of a model small-loan law, was to recognize the purveyor of credit as a responsible businessman who must make a reasonable profit. This has formed the basis of much state legislation in the U.S. Such legislation requires verification of the lender's fitness and good character, along with licensing, supervision, and annual reports; imposes severe penalties for infractions; sets a maximum size for loans; and permits a rate of 2 to 3.5 percent a month on the unpaid principal balance on the smallest loans, with graduated lower rates on larger loans. Such lenders are called loan companies, licensed lenders, or consumer finance companies.

The loan and investment approach known in the United States as the Morris Plan, founded in 1910 by Arthur J. Morris, follows the traditional bank method for business loans. The borrower contracts for a term loan at the statutory rate and agrees to make periodic deposits or buy an investment certificate on the installment plan. Deposits or the paid-up certificate are sufficient to liquidate the loan at maturity. The economic effect of the dual transaction is that the initial loan is approximately twice the average loan outstanding, and the finance charge is double the contract rate.

Another way was to amend the state laws to permit the formation of credit unions by groups of employees, members of fraternal organizations, etc. The first credit union in the United States was organized in a parish of cotton mill workers in Manchester, N.H., in 1909. Credit-union members contribute to a common pool from which loans are made to those who need them. Overhead is kept low because space, time, and effort are contributed, permitting rates of about 1 percent a month.

As the production of automobiles and other household appliances grew, retailers often allowed customers the privilege of extended payments in return for a higher price, and out of this practice developed a legal notion called the time-price doctrine. Under this doctrine, courts in the United States have held that the difference between a cash price of $100 and a sales credit price of $120 ($10 a month for 12 months) is not interest, since no loan of money has taken place, but simply a time-price differential and not subject to the usury laws. This legal distinction has influenced the regulation of retail sales financing in many states.

An indirect type of financing has been developed by sales finance companies, which buy, at a discount from dealers, notes given by consumers in connection with time-payment purchases, which they are unable or unwilling to pay in a lump sum. Effective rates in the early 1960s ran from 12 to 24 percent. Starting in the late 1930s, commercial banks also became heavily engaged in consumer credit.

A consequence of these ad hoc statutes and piecemeal regulations in the United States has been to fragment the consumer credit market, which from an economic point of view is essentially a single market. In 1970 over two-thirds of the states had separate laws for retail financing, industrial loans, and installment loans; nearly all the states had special small loan and credit union laws. Many lenders were restricted by these laws to state-wide or even to local markets, to a particular kind of loan, or to a particular kind of transaction. Thus commercial banks and credit unions were not allowed to deal directly in second-mortgage real estate financing. Retailers could not make cash loans. Finance companies could not make revolving loans, although retailers could. Various proposals were made to remove some of the restrictions on the ground that they would produce a more efficient consumer-credit market.

Finance charges

The cost of credit. The finance charges on consumer loans generally run substantially higher than the interest costs of business loans. There are several reasons for this. Most consumer loans are for relatively small amounts, and the costs of lending do not decrease for small loans but, on the contrary, increase as a proportion of the total amount borrowed. These costs include credit investigation—necessarily more expensive in the case of a private household than of a business—and the expense of processing the loan. Actual charges range from 12 percent per annum to as high as 36 percent on small, unsecured cash loans. The cost of borrowing is usually disguised by the way in which the charges are quoted. A loan of $100 for a year may call for repayment in monthly installments totalling $106; this is often described as $100 plus a 6 percent "add-on rate." Or the loan may be of $100 minus a 6 percent "discount," with monthly repayments totalling $100. Or the interest may be charged at "1 percent per month" for a year. In each case, however, the actual charge to the borrower is 12 percent or more because he is assessed for interest on the full loan throughout the period of repayment.

In the case of the simple retail charge account, where there is no interest fee, the costs are generally borne by the seller or, in a sense, by the seller's customers in the form of higher prices.

In the United States in the late 1960s, there was much

support among consumers for "truth in lending" legislation that would require lenders to disclose the actual cost of borrowing. This led the U.S. Congress to pass the Truth-in-Lending Act, which took effect in July 1969. One of its more important provisions requires lenders to state the effective finance charge in terms of an annual percentage rate as a measure of the relative cost of credit in percentage terms so that borrowers may compare a finance charge stated in terms of dollars per month with, say, another finance charge stated as an annual add-on rate.

It has been argued that changes in down payments, loan-to-value ratios, and maturities have a bigger effect on monthly payments than do changes in the finance rates and that borrowers may therefore be more sensitive to these other credit terms than to finance rates. The disclosure requirements under the 1969 law may provide some evidence on this point.

Other proposals have sought to simplify the rules under which lenders operate in the consumer credit market. These would include a uniform set of regulations applying to all lenders and to all consumer credit transactions as well as measures to encourage the entry of new lenders into the market, to facilitate more effective competition among existing lenders, and to provide borrowers with legal remedies for the redress of some abuses. In Great Britain a committee headed by Lord Crowther recommended in March 1971 that the existing accumulation of laws relating to consumer credit be replaced by uniform legislation. It proposed, among other things, that government controls over the terms of credit be abandoned and that lenders be required to state the true annual rate of interest being charged.

THE DIMENSIONS OF CONSUMER CREDIT

Credit and the economy

Quantitative estimates of the relative importance of various kinds of consumer credit are available for the United States, where it plays a larger part than it does in most other countries. The total amount of credit outstanding grew from $21,500,000,000 in 1950 to $126,800,000,000 in 1970, or from 7.5 percent of the gross national product to 13.0 percent. The important category of installment loans grew from $14,700,000,000 in 1950 to $101,200,000,000 in 1970, as shown in the Table. The relative amount of credit supplied by retailers declined during the same period from about 20 percent to less than 13 percent, while the proportion supplied by commercial banks, credit unions, and other financial institutions rose to more than 85 percent. More than half of the households in the U.S. had installment loans outstanding in 1970, as compared with less than a quarter in 1950.

Major Categories of Consumer Credit in the United States, Selected Years
($000,000,000)

year	total consumer credit	installment credit		noninstallment credit
		total	automobile	
1950	21.5	14.7	6.1	6.8
1955	38.8	28.9	13.5	9.9
1960	56.0	42.8	17.7	13.2
1965	87.9	68.6	28.9	19.3
1970	126.8	101.2	35.5	25.6

Source: *Federal Reserve Bulletin*

In Great Britain the Crowther committee found that consumer credit extended in 1969 totalled £4,228,000,000. Almost half of this was in loans for house purchase and improvement, the bulk of it financed by building societies. Of the remainder, which was used to finance purchases of automobiles, clothes, household durables, and miscellaneous purchases, the chief suppliers were banks, finance houses, mail order houses, and retail stores. Comparisons among countries are difficult to make; according to one estimate, in 1967 the debt outstanding on passenger cars and other consumer goods was 4.2 percent of disposable income in Great Britain, 7.3 percent in Australia, 11.8 percent in Canada, and 9.6 percent in the United States (N. Runcie, *The Economics of Instalment Credit*, 1969, p. 94).

The immense growth in consumer credit has raised the question whether this mass of debt may have a destabilizing effect on the economy. If to consumer credit one adds mortgage debt, then total household debt in the U.S. in 1970 reached $450,300,000,000 (as compared with $75,500,000,000 in 1950), or about two-thirds of total disposable income. It is held that consumer credit tends to increase when business is on the upswing, until consumers become so committed that their purchasing power for other purposes is reduced. If this were true it would make the economic system more vulnerable to shocks and exert a downward pressure on economic activity. The fact that peaks and troughs in consumer credit accompany the peaks and troughs in business activity has been cited in support of this theory. But there is also some evidence suggesting that the supply of automobile loans may actually function as a stabilizer—by becoming more restrictive during peak levels of business activity and less so during recessions and the early stages of recovery. In general, changes in consumer credit do not appear as important an influence on economic activity as changes in inventory accumulation or investment spending by business.

Credit controls

Efforts to control economic activity often include controls on the supply of consumer credit. Selective controls over consumer credit have been introduced in wartime in the United States and Great Britain, to discourage expenditures on consumer durable goods. The British monetary authorities also administered selective consumer credit controls in the 1960s as a stabilization measure to help reduce inflationary pressures. The most common methods of control are to require a larger down payment, to set a maximum loan-to-value ratio, and to shorten the period of repayment. Selective controls are not popular, and their administration requires an extensive monitoring apparatus. More recent stabilization efforts have tended to rely on monetary measures to influence the cost and availability of credit. Changes in the supply of money have a significant impact on the supply, availability, and cost of consumer loans, although the effects seem to be felt even more strongly in other areas such as the home mortgage market.

BIBLIOGRAPHY. General surveys of consumer credit include: JOHN M. CHAPMAN and ROBERT P. SHAY (eds.), *The Consumer Finance Industry: Its Costs and Regulation* (1967); D.I. FAND, *Savings Intermediaries and Consumer Credit Markets* (1970); RALPH HARRIS, MARGOT NAYLOR, and ARTHUR SELDON, *Hire Purchase in a Free Society*, 3rd ed. (1961); Report of the Crowther Committee in Great Britain, *Consumer Credit* (1971); FRIEDRICH A. LUTZ, *Der Konsumentenkredit* (1954); and N. RUNCIE, *The Economics of Instalment Credit* (1969). For the development of legislation, see BARBARA A. CURRAN, *Trends in Consumer Credit Legislation* (1965); NATIONAL CONFERENCE OF COMMISSIONERS ON UNIFORM STATE LAWS, *The Uniform Consumer Credit Code* (1968); F.R. OLIVER, *The Control of Hire-Purchase* (1961); and the SENATE COMMITTEE ON BANKING AND CURRENCY, *Truth in Lending, 1967* (1967).

(D.I.F.)

Consumerism

Widespread interest and concern about consumer protection and the growth of organizations whose objectives are to disseminate information about products and to persuade industry to provide better quality goods and services—both parts of a phenomenon usually described as consumerism—were largely developments of the 1960s and 1970s. It is true, however, that it is possible to point to common examples of laws in ancient cultures whose purpose was to protect the purchaser. These included proscriptions against selling short-weight or adulterated goods, medieval concepts of a "just price," and laws against usury and the manipulation of markets. In some countries, legislation of the later 19th and early 20th centuries still serves as the basis of consumer protection, though such measures in the past were directed largely at marginal abuses or, in the case of more recent

legislation, at products, such as food and drugs, that may constitute a threat to public safety if certain minimum standards are not observed.

The consumer's position in common law

The fundamental relation between buyer and seller, however, was, and to some extent remains, that which is implied in the common-law maxim *caveat emptor* ("let the buyer beware"), according to which the onus is placed on the buyer to protect his own interests by ensuring that what he purchases is of sufficient quality to meet his needs. This doctrine, which has been progressively modified in the interests of the consumer to the extent that the law is governed more by the exceptions than by the rule itself, accorded well with the principle of laissez-faire capitalism that the buyer is served best by free competition between sellers. Certain changes that have occurred over the past century in the nature of industrial corporations and in the types of products available to the consumer have rendered this assumption invalid. First may be cited the growth of large corporations exerting monopoly influences—*i.e.*, setting prices and standards of quality for a particular product, so that the consumer is no longer afforded the opportunity to choose among effectively competing suppliers. Such corporations do not have to control a product's entire market, which may in fact be divided among several suppliers who offer the consumer products so similar in terms of price and quality that effective choice is impossible. Sometimes the impression of competition may be created by heavy advertising of "competing" brands or makes. Such a situation, known to economists as one of oligopoly, has often been typical of the markets for detergents and automobiles.

A second factor that has weakened the effectiveness of the consumer in protecting his own interests by rational choice among competing alternatives is the complexity of the products that modern technology has made available. There is, for example, little basis on which the consumer can decide between the merits of two different television sets, vacuum cleaners, or patent drugs without a mass of technical information, which he would probably be unable to evaluate even if it were given to him.

A third factor that is commonly alleged to have eroded the consumer's capacity to choose effectively in his own best interest is his increasing exposure to advertising and to new sales techniques. Advertising, it is argued, is predominantly persuasive rather than informative, serves to create demand for products that are unneeded, and often exploits the hidden fears, insecurities, and prejudices of the consumer; further, modern sales methods, ranging from so-called "introductory offers," "free gifts," and trading stamps to the "high-pressure" techniques practiced by door-to-door salesmen, are alleged to subject the consumer to an unjustified degree of influence.

It is against this background that a variety of pressures and methods for protecting consumer interests have developed.

MAIN AREAS OF REGULATION

As mentioned above, the common law has long provided certain safeguards to buyers, some of which have subsequently been incorporated into legislation. Of the safeguards derived from common law, the most important one—embodied in England's Sale of Goods Act (1893)—is the one that states that, whenever a buyer, expressly or by implication, makes known to the seller the particular purpose for which the goods are required, thus relying on the seller's skill or judgment, there is an implied condition that the goods sold shall be reasonably fit for such a purpose.

Early legislation concerning food and drugs

Other early legislation dealt mainly with adulteration of food and drugs. This was true, for example, of the Adulteration of Food and Drugs Act of 1872 (England) and similar, cumulative measures of 1848, 1890, and 1906 in the United States. The scope of such acts has been enlarged from time to time to include, for example, goods such as electrical products and automobiles, which could endanger the safety of the consumer if certain standards are not met. The provisions of such legislation are necessarily complex and vary from country to country, as well as, in the United States, from state to state. Various non-statutory controls, such as standards laid down by national-standards institutions (see below), also interact with statutory controls. In the following survey both the statutory and non-statutory aspects of this subject are considered together.

Controls on manufacturing and design. *Legislative controls.* Of all industries, food and drugs are the most controlled by legislation. Other products in general are controlled by standards institutions, which lay down basic minimum standards for many different kinds of products. Legislative controls applying to food and drug manufacturers prohibit them from adding or removing anything from the product they sell that would make it injurious to health. Although this might appear to afford absolute protection for the consumer, manufacturers sometimes unwittingly add ingredients that are subsequently found to be harmful—for example, cyclamates, which were used for some years as an artificial sweetener. The frequency of such occurrences will clearly depend on the rigour of the standards of the official testing agencies concerned and the stringency with which such standards are applied.

Standards institutions. For nonfood products, legislation is less easily devised and far less easily enforced. Most countries, nevertheless, have developed minimum applicable standards. National-standards institutions were, in many instances, set up more for the benefit of manufacturers than for that of the ordinary, domestic consumer. In addition, government bodies were often formed to better control government purchasing. In the United States, for example, the General Services Administration laid down specifications and quality standards that had to be satisfied before the federal government would buy supplies. Other standards bodies, such as the British Standards Institute, started in 1901, were set up for the convenience of manufacturers so that one manufacturer's goods could be used in conjunction with another's, as in the standardization of electrical fittings.

By the 1950s, standards organizations had become far more aware of the needs of the ordinary consumer, but their legal status, for the most part, remained unaltered. Most recommendations are devised with the cooperation of industry, government departments, and consumers. The standards themselves are not usually legally enforceable but remain voluntary. They usually do not reflect the quality of the product as a whole but deal only with a specific aspect of it. The mark of a standards institution, for example, may well indicate that a hair dryer is sufficiently insulated against electrical-shock hazards, but not that it dries hair satisfactorily.

Usually unenforceable voluntary standards

Though standards institutions in various countries may differ in some aspects, they are all basically the same. Some are entirely financed by the government (as in India); some are part of a government department (as in Japan); others are a mixture, as in the United States, where the General Services Administration is a federal agency and the American National Standards Institute (American Standards Association) is financed by manufacturers. For the world as a whole, the International Standards Organization draws up standards that can be adopted by all countries. Again, compliance with such standards is optional.

Weaknesses of standards controls. Although the standards institutions have assisted in raising the quality of many consumer products, their grip is weak. Most standards result from decisions of committees in which manufacturers usually have the final say. The recommended standard is thus more often a reflection of the industry's conscience than of the standard that would be required to provide satisfaction for the consumer. The standards laid down by manufacturers for a product can be so low that the consumer benefits little, if at all. Further, almost all standards refer to the safety of a product and not to its efficiency; and, with only a few exceptions, the recommendations of standards agencies are voluntary. The decision whether to adopt the standard is up to the company that markets the product, and such a decision necessarily involves an assessment of possible costs and returns. It is unfortunate that, in many

countries, the selling power of the standards symbol is less substantial than that of a good promotion campaign. Consumers, it would appear, are not sufficiently aware of the presence and significance of these symbols, perhaps because they tend to be little publicized by the manufacturers.

Apart from the formulation of standards, testing by various bodies occasionally results in the redesigning of certain products. Such testing has been most apparent in the automobile industry, in which cars have been recalled by their manufacturers so that alterations and improvements could be made. In the United States, for example, General Motors Corporation voluntarily recalled some 1,000,000 cars in 1971–72 for such changes.

Controls on advertising. Of all the criticisms levelled at manufacturers, those against their advertising probably have been the most vociferous. Advertising is necessarily vulnerable to these attacks: it is experienced by everybody, its products are on show for a long time, and its purposes are materialistic. Although the major purpose of company advertising, which is to attract members of the public toward buying a particular product, is fairly straightforward, the methods employed in this process have become increasingly complex. As business has become more competitive, so has the advertising that sells its products. Coupled with this increased competition has been the development of more powerful media—the most important of these being television.

Criticisms of advertising can be broadly divided into two: those that affect the consumer directly, and those that are concerned with economic matters, particularly the structure of industry.

From the consumer's point of view, the basic criticism of advertising is that it leads him to purchase goods that he has no wish to purchase by presenting misleading and untruthful statements or by creating wants, needs, and desires in his mind that might not otherwise exist. In the first instance it is accepted that the consumer, of his own volition, has a need that is filled by the description of the advertised product (but not necessarily by the product itself), whereas in the second the need is artificial and is stimulated entirely by the media.

From an economic viewpoint, critics of advertising point to the enormous amount of money involved—money that, they state, does not benefit the consumer although he is compelled to pay it. A second criticism is that advertising restricts competition because only large companies can afford expensive, nationwide campaigns, thus limiting freedom of entry of new firms into an established market.

A definitive answer to both these questions is obviously impossible. Regarding the first, it might be fair to say that economic growth and the creation of wealth might come about far more slowly without the aid of advertising. The development of national rather than regional brands—and the economies of scale implicit in this development—might be retarded. For all its drawbacks, advertising informs the consumer and enables him to make not only a choice between products but also a choice between the stores at which he can buy those products. For the manufacturer it justifies a heavy investment in capital and manpower in that it assures (to some degree at least) the quick development of sales.

Regarding the second major criticism—that advertising encourages the concentration of industry—there is no doubt that this is true. But not everyone agrees that industrial concentration necessarily acts against the interests of the consumer, particularly in the absence of outright monopolies or cartels. In some countries, such as the United States and Great Britain, anti-trust or monopoly laws act to restrain the more flagrant abuses of industrial power. Other countries, especially some in western Europe, have established monopolies boards, which monitor or oversee activities of large corporations in the field of takeovers and mergers.

The advertising industry has for many years been aware of the various criticisms and has accepted the need for some control over advertising methods in addition to the provisions of statutory regulations that exist in many countries. In the United States the Federal Trade Commission can stop advertising it considers misleading. In the United Kingdom the Department of Trade and Industry has some powers to do this under the Merchandise Marks Act. These powers are necessarily used in conditions of the last resort, the advertising industry itself preferring to control its members voluntarily. One method of self-control has been the establishment prior to World War II of the International Code of Standards of Advertising Practice under the auspices of the International Chamber of Commerce; it codifies the views common to national advertising groups on what constitutes good practice.

The country with the most stringent advertising standards is usually thought to be Great Britain, where, for example, all television advertising is controlled by the Independent Television Authority (ITA). The ITA lays down some 32 separate controls on advertising, banning the use, for instance, of subliminal advertising (methods by which the viewer might be influenced without his becoming aware of it) and of advertising that plays on fear and on the minds of the superstitious.

The ITA has a further list of unacceptable products and services; in the early 1970s it included cigarettes, bookmakers, undertakers, fortunetellers, and matrimonial agencies. Other regulations involve methods of television reproduction, the wording and advertising of guarantees, and the enforcement of prices and other offers; furthermore, special conditions exist in specific cases—the viewing child, the employment of children in advertisements, and the advertising of certain products such as medicines and drugs and also finance.

The general character of governmental and private controls over the claims and methods of advertisers may be said to be one of considerable laxity. It seems likely that this situation will be changed not so much by the introduction of more stringent codes as by challenges to particular advertisers by consumer interest groups within the framework of existing legislation regarding truth in advertising.

Labelling standards. Labelling can be used either to inform or to deceive the consumer, and manufacturers, in their sales efforts, are often tempted by the latter expedient. Minimum standards of labelling exist for some products, but, as with controls on manufacturing quality, legislation tends to concentrate on food and drugs. Usually, every container carries a statement of contents, but, apart from food and drugs, content identification is not usually required. If it is provided, however, it must not misrepresent. In general, this means that labelling, when it is present at all, tends to be accurate.

Regulation of labelling in Sweden

Consumer movements and official bodies have, in many countries, seen the need for better systems of product labelling. Of methods proposed, one of the best may be the Varudeklarationsnämnden (Quality Labelling Board) system adopted by Sweden, which is financed by the Swedish government in conjunction with various national business and consumer organizations. In this system, labels must describe the most important characteristics of the product and must do so in words that by any standards are explicit, many being more than 200 words long. Price labels are of further importance to the consumer. With the advent of self-service shopping, the need for goods to be priced correctly is essential. Vendors, however, are under no legal obligation to indicate prices, and a major criticism by consumer groups has been that, even when prices are indicated, it is often difficult to make price comparisons because of the lack of standardization of the weights or volumes of packages in which a product is sold.

Controls on sales methods. Generalizations cannot be made concerning statutory controls on sales methods because they vary from place to place. Sales practices have been controlled for over a century; early regulations were largely concerned with peddlers and hawkers. Legal progress has, in general, imposed a stricter control of selling methods to reduce the incidence of deception.

Particularly difficult to control is door-to-door selling, a method that for many years has drawn criticism from the

general public, even though the majority of door-to-door salesmen are fair and reputable tradesmen. Vacuum cleaners, floor polishers, sewing machines, and encyclopaedias have been sold by this method, some by salesmen who have exploited the purchasers' vulnerability. Salesmen's tactics often go far beyond the common foot-in-the-door technique. To persuade people to enter into a heavy financial committment, salesmen have been known to misstate the terms of payment or the trade-in allowance, to conceal figures on the order form or agreement, and to resort to other deceptive practices. Some countries have outlawed such deceptions. In Sweden a Door-to-Door Selling Act leaves buyers free to withdraw from contracts signed in their own homes and at exhibition stands within a period of seven days. A similar law in Britain specifies a period of four days. Some American states likewise provide statutory protection. The Swedish legislation is part of a package designed to help the consumer and includes an Undesirable Terms of Trading Act (for monitoring standard contracts) and further legislation regarding credit sales and hire purchase, legal aid, holiday travel, and foodstuffs.

Switch selling

Another technique used in direct sales is that of switch selling. The salesman attracts his victims by placing an advertisement offering a domestic article at a remarkably low price; this is known as the "bait." Inquirers are personally visited by a salesman, who, from the outset, makes no attempt to sell them the product advertised. Having convinced the inquirer that the model is not worth buying, the salesman goes on to offer the customer another model that he happens to have with him at, of course, a higher price. Although this and similar methods often are in violation of statutes governing the sale of goods, enforcement is difficult. Extra protection is provided by legislation in some countries, and, in others, non-statutory regulations protect the consumer.

THE CONSUMER MOVEMENT

The development of private consumer associations. In the four broad fields of manufacturing, advertising, labelling, and selling, legislation in most countries provides a background of protection for the consumer. Nevertheless, gaps remain, partly from ignorance on the part of consumers themselves and partly from the reluctance of most governments to unduly restrict the activities of business.

As industry and commerce have developed during the 20th century, pressures on the consumer have adversely affected the balance of power between buyer and seller. The recognition of this imbalance and the desire to restore it originated in the United States. In 1927 an economist, Stuart Chase, and an engineer from the American Standards Association, F.J. Schlink, wrote *Your Money's Worth*, a book about the activities of the National Bureau of Standards—the organization that tests all goods supplied to government departments and services. Their book, a best seller, concluded that the government was indeed getting its money's worth; it furthermore concluded that most consumers were not—mainly because they did not pay sufficient attention to the real values of the goods they were buying. This was the beginning of an association, run by Schlink, called Consumers' Research, from which a splinter group, formed in 1935 by a close associate of Schlink, Arthur Kallet, emerged as the Consumers' Union of United States in 1936. The character of this organization, which proclaimed that the function of the union was "... to provide consumer information and counsel relating to consumer goods and services, to give information and assistance on matters relating to the expenditure of the family income ... to initiate and to cooperate with individual and group efforts seeking to create and maintain decent living standards...." has been emulated and adapted by consumer groups operating in nearly every country in the world.

Objectives of the Consumers' Union

The Consumers' Union always was, and remains, totally independent of either industry or government. Its finance is derived solely from its members, who in 1971 numbered just over 1,800,000. Other associations have tended to differ in this regard, many of them government financed either wholly or in part. In Sweden, for example, the government took over the Institute of Consumer Information in 1957. The Swedish organization aims not only to inform the shopper of products that provide value for money but also to influence and cooperate with industry so that better products are made. The influence of the Consumers' Union in the United States on manufacturers is far more indirect and derives from its influence over consumers.

Until 1957 the Consumers' Union concentrated much of its effort on the production of comparative test reports on consumer goods. In that year, Arthur Kallet, the original director, was not reappointed. Instead, Dexter Masters, a journalist, was asked to head the team. This move from engineer to journalist was mirrored by the broadening interest of the union in the wider aspects of consumer protection. It became more concerned with the problems of obtaining justice for the consumer—attacking, for example, such practices as selling cosmetics in jars with false bottoms, cereals in packets only three-quarters full, and goods at odd weights and at odd prices. Further, it has sometimes been argued that from the late 1950s the Consumers' Union began to adopt a slightly ideological approach to consumer protection, reflecting to some extent a general antipathy to big business and profit making.

Consumer associations in other parts of the world do not generally share this outlook, although they do tend to concentrate either on product testing or on the wider ranging issues of consumer protection. In Great Britain, for example, the Consumers' Association, an independent, self-financed organization, is largely oriented toward product testing, whereas in Germany, the Arbeitsgemeinschaft der Verbraucherverbände (Consumers' Union of the Working Community), financed by 20 member bodies, takes after the modern-day Consumers' Union, campaigning for consumers particularly on issues such as labelling and standardization. In France, the Organisation Générale des Consommateurs is concerned with general aspects of interest to consumers as much as with comparative testing. Also in France, the Union Fédérale de la Consommation, an independently financed organization that published comparative tests, was forced to cease operations because of a lack of sufficient interest.

Consumers' organizations in Japan

Consumer organization in Japan differs considerably from that in other countries. First, it is extensive and increasingly well organized, particularly in consumer-boycott activity. Second, it is largely made up of women —500 housewives' groups, for example, comprise the biggest organization, called Shufuren. Through its quarterly magazine, Shufuren keeps its members informed of policy decisions—and also of which manufacturers are on its blacklist. Shufuren is active in product testing as well as in boycotts, which, in fact, remains its main work. Its members, however, see the need for a coordinated effort to be conducted on many fronts. The rapid development of the Japanese economy, which has been conducted largely behind the shield of high tariff barriers, has enabled manufacturers in Japan to have their own way for much of the time. Groups such as Shufuren are ensuring that this will change.

The increasing activism of certain sectors of the consumer movement has led to the formation of splinter groups of the well-established consumer associations. Of these, the U.S. consumerist Ralph Nader and his "raiders" (a group of mainly young helpers who gather information concerning suspected abuses) have received the greatest publicity. Nader has applied his own individual style to consumer protection, directing his campaigns mainly at particular manufacturers and their products. Nader's successes have been considerable and have been achieved largely by the effect of adverse publicity on manufacturers, rather than through the courts. His group also has acquired single shares of the stock of large industrial companies in order that it may monitor and, if necessary, criticize from the inside the operations and activities of these companies. Other similar groups have been formed on both sides of the Atlantic. Characteristically, their scope of operations is far wider than that of the original consumer organizations. They are, for ex-

ample, directing attention to the social, ecological, and environmental problems affecting society.

International organization of consumer movements. The development of uniform business techniques throughout the world, particularly in marketing, advertising, and manufacturing, coupled with the growth of multinational companies, increasingly has brought about the supplying of international markets with the same goods. Consumer organizations consequently have begun to pool their resources. Thus, in the European Economic Community (EEC), or Common Market, the member countries have set up, within the EEC's commission, a Committee of Contact that is designed to represent the interests of consumers throughout the Common Market. The Committee of Contact also organizes comparative tests of products common to all markets, known as Euro-tests.

In addition to regional groups, the International Organization of Consumers Unions (IOCU) provides for international coordination and cooperation. The IOCU, set up in 1960, was originally devised as a clearing house for testing methods but over the years has developed into an international forum for all kinds of consumer problems. Not the least of its functions is the representation it provides consumers within international agencies such as the Economic and Social Council of the United Nations. Strict rules limit the membership of the IOCU, particularly with regard to members' commercial and political links, and these constraints have tended to prevent the more politically oriented organizations from joining. The increasing interest in political matters shown by some of the founder members, such as the Consumers' Union, may, however, force the international organization to change its policy. It nevertheless seems likely that the influence of the IOCU will continue to develop, because its operations and methods of procedure are responsible and its suggestions tend to be practical and constructive.

BIBLIOGRAPHY. The literature on the consumer and his problems transcends national boundaries. GUNTER VOIGT and WALTER BERNAUER, *Der Verbraucher im gemeinsamen Markt* (1963), deals with the countries of the European Economic Community. A broader approach is JEAN MEYNAUD, *La Défense des consommateurs dans les sociétés capitalistes* (1968). Two Japanese books on the consumer are NORIYUKI OGI, *Shōhisha mondai o kangaeru* (1969); and KAZUKI DAIMON, *Tōdatsu no ronri* (1970). A French study is UNION FEMININE CIVIQUE ET SOCIALE, *À la recherche du bien commun* (1966). British books include GORDON J. BORRIE and AUBREY L. DIAMOND, *The Consumer, Society and the Law*, 2nd ed. (1968); and JOHN MARTIN and GEORGE W. SMITH, *The Consumer Interest* (1968). Among books published in the United States are: JAMES BISHOP, JR. and HENRY W. HUBBARD, *Let the Seller Beware* (1969); CURT GENTRY, *The Vulnerable Americans* (1966); GRANT S. MCCLELLAN (ed.), *The Consuming Public* (1968); WARREN G. MAGNUSON and JEAN CARPER, *The Dark Side of the Marketplace: The Plight of the American Consumer* (1968); SIDNEY K. MARGOLIUS, *The Innocent Consumer vs. the Exploiters* (1967); DAVID SANFORD (ed.), *Hot War on the Consumer* (1969); RALPH LEE SMITH, *The Bargain Hucksters* (1962); FRED TRUMP, *Buyer Beware!* (1965); and RALPH NADER, LOWELL DODGE, and RALPH HOTCHKISS, *What to Do with Your Bad Car* (1971).

(Mi.Do.)

Consumption, Economic

In economics the word consumption means the using up of goods and services. In modern economic writing it means, specifically, "final" consumption as distinguished from the using up of goods to produce other goods in industry. Final consumption must also be distinguished from the purchase by industry of fixed assets such as buildings and machinery, which is known as capital formation or investment. On the other hand, consumption expenditure by private persons is understood to include durable goods, such as furniture or vehicles, as well as works of art that may even increase in value with the passage of time. Strictly, the acquisition of such goods should be considered asset formation rather than consumption and should be classified together with the acquisition of other assets such as houses, schools, roads, and hospitals.

In modern industrial economies, consumption as defined above accounts for 70 or 80 percent of total national expenditure. Table 1 shows that even in Western capitalist countries a significant part of total consumption is determined directly by the expenditure of public authorities. Some of the benefits of this part of consumption, such as expenditure on defense or on public hygiene, are widely diffused; others are directed by common consent to the benefit of particular sections of the community. These consist in part of specialized services such as education or the medical care of the infirm or of expectant mothers; but other services—such as the provision of assistance to the unemployed, state pensions for old people, and assistance to children and wives deprived of the support of a family breadwinner—are designed to create greater equality in levels of consumption than would otherwise be obtained.

PATTERNS OF NATIONAL CONSUMPTION

The ways in which people spend their incomes show much uniformity among countries at the same economic level. Expenditure patterns in Great Britain, for example, are typical of western Europe. Table 2 shows how British consumers distributed their expenditures among the main categories of goods and services in 1949, 1959, and 1969. In 1949 the pattern was still affected by postwar shortages and rationing, but the level of total consumption was not very different from what it had been before the war. The 20 years that followed brought nearly a 50 percent increase in the level of private consumption expenditure per person (measured at constant prices), in addition to a great increase in public services such as health and education. Yet the broad distribution of expenditures remained strikingly constant in spite of the introduction of many new commodities and considerable changes in their relative prices. A few differences can be seen in Table 2. The percentage of total expenditure devoted to food fell, a phenomenon that usually accompanies a rising standard of living, and the largest proportionate increases were in the purchase and maintenance of private motor vehicles and of radio, television, and electrical goods. Both of these categories represent in part net additions to private wealth in the form of durable goods, and both reflect the effect of technical progress and cost reduction in the automobile, electric, and electronic industries. As in other industrial countries, much of the improvement in living standards has taken the form of more travel, better communication services, and the acquisition of labour-saving equipment.

Long-run changes

This trend, which may be found in all industrial countries, is not new. Over the past 150 years one may see a similar process at work on a still larger scale. In 1824 an English domestic economist suggested ways in which families with different incomes ought to budget their expenditures (Table 3). From actual budgets collected around that time, it would appear that such patterns were not untypical. Food expenditure then absorbed around one-half of the total, the fraction declining as income increased. Expenditure on clothing then was about twice the proportion of 1971, even though more clothes were made in the home. Fuel and light (in 1824 consisting of coal and candles) also absorbed about double the 1971 proportion of expenditures in working class families. On the other hand, the relative proportion spent on housing has remained the same.

The category "other expenditure" increased most rapidly with income in 1824; in 1971 it accounted for 42 percent of all private expenditure. In 1824 expenditure in this category largely represented purchases of direct labour services for the household and to a lesser extent purchases of horses, carriages, and the services of grooms and stable boys. In 1971 such direct labour services accounted for less than half of 1 percent of total expenditure, though somewhat more than this if one includes similar services that were then provided mainly in the home and are now obtained almost exclusively outside, such as hairdressing, catering, and vehicle maintenance. But in 1971 much of the "other expenditure" category was devoted to commodities either not invented or not

Table 1: Analysis of National Expenditures, 1970
(in percent)

type of expenditure	U.K.	France	West Germany	Netherlands	Sweden	U.S.	Japan*	U.S.S.R.
Consumption								
Consumers' expenditures on goods and services	63	59	54	56	55	63	51	61
Public authorities' current expenditure on goods and services	18	12	16	16	21	20	8	6
Total	81	71	70	72	76	83	59	67
Gross domestic capital formation								
Investment in fixed assets by government and industry	18	26	26	27	22	17	35	18
Value of physical increases in stocks and work in progress	1	3	2	2	3	0	4	11
Total	19	29	28	29	25	17	39	29
External transactions								
Exports of goods and services	23	16	22	49	25	6	11	...
Imports of goods and services	22	16	20	50	25	6	9	...
Difference	1	0	2	—1	0	0	2	2

*1969 figures.
Source: United Kingdom, *National Income and Expenditure*, 1972; United Nations, *Statistical Yearbook*, 1971.

distributed on a mass scale in 1824. These include books, newspapers, magazines, most drugstore items, recreational equipment of all kinds, and motor vehicles and electrical equipment. Another substantial part of "other expenditure" is absorbed by specialized nondomestic services such as insurance and the provision of entertainment, travel, private education, and health care.

Differences among countries

The extent to which recent changes in British consumption patterns are typical of other countries may be seen from Table 4. In most industrialized countries there has been a compound rate of increase in the total volume of consumption expenditure per person of 2 to 3 percent per year, the main exceptions being West Germany, where consumption has grown at double this rate, and Japan, where it has grown three times as fast, but the pattern of change is similar in almost all countries. Food consumption has grown less rapidly than total consumption, particularly in Scandinavia, the United Kingdom, and North America, where the rate of increase is by now about 1 percent per year; expenditure on clothing has been growing at about the same rate as total consumption. Changes in rent outlays reflect rent-control and public-housing programs in some countries. The acquisition of durable goods still continues at a very high rate in all countries except in North America, where the process of accumulation began much earlier.

Comparable data on consumption in poorer countries are much harder to obtain and are usually less reliable, but it is probable that, expressed as a proportion of total consumption, food expenditure is roughly twice as important over large areas of Asia, Africa, and Latin America as it is in western Europe and North America. In the economically most advanced countries, food expenditure represents only one-quarter to one-third of the total, whereas in countries where the total expenditure per household is less than the equivalent of U.S. $1,000, the proportion rises to one-half or even greater (Table 5). It is important to note that in the rural regions of poor countries housing expenditure is minimal; in these regions shelter is rudimentary and largely self-provided.

Food consumption varies in character from country to country. In part this is due to climatic factors and reflects differences in long established food habits. The diet in northern Europe and Scandinavia is relatively poor in fruit and vegetables but contains a high proportion of milk, fats, and sugar. In France the consumption of vegetables and meat is high. Fruit and vegetable consumption is generally high in southern Europe, and milk consumption is low. In Mediterranean countries food grains are preferred to potatoes and sugar as sources of carbohydrates. But over and above these regional variations, the influence of general living standards is evident. The North American diet, for example, with its low grain and potato consumption and high consumption of sugar, meat, eggs, and fats is attributable more to a high standard of living than to regional peculiarities of taste. These characteristics can be observed in the diets of the wealthier classes of most countries.

Table 2: Private Consumers' Expenditures in the United Kingdom
(in £ at 1963 market prices)

item	£ per person per year			percentage increase of 1969 over 1949	percentage distribution in	
	1949	1959	1969		1949	1969
Food	73	85	89	21	26	22
Alcoholic drink and tobacco	37	43	47	26	13	12
Clothing and footwear	27	31	37	38	10	9
Housing services	32	37	47	48	11	12
Fuel and light	12	15	22	77	5	5
Furniture and household goods	7	9	8	43	2	2
Electrical goods	3	8	9	177	1	2
Motor vehicles and running costs	15	17	37	149	5	9
Other goods	25	32	36	39	9	9
Other services	46	63	75	63	17	19
Total*	278	341	407	46	100	100

*Items do not always add to totals because of rounding; percentages have been computed before rounding the expenditure figures.
Source: United Kingdom, *National Income and Expenditure*, publications for various years.

Table 3: Patterns of Consumption Expenditure in the United Kingdom, 1824 and 1971
(in percent)

item	recommended budget proportions in 1824 for families with annual incomes of:*			average budget proportions, United Kingdom, 1971
	£55	£86	£250	
Food	51	49	43	21
Alcoholic drinks	6	7	8	12
Clothing	18	18	16	9
Housing services	12	12	11	11
Fuel and light	11	8	5	5
Other expenditure	2	6	17	42

*Each family is assumed to consist of a man, his wife, and three children. The first two budgets are for families of skilled labourers, the third for a middle class family keeping one servant.
Sources: Figures for 1824 are from *A System of Practical Domestic Economy* by Mrs. Rundell, 1824. Figures for 1971 are from United Kingdom, *National Income and Expenditure*,

Table 4: Percentage Increases in the Volume of Private Consumption per Person 1959–1968

country	food	clothing	rent	durables	other	total
France	38	62	66	127	70	62
West Germany*	43	55	90		82†	70
Netherlands	33	62	27	80	77	66
Sweden	22	21	44	55	61	43
U.K.	18	31	46	9	13	27
U.S.	22	49	54	84	48	47
Japan	64	127	68		215†	118

*Percentages computed for the 1958–1967 period. †These figures represent percentage increases for durables and other products combined.
Sources: Organization for Economic Cooperation and Development, *National Accounts Statistics* 1950–1968: United Nations, *Statistical Yearbook, 1971.*

Table 6: The Income Elasticity of Food Expenditures in Rich and Poor Countries*

country	bread	fruit	vegetables	meat	fish	milk	all food
U.S.	−0.3†	0.3	0.1	0.2	0.2	−0.3	0.2
U.K.	−0.2†	0.5	0.3	0.3	0.3	0.0	0.3
Italy	−0.1†	0.8	0.4	0.8	0.4	0.6	0.5
Poland	0.1†	0.9	0.6	0.6	0.7	0.3	0.7
Japan	0.7‡	0.8	0.7	1.2	0.6	0.7	0.6
Philippines	0.4‡	0.7	0.7	1.2	0.7	1.5	0.8
Argentina	0.1‡	0.6	0.3	0.3	0.4	0.2	0.3
Ivory Coast (urban)	0.5‡	0.2	0.3	0.6	0.5	0.0	0.5
India (urban)	0.4‡	0.8	0.7	0.5	1.5	1.6	0.8

*The income elasticity of food expenditure is defined as the percentage increase in expenditure on food associated with a 1 percent increase in income. †Wheat. ‡Cereals.
Source: The United Nations Food and Agriculture Organization: unpublished analyses based on official household surveys.

The influence of the general standard of living is also shown in the relative priorities accorded to increased consumption of particular foods as incomes increase. These priorities are measured by economic statisticians in the form of income elasticities of expenditure, defined as the percentage increase in the consumption of an item divided by the percentage increase in income that makes it possible. These elasticities are usually calculated for a given country by comparing the budgets of rich households with those of the poor. Table 6 shows that the income elasticities for most items of food expenditure in the United States and the United Kingdom are of the order of 0.3 or less. This means that an increase in income of 10 percent will produce an increase of 3 percent or less in food expenditure. The consumption of cereal foods in those countries actually decreases as incomes increase. In the less developed countries the elasticities are usually considerably higher, particularly for products of animal origin and fruit. In these countries even the consumption of carbohydrate foods is still increasing fairly rapidly as incomes rise. Surveys of household expenditures in Africa, Southeast Asia, and Latin America still show total consumption levels of no more than 2,000 kilocalories per person per day in rural areas and low levels of consumption of proteins and important vitamins. The difficulty of raising these nutritional levels is enhanced by the rapid rate of growth of population.

FACTORS INFLUENCING CONSUMERS

The theory and measurement of consumer behaviour forms an important part of modern economic theory. It was first developed during the 19th century on the basis of the following conception: from the purchase of any commodity the consumer derives a positive satisfaction or utility; but the additional satisfaction that he derives from additional purchases of the same commodity declines as his supply of that commodity increases; and with a given amount of money to spend, the consumer distributes his expenditure among commodities so as to maximize the total satisfaction or utility that he can get from all of his purchases. This rather crude model of consumer behaviour has undergone considerable refinement by modern mathematical economists. The virtue of this approach, which has had a profound and enduring effect on the theoretical and empirical work of economists, is that it separates the main economic variables influencing consumer behaviour—that is, income and prices—from all the remaining influences, such as individual preferences, social pressures, customs, habits, and so forth, but at the same time it unites them all in a single analytical apparatus. Critics have often objected that the model assumes a rational person bent on scrupulously maximizing his satisfaction and that the model is thus part of a mechanistic stream of thought that has been substantially undermined by 20th-century advances in psychology. A concise and clear defense of the economist's approach is provided in the argument that the only useful criterion of any hypothesis is the range of situations in which the derivative model is shown validly to predict events. For example, it is useful to assume that the leaves of a tree attempt to maximize the amount of sunlight they receive, since the assumption implies that leaves are denser on the sunny than on the shady side of trees and this can be checked from experience, or that billiard players make their shots as if they knew the mathematical formulas of mechanics. Similarly, to assume that consumers behave as if they were rational utility maximizers helps to provide accurate predictions of a broad range of market phenomena; *e.g.*, a fall in the price of a commodity will generally lead to increased consumption of that commodity, and an increase in consumer income will lead to increased consumption of most commodities. Only persistent discrepancies between predictions and events will require a modification of the model's assumptions; some examples of such cases are discussed below.

Spending in relation to incomes

The theory points to the income of the consumer as the most important single determinant of his consumption pattern. It follows that in any community both the average level and the distribution of incomes are important influences on total consumption. A community in which incomes are equally distributed will consume fewer luxury goods and fewer poor quality goods than one containing a few rich individuals and many poor. Among wealthy people in early 19th-century England, a dinner with five main dishes—fish, meat, game, poultry, and ragout with truffles—was described as the minimum, while in poor years the families of agricultural labourers ate mainly oatmeal and potatoes; today the standardized produce of modern agriculture is present on all tables.

The classic model of consumers' behaviour implicitly assumes that the individual enjoys a constant income. In practice it may fluctuate according to the season, from year to year, or more generally over his span of life. In the short run the consumption of some commodities is much affected by these income fluctuations, while the consumption of others is affected very little. Observers of wage-earning households have noted that it is common for the husband to give a constant weekly housekeeping allowance to his wife, out of which the necessities of food and clothing are bought, while the variable excess of his

Table 5: Consumption Expenditures Shown by Family Budgets in the mid–1960s

households surveyed	total annual consumption expenditure per household (U.S. $)	pattern of expenditure (percent)			
		food	clothing	housing	other
U.S.	5,522	24	9	12	55
U.K.	2,840	33	9	11	47
Italy	2,303	43	10	12	35
Poland	1,569	46	16	2	36
Japan	1,266	37	11	7	46
Kenya (urban)	1,007	40	9	9	42
Argentina	875	59	20	3	18
Ivory Coast (urban)	793	50	8	11	31
Ghana (Accra)	699	54	12	10	25
Ghana (rural)	456	60	15	1	24
India (urban)	294	62	8	4	26
India (rural)	235	70	10	1	19

Source: International Labour Organisation, *Household Income and Expenditure Statistics* (1967).

earnings is spent on drink, tobacco, and entertainment. The expected average level of future income therefore influences consumption habits as much as actual present income, and one may divide commodities into two classes: the first consisting of goods men buy when temporarily affluent but give up when temporarily poor and the second consisting of goods for which the pleasure of a temporarily higher level of consumption would not be worth the financial or psychic cost of giving them up in the future.

Consumers are also influenced by the history of their incomes. A man who owns an expensive car may continue to use it after his income falls, though at the lower level of income he would not in the long run choose to replace it. This may be rational, in the sense that the value of the car to him in use may be greater than what he can get for it in the second-hand market, or it may be irrational, in the sense that he is simply continuing a habit that he should abandon. The distinction is largely subjective and cannot be clearly made by an outside observer.

Over the life cycle as a whole, consumption patterns are markedly different in different occupations. In most unskilled or semiskilled occupations, the course of earnings is fairly stable: a man of 21 may earn as much as anyone older. But in many professions the man of 50 or 60 may earn many times the income of the man of 21; this gives the young man a strong incentive to incur considerable debt with the expectation of amortizing it steadily through life, so that the typical consumption pattern of the occupation can be achieved earlier than otherwise. This applies particularly to the purchase of houses, household furniture and equipment, and vehicles. The manual worker on the other hand, whose expectations are little greater than his present consumption, will prefer to rent rather than to own a house and will not be willing to raise his current consumption standards by incurring commitments of a longer term than that of ordinary installment credit.

Non-rational influences

To be fully rational and consistent, the consumer would need to have access to sufficient information on goods and their prices so that he could choose those with the lowest unit price for a given quality. Many observers have pointed to the fact that consumers do not always behave this way. Natural pearls are sold at a much higher price than cultured pearls, though the difference between them is demonstrable only by dissection or with X-rays, and their quality in use is identical. Proprietary brands of drugs sell better and at higher prices than unbranded drugs made from the same standard formula. To some extent this is due to what an American economist, Thorstein Veblen, called the desire for conspicuous consumption: part of the attraction of the good is simply its high price. It is also the result of consumers' ignorance, made more acute by the increasing sophistication of commodities whose qualities must be measured in many dimensions. If it is costly in time for the individual fully to inform himself on the comparative qualities of competing products, it is not wholly irrational for him to take the market price as an indicator of quality. The lack of information has given rise to consumers' organizations in most industrialized countries that test and report on a wide range of products for their subscribers.

The influence of modern advertising techniques must also be considered. Insofar as advertising informs the consumer of the range of alternatives, it merely increases his information; and insofar as advertising consciously or subconsciously changes consumer preferences, it remains one of many factors determining consumer preferences that the economist takes as given. But advertising cannot persuade the public to buy whatever the producer offers. It is likely to be most effective in influencing consumers to choose one from among several almost identical products, such as toothpaste, cigarettes, or gasoline. But it may also raise the demand for the group of competing products as a whole. It can be argued that the total effect of modern advertising is to shift the preferences of consumers in favour of luxuries rather than necessities, in favour of consumption rather than saving, and in favour of income as against leisure.

NECESSITIES AND LUXURIES

The distinction between necessities and luxuries is hard to make precise. The dividing line varies with the income and social class of the classifier and shifts over time as technology develops and social mores change. Only in the most primitive and poorest communities can necessities be defined purely in terms of physiological needs. Adam Smith wrote in 1776:

> By necessaries I understand not only the kind of commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without. . . . Under necessaries, therefore, I comprehend not only those things which nature, but those things which the established rules of decency have rendered necessary to the lowest rank of people. All other things I call luxuries; without meaning by this appellation to throw the smallest degree of reproach upon the temperate use of them. Beer and ale, for example, in Great Britain, and wine, even in the wine countries, I call luxuries. A man of any rank may, without any reproach, abstain totally from tasting such liquors. Nature does not render them necessary for the support of life, and custom nowhere renders it indecent to live without them.

In the 19th century, with the development of more mathematical methods of reasoning based on a utilitarian calculus, the distinction came to be phrased differently. Necessities were defined as those commodities the demand for which has an income elasticity less than unity, and luxuries as those with an income elasticity greater than unity. These definitions imply that as a man's income increases his expenditure on necessities increases less than, and his expenditure on luxuries more than, proportionately. But even with the elasticity approach the distinction must vary over time. In 1950 the demand for television sets was highly income elastic, whereas now, in some countries, television is regarded as a necessity in the great majority of households.

Changing attitudes toward luxury

Economists of the early 19th century all believed that under capitalism the living standards of the working classes would remain close to a subsistence level, meaning that luxuries would be more or less permanently denied them. But in modern industrialized economies even the poor consume goods that the early economists would not have considered in any sense necessary.

The historical and social role of luxury consumption is a subject of much interest. In the Mediterranean city-states during the Renaissance, the demand for luxuries provided a mainspring for the specialization of skilled labour and for the development of foreign travel and long-distance trade. The duke of Milan, Filippo Maria Visconti, possessed valuable English dogs, leopards from all parts of the East, and hunting birds from northern Europe. Some writers have argued that the luxurious consumption of the rich benefits the poor through the provision of employment opportunities that would not otherwise exist. A subtler version of this idea was entertained by Adam Smith, who contrasted the uselessness of menial labour employed by the rich for their personal service with the benefits flowing from the employment of craftsmen who created luxurious products of enduring merit that eventually became available to society as a whole:

> The houses, the furniture, the clothing of the rich, in a little time, become useful to the inferior and middling ranks of people. . . . What was formerly a seat of the family of Seymour is now an inn upon the Bath road. The marriage-bed of James the First . . . was, a few years ago, the ornament of an ale-house at Dunfermline.

But Smith and most of the economists who succeeded him believed that if the wealth expended on luxurious consumption by the rich were saved and invested in useful production, society would benefit as a whole. The Industrial Revolution brought an ever increasing demand for funds for productive investment and made possible a more rapid rise in general standards of living than the world had known before. The classical economists argued therefore that all luxury consumption involved a selfish diversion of labour and capital and acted as a brake on human progress.

This view was not seriously challenged until an English economist, J.M. Keynes, published his *General Theory*

of Employment, Interest and Money in 1936. Writing at a time when millions of workers were unemployed, Keynes argued that the consumption of luxuries was socially desirable if it provided jobs that would otherwise not exist. He also suggested that capitalism might be outrunning its investment opportunities, so that in the long run the problem of finding employment for capital itself would arise—a difficulty that might be postponed if the rich were to spend more on themselves:

> In so far as millionaires find their satisfaction in building mighty mansions to contain their bodies when alive and pyramids to shelter them after death, . . . the day when abundance of capital will interfere with abundance of output may be postponed.

In industrial countries since World War II, this pessimistic view has been overborne by a seemingly endless boom in consumer industries. As fast as consumers accumulate durable goods, these become technologically or conventionally obsolete and are replaced by new goods. Instead of seeking more leisure, previously thought to be one of the main benefits of technical progress, the populations of the industrialized countries seem to prefer to work in order to buy more luxuries. To this extent the desire for leisure and the demand for luxuries are in direct competition.

In communist countries, public consumption has long been treated as more important than private luxury. In the 1960s this emphasis seemed to be giving way to the aim of catching up with the standards of consumption prevailing in capitalist countries. In the poor countries of the Third World, the tension between the demand for luxuries and the low standard of living gives rise to acute economic and social problems. The rapid growth of international travel and communications since World War II has led the literate and skilled classes of every nation to seek similar standards of private consumption regardless of their national environment. This, in the poorer countries, leads either to a highly unequal distribution of the national income or to the emigration of the skilled population. Thus the increasing awareness of the consumption habits of the most fortunate sections of the world's population is both a spur and a hindrance to general progress.

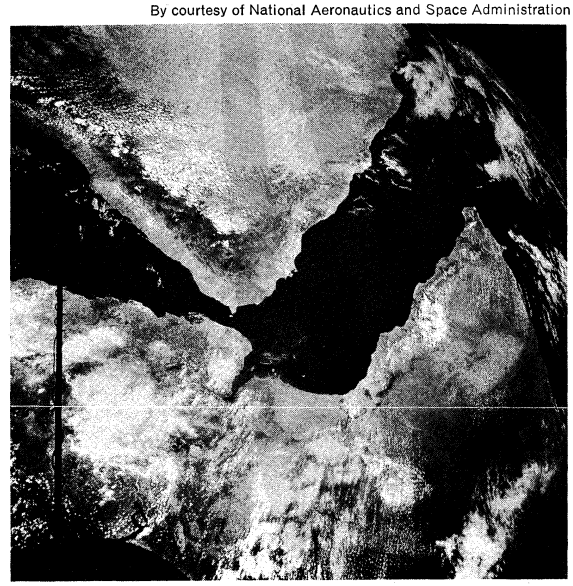
BIBLIOGRAPHY. J. BURNETT, *Plenty and Want* (1966), a description of the food habits and dietary standards of the British in the 19th century; C.M. CIPOLLA, *European Culture and Overseas Expansion* (1970), an account of the cultural and commercial contacts between Europeans, Africans, and Asians during the Renaissance; JOHN KENNETH GALBRAITH, *The Affluent Society* (1958), a critical study of the social and economic consequences of advanced industrialism in which basic human needs can be easily satisfied, but in which the economy is geared to the pursuit of luxury consumption by the middle and higher income groups; S.S. KUZNETS, *Modern Economic Growth* (1966), a comparative study of the historical development of modern economies, based on exhaustive statistical studies; E.J. MISHAN, *The Costs of Economic Growth* (1967), a critique of the goal of economic growth as a major aim of policy; VANCE PACKARD, *The Hidden Persuaders* (1957), a somewhat partisan attempt to show that advertising techniques create artificial wants through psychological manipulation; THORSTEIN VEBLEN, *The Theory of the Leisure Class* (1899, reprinted 1967), an argument that the consumption of the wealthy classes is aimed mainly at demonstrating status rather than satisfying basic needs.

(J.A.C.B.)

Continental Drift

The large-scale horizontal displacements of continents, relative to each other and to ocean basins, during one or more episodes of geological time is termed continental drift. The opposing view is that continents have grown *in situ* from nuclei that had fixed positions of latitude and longitude at the original formation of the Earth's crust. Continental drift was first proposed in a comprehensive manner by Alfred Wegener, a German astronomer and meteorologist, in 1912. His radically new interpretation of Earth history aroused international interest and violent controversy. By 1930 the majority of earth scientists had abandoned the drift hypothesis, mainly because of geophysical evidence that the crust and mantle are both solid and possess adequate strength to resist displacement by any known tangential forces. Interest in continental drift has revived dramatically in recent years as a result of advances in rock magnetism (*q.v.*), geochronology, and oceanography. Sea-floor spreading (*q.v.*), a concept first introduced about 1960, has given rise to plate tectonics, a synthesis of geological and geophysical observations that views the Earth's surface as divided into a few rigid plates undergoing rotational motion relative to each other. Today, the early model of the continents as icebergs floating on a dense fluid has become obsolete, but the weight of evidence does favour a modified continental drift theory. Estimated rates of movement between continents, or crustal plates, range from 1 to 15 centimetres per year.

By courtesy of National Aeronautics and Space Administration



Red Sea and Gulf of Aden rifts photographed by the Gemini XI space mission at a distance of about 470 nautical miles.

HISTORICAL BACKGROUND

The apparent fit of the bulge of eastern South America into the bight of Africa, which can be seen on any map or globe, is a striking physical feature that has caused many laymen and scientists to speculate on continental drift.

Early theories. The English philosopher and man of letters Francis Bacon was possibly the first observer to comment in writing on the correspondence of the shorelines across the Atlantic; he suggested in 1620 that the similarities were not accidental. The German naturalist Alexander von Humboldt proposed around 1800 that the lands bordering the Atlantic were once joined and had been scoured apart by the waters of the ocean. During the 19th century, several investigators published discourses on drifting continents; they were largely ignored because scientific thought was then dominated by the contraction theory and its model of a cooling, shrinking Earth, its crust crumpling and shearing under compression, but with continents and ocean basins maintaining a permanent configuration.

The concept of continents floating in isostatic equilibrium in a dense substratum was proposed because of measured gravity anomalies during the latter half of the 19th century. By this principle, vertical movements are attributed to mass changes in crustal columns that arise because of factors such as erosion, intrusion, and chemical or phase changes at depth. Although isostasy is fundamental to the continental drift hypothesis, it does not of itself imply that large-scale horizontal motion is possible.

Between 1885 and 1909, an Austrian geologist, Eduard Suess, published his masterwork, *Das Antlitz der Erde* (English translation *The Face of the Earth*), in which he described the contracting Earth as consisting of three rock shells, for which he coined the terms "nife" (Ni-Fe, for

Nickel-Iron) for the core, "sima" (Si-Mg, for Silicon-Magnesium) for the mantle, and "sal" (Si-Al, for Silicon-Aluminum) for the crust. The last two names (with "sial" substituted, by Wegener, for "sal") are now in common usage in the geological literature, but their meanings are variable. In outlining the distribution of shield areas (stable continental nuclei) and orogenic belts (zones of mountain building), Suess concluded that until the end of the Paleozoic Era (225,000,000 years ago) there had been several very large continental masses, two in the Northern Hemisphere and one, or possibly two, in the Southern Hemisphere, separated by an east-west seaway termed Tethys. He believed that large segments of these continental plates foundered during the Mesozoic Era (*q.v.*), forming new ocean basins, and that the Tethys closed under compression during the Tertiary Period (*q.v.*), producing mountain chains from Morocco to China.

Gondwana

Suess' southern continents included one he named Gondwana, meaning Land of the Gonds, for the typical locality of a distinctive series of Late Paleozoic sediments in India. He conceived of Gondwana, or Gondwanaland, as a vast sialic plate spanning over 150° of longitude, from South America across the Earth to the far side of peninsular India. Suess was uncertain whether to include Australia and Antarctica in the same crustal mass, but his followers tended to do so. Suess concluded that the continuity of Gondwana was destroyed by the faulting and foundering of the South Atlantic and Indian Ocean segments during the Mesozoic Era. His deductions that India belonged with the southern continents and that these landmasses had been continuous were later adopted into the canons of the continental drift hypothesis. However, his view that the oceans are fault valleys floored by sunken continents is clearly incompatible with drift. Indeed, this idea was one of the more powerful arguments used against the drift hypothesis until after World War II, when extensive gravity and seismic measurements made at sea demonstrated that the Atlantic and Indian oceans have basaltic floors.

Continental drift has repeatedly been ascribed to catastrophes involving the Moon. In 1882 Osmond Fisher, an English physicist, elaborated on the 1879 thesis of Sir George Darwin that the Moon was torn from the Earth, early in its history, by resonance effects between the rotating planet and the solar tides. Fisher proposed that the Pacific Ocean Basin is the scar produced by this event and that the other oceans were formed as rifts in the remaining sial. In 1910 F.B. Taylor explained the world distribution of Tertiary mountain ranges by suggesting that two proto-continents, located over the Earth's poles, had ruptured and moved toward the Equator under the force of lunar tidal action. He knew that, at its present distance, the Moon could not raise tides sufficient to the purpose, and so he postulated capture of the Moon at close range in the Late Cretaceous Period.

Wegener's hypothesis. Alfred Wegener published two articles in 1912 outlining his wholly original continental drift hypothesis. His hypothesis was by far the most detailed and comprehensive then proposed. It sparked worldwide interest, much support, and violent opposition. Its effectiveness shocked orthodox scientists into a defense of their positions; on the other hand, it was hailed with enthusiasm by those who had found the contraction hypothesis insufficient to explain geological observations. Wegener continued his research and revisions until 1930, when he died while on an expedition to Greenland.

The international controversy

Meanwhile, the international controversy over his concept made his name synonymous with continental drift, which is now universally referred to "Wegener's hypothesis."

Wegener deduced from gravity measurements and the bimodal distribution of the Earth's surface elevations that the ocean floors consist of denser, more iron-rich rock than do the continents. To him, the geological similarities of widely separated continents indicate the splitting and drifting apart of formerly continuous landmasses. According to Wegener's model, the continental sial (granitic rock with an average density of 2.70 grams per cubic centimetre) is balanced isostatically in the denser oceanic layer (basalt with an average density of 2.95 grams per cubic centimetre), which he called sima and regarded as the top of the Earth's mantle. He compared the continents with icebergs that are buoyed in water at heights proportional to their mass and moving through the yielding sima.

In support of his hypothesis Wegener plotted the distribution of orogenic belts, geologic contacts, paleoclimatic zones, and living and fossil plants and animals. He concluded that until the Jurassic Period (136,000,000 to 190,000,000 years ago) all the sial had been massed in one large protocontinent, termed Pangaea, which covered about one-half of the globe and was surrounded by the primeval Pacific Ocean. Pangaea was like a grounded ice floe undergoing internal rifting and healing throughout most of Earth history until a definite breakup began in the Jurassic. In Wegener's day this was believed to be about 40,000,000 years ago; today the beginning of the Jurassic is dated at about 190,000,000 years ago. Wegener believed that at that time separate continental blocks began floating apart in the partially liquefied sima, with both an equatorward and a westward component to their motion. He postulated that as the continents moved forward, new ocean basins opened behind them; that the western margins of the American continents finally en-

Base from E. Bullard, "Symposium on Continental Drift," *Philosophical Transactions*, (1965); Royal Society of London; other data from P. Hurley and J. Rand, "Pre-drift Continental Nuclei," *Science*, Vol. 164, June 13, 1969; American Association for the Advancement of Science

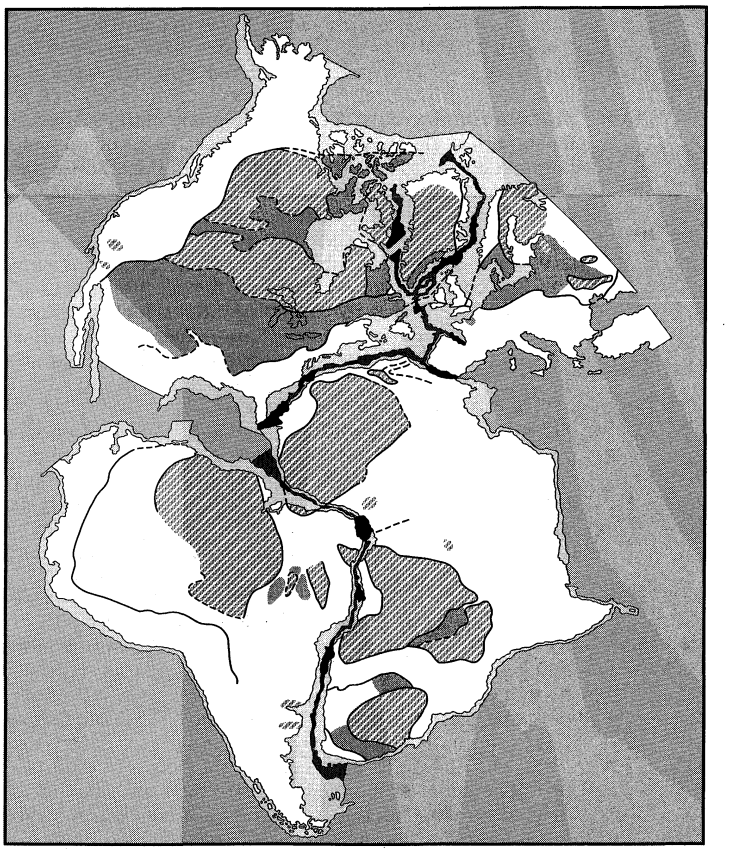


Figure 1: Reconstruction of the fit of the continents around the Atlantic prior to drift, showing Precambrian continental nuclei and continental margins to a depth of 1,000 metres.

countered resistance from recongealed sima that in turn caused the crumpling and uplift of marine sediments and continental margins to form the major mountain systems; and that the island arcs (*q.v.*) of the western Pacific are festoons that dragged behind as Asia moved westward away from them. In order to explain the distribution of paleoclimates he proposed a long history of polar wandering, with the Mesozoic south pole located in southern Africa.

Wegener originally suggested two possible causes of drift: a pole-fleeing force, which tends to move matter equatorward; and lunar–solar tidal force, which moves it westward. Both forces exist but are very small indeed. Wegener believed that these forces would become effective if applied over many millions of years. Most geophysicists rejected these forces as totally inadequate regardless of time.

Wegener's views made a direct appeal to numerous geologists, particularly to Alexander Du Toit in South Africa and to Arthur Holmes in England. For many years after the death of Wegener both men continued to develop evidence for continental drift and to work on the problem of a driving mechanism. Each concluded independently that the Earth is neither shrinking nor expanding (the latter having been widely believed after the discovery of radioactivity in crustal rocks) but that horizontal displacements occur, largely as a result of thermally driven currents in the mantle. Du Toit developed the idea that gravity sliding, the creep of continental masses toward rimming geosynclinal depressions, was an important factor in drift.

Opposition to drift

By the mid-1920s proponents of continental drift were working against a great tide of opinion in North America and Europe. The hypothesis suffered primarily from the lack of a theoretically sound mechanism adequate to produce continental drift, from too many enthusiastic but unsupported claims of matching shorelines and moving landmasses, and from the fact that it explained only one very late episode in Earth history, whereas the continents show a long record of tectonic activity. A definite turning point in scientific attitude occurred in 1926 at an international symposium held in New York. After this symposium, continental drift suffered a widespread loss of support and, ceasing to be a subject of serious investigation, became instead a source of amusement and derision. This overwhelmingly negative reaction is difficult to understand, particularly because the chairman and about half of the participants favoured drift.

The reaction was essentially due to the opposition expressed in 1924 by Harold Jeffreys, an English geophysicist, whose enormous prestige proved decisive with many geologists. Jeffreys argued that the force of gravity is stronger than any known tangential force acting upon the Earth's crust and that because the continental and oceanic layers are sufficiently strong to maintain topographic features such as Mt. Everest and the deep ocean trenches without slowly spreading out under the pull of gravity they are clearly too strong to permit horizontal drifting of sialic blocks through the sima. He was particularly opposed to the apparent inconsistency in Wegener's view that as continents drifted apart in the yielding sima they eventually met resistance enough to cause the crumpling and uplift of mountains on their forward margins. The sima, he pointed out, is either soft enough to yield to moving rafts of sial or it is not, and seismic evidence indicates that it is not. Jeffreys expanded upon his arguments in numerous journals and through five editions of his book *The Earth*, and he remains today one of the strongest opponents of continental drift and also of the concept of convection in the Earth's mantle.

Creative thought on the continental drift hypothesis lapsed for a number of years, and the arguments became increasingly repetitive and sterile until new types of evidence began to appear after World War II and to proliferate during the 1960s. Once again it was a symposium, held in London by the Royal Society in 1964, that generated a marked change in the attitudes of scientists toward the subject. The papers presented included reviews of paleomagnetic studies, new information on ocean-floor topography, and other relevant subjects. One memorable offering was a predrift reconstruction of the continents bordering the Atlantic Ocean prepared by Sir Edward Bullard and his colleagues (Figure 1). This reconstruction, produced by a computer programmed to find the best fit of the Atlantic margins, disposed of the earlier arguments that the fit is very poor or nonexistent. It also prompted investigators to use it as a base map in searching for geological contacts truncated by the ocean.

EVIDENCE FOR CONTINENTAL DRIFT

Continental drift is postulated to explain such geological and geophysical observations as the following: the apparent matching of continental margins, of truncated mountain belts, and of dated contacts between geological formations that are separated by oceans, particularly the Atlantic; the bizarre pattern of Late Paleozoic glacial deposits, plants, and animals; the apparent migration of the earlier magnetic poles along paths that differ for different continents, as suggested by remanent magnetism (that which is related to the Earth's magnetic field at the time of formation of a given rock) in rocks older than mid-Tertiary; the young ages of marine sediments and of the ocean floors; the linear patterns of magnetic anomalies that parallel the oceanic ridges, coupled with seismic evidence for transform faulting and sea-floor spreading (*q.v.*).

Geological data. The Atlantic shorelines of Africa and South America suggest complementary pieces of a fractured block, as do those of Europe and North America (Figure 2) when allowance is made for their broad

From "The Confirmation of Continental Drift," by Patrick M. Hurley. Copyright © 1968 by Scientific American, Inc. All rights reserved

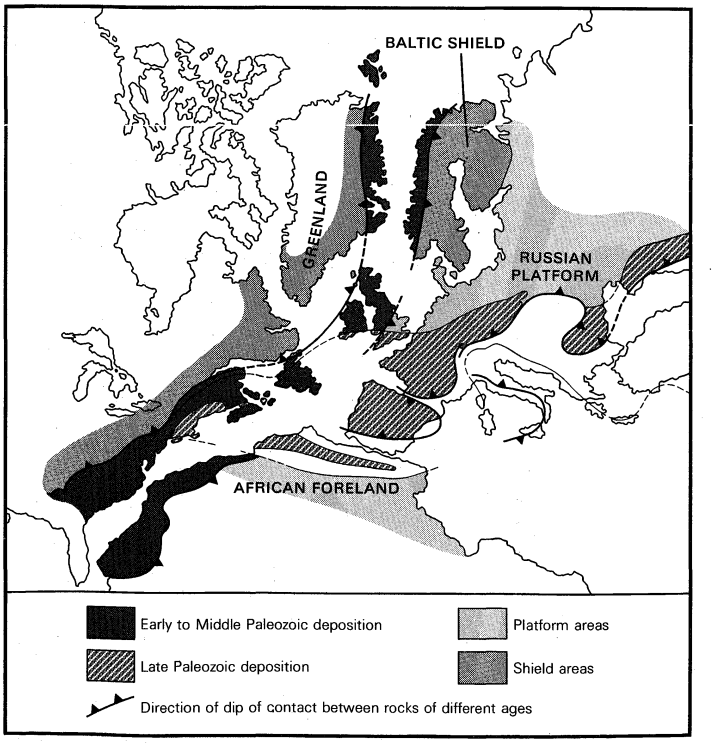


Figure 2: Geological fit across the North Atlantic, showing structural trends and ages of rocks.

continental shelves. Figure 1 illustrates the match along the 500-fathom line of the continental slopes. This configuration requires the rotation of Spain into the Bay of Biscay and the omission of most of Central America and the Caribbean Islands, including areas where Paleozoic rocks are known to occur. It is, nevertheless, so excellent a fit that it encourages a search for the matching of geological provinces.

Ages of rocks. Before the development of radiometric dating methods, the equivalent ages of orogenic belts

and rock contacts truncated by the Atlantic coastlines were often in question. Today, dated contacts showing a remarkable degree of continuity have been mapped across both the North and the South Atlantic. Joint investigations carried on by geochronologists in the United States and Brazil, for example, involved a contact in the crystalline basement rocks of Ghana. This was dated by both the potassium-argon and the rubidium-strontium methods, was plotted on the map shown in Figure 1, and then extended to northeastern Brazil. Subsequent sampling and dating disclosed the contact in its predicted location.

Farther south along the Brazilian coast a second belt of ancient rocks (2,000,000,000 years old) matches one in West Africa. The elegant fit of these contacts, in addition to that of the continental margins, makes a strong case for the continuity of Africa and South America during Precambrian time (*q.v.*). That this continuity persisted into the Mesozoic is suggested by the matching of Late Paleozoic fossiliferous strata and by the parallelism of Triassic dike systems in the two continents. Furthermore, the earliest marine deposits along the Atlantic coastlines of either Africa or South America are Jurassic in age, suggesting that the ocean did not exist prior to that time. Ancient rocks that are 1,700,000,000 to 3,500,000,000 years old occur in all continents and on present maps appear random in distribution. When they are plotted on the predrift reconstruction map of Figure 1, they fall into two groups that appear to represent ancient and rather small continental nuclei. These nuclei are rimmed and embayed by rocks 800,000,000 to 1,700,000,000 years old, which in turn are surrounded and crosscut by younger rock. Two conclusions have been derived from these data: (1) the continental crust has evolved from two ancient Precambrian nuclei and has grown at an accelerating rate throughout Earth history; and (2) these nuclei grew in place until the Jurassic when, according to the geological evidence cited above, they began to fragment and drift apart.

One of the most potent arguments against Wegener's hypothesis was that it accounted for only one late episode of continental drift in an Earth with a very long record of tectonic activity. The latest evidence from geochronology tends, however, to support Wegener's view that drift has occurred only once, beginning in the Mesozoic. It also supports Du Toit's conclusion that there were two original continental masses.

Gondwana tillites. Although the name Gondwana was chosen by Suess for a hypothetical continental plate stretching from South America to India, it is now commonly used for the much smaller protocontinent envisaged by Du Toit. The name comes from the typical locality in peninsular India of a highly distinctive series of flat-lying strata, about 20,000 feet thick, ranging in age from Late Pennsylvanian to Early Cretaceous (about 290,000,000 to 130,000,000 years old). The basal bed of the series is a thick tillite, indicating extensive and prolonged continental glaciation. Overlying the tillite are thousands of feet of sandstone, shales, and clay ironstones of predominantly continental origin, interbedded with minor marine deposits and thick seams of coal. Capping the sediments are massive flows of basalt. This succession, the Gondwana System, has counterparts in six landmasses of the Southern Hemisphere: Africa, South America, the Falkland Islands, Madagascar, Antarctica, and Australia (Figure 3). In some cases, the source areas of these sediments lay beyond the present continental margins. The overall resemblances of the Gondwana-type sediments demonstrate clearly that seven landmasses, now separated by oceans and distributed over 120° of latitude, had remarkably similar histories from the Late Paleozoic into the Cretaceous. Nothing resembling the Gondwana record has been found in the Northern Hemisphere except in India. While glacial tillites were being deposited at sea level on lands now straddling the Equator, the northern continents were free of continental ice caps. This pattern is so unlikely as to force a choice between continental fixity with extreme irregularity of climate or continental drift with climatic zoning based on latitude. The latter choice seems the more reasonable.

Similar histories among separated landmasses

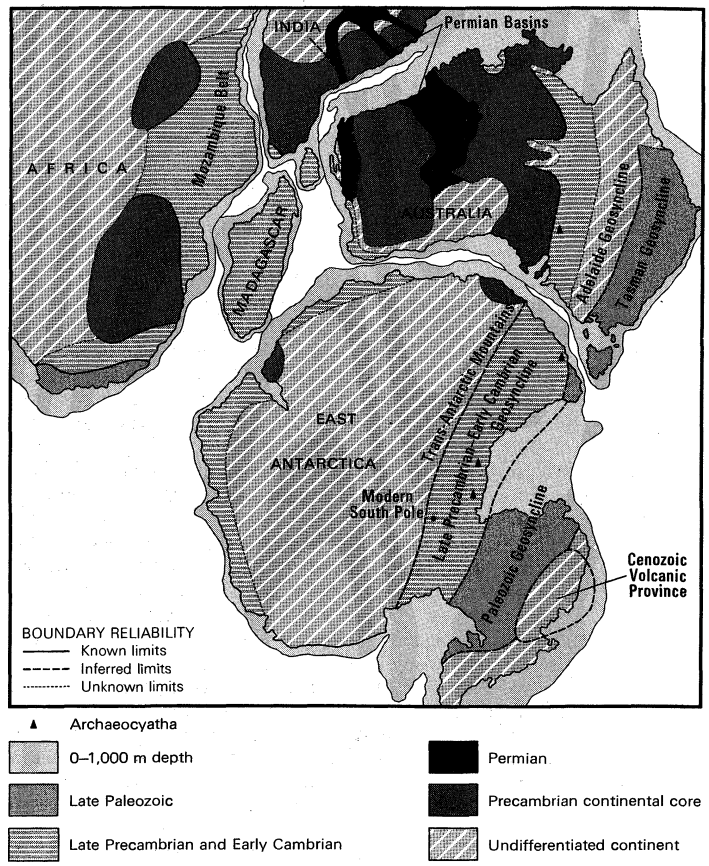


Figure 3: Geological fit of Antarctica, Australia, and Africa, showing structural trends and ages of rocks.

From "The Confirmation of Continental Drift," by Patrick M. Hurley. Copyright © 1968 by Scientific American, Inc. All rights reserved

Many advances have been made in the study of climatic changes (see CLIMATIC CHANGE) based on the distribution of coral reefs, evaporites, red beds, and the temperature-sensitive oxygen isotope ratios in fossil marine shells. Some evidence appears to favour a Paleozoic climatic zoning similar to that of the present. In the end, however, the glacial tillites rank among the most dramatic and unambiguous indicators of paleoclimates, and those of the Permocarboniferous appear most understandable if they are grouped together near the South Pole.

Distribution of fossils. It is well known that the migration of some types of organisms can be prevented by seemingly minor environmental barriers of water, climate, or altitude. Famous examples of specialization due to short-distance marine barriers are the unique fauna of Australia and of several of the Galápagos Islands. A major biological problem arises, therefore, when identical species of plants or animals, strictly adapted to living on land or in freshwater, are found on both sides of very large marine barriers. The observation that lemurs occur in India, Madagascar, and Africa led in the last century to the postulated existence of Lemuria, an ancient landmass linking these three areas. Land bridges became common in geological thought as more and more faunal identities were discovered between widely separated continental masses. Wegener and Du Toit explained these similarities in terms of continental drift. Their arguments have, if anything, been strengthened by recent paleontological discoveries. Meanwhile, ocean-floor topography has become too well known to allow for sunken continental slabs or land bridges.

Similarities among faunas and floras

Glossopteris, a flora including two genera of seed plants totaling 58 species, is associated with the tillites in all the occurrences of the Gondwana formations. The Glos-

sopteris flora apparently thrived in subarctic to temperate conditions. Their present distribution, however, would require them to have crossed vast expanses of open ocean.

Vertebrate fossils are generally too limited in numbers in most Gondwana beds to yield useful evidence of continental relationships. Exceptions include the Permian reptile *Mesosaurus*, a Triassic amphibious labyrinthodont, and Devonian to Cretaceous freshwater fish. *Mesosaurus* was a small reptile adapted to shallow, brackish waters, whose skeletons are found in only two localities: the Early Permian Dwyka (tillite-bearing) Formation in Western Cape Province, South Africa, and 3,000 miles directly across the Atlantic in the similar Irarare Formation of Brazil. A mandible fragment from a Triassic labyrinthodont, discovered in 1968, is the first land vertebrate fossil to be found in Antarctica. It represents an amphibian, four feet long, whose remains also occur in Gondwana beds in Africa, South America, and Australia. Neither *Mesosaurus* nor the labyrinthodont was a marine creature, and without appeal to sunken land bridges it is difficult to account for their distribution other than by continental drift.

The modern freshwater fish of the Old and New Worlds differ substantially, with only 21 percent of North American genera represented in Europe. In the Devonian red beds, however, 57 percent of North American genera are also present in Europe, and a similar degree of identity persists throughout the Paleozoic.

In the Southern Hemisphere a striking correspondence has recently been found between linear geosynclinal troughs in Brazil and Gabon. These troughs, which appear to be segments of a once continuous basin, contain Jurassic to Middle Cretaceous nonmarine formations, including similar hydrocarbon occurrences, several identical species of freshwater fish, and 30 identical species of freshwater ostracods. The earliest marine bed in each trough is an Upper Cretaceous salt deposit that apparently marks the first appearance of the Atlantic Ocean between Brazil and Gabon.

Although these types of evidence support the concept of continental drift, other data on the distribution of plants and animals show anomalies that cannot be accounted for by any reasonable pattern of drifting landmasses. Furthermore, a careful study of the distribution of fossil and modern marine invertebrates by Francis G. Stehli indicates that numerous classes of these invertebrates exhibit much greater diversity of genera in the warm seas near the Equator than they do farther north and south. The diversity gradients for Permian marine invertebrates show a pattern of climatic zoning by latitude similar to that of today. Although this evidence does not contradict the idea that a westward drift of the Americas or a spreading of the Atlantic has taken place, it runs counter to the evidence of polar wandering and latitudinal migration of continents as deduced from paleomagnetic measurements.

Paleomagnetic data. The first indications that the continental drift hypothesis might be revived and re-examined came in the 1950s as a result of remanent magnetism measured in rocks by researchers in England and Japan. Paleomagnetic studies have produced three lines of evidence of the utmost importance to continental drift and plate tectonics: evidence of polar wandering, evidence of continental displacements and rotation, and evidence of reversals of the geomagnetic field.

Measurements of the fossil magnetism in unweathered rock samples from all continents show that for the past 20,000,000 years the locations of the Earth's magnetic poles have remained essentially unchanged.

Polar wandering. According to data assembled by S.K. Runcorn and his co-workers at Newcastle-on-Tyne, England, the position of the north magnetic pole moved at least 21,000 kilometres over a long curving path from western North America in the late Precambrian, through the present Pacific Ocean to northern Asia in the Mesozoic, and to the Arctic in the mid-Tertiary.

If rocks of the same age in all continents were to point to the same pole throughout a long migration, this phenomenon could be ascribed to polar wandering, as though an outer shell of the Earth, including the crust and part of the mantle, had become decoupled and moved as a unit over the interior. The actual evidence from rock magnetism, however, permits no such simple solution.

Rocks older than 20,000,000 years from different continents point to different pole positions and describe systematically divergent paths of migration. This suggests that landmasses that were once joined have rifted and moved relative to one another and that their former positions can be reconstructed by superimposing their polar-migration curves. Many investigations have shown that continental reconstructions based on paleomagnetism coincide well with the requirements of the fit of continental blocks and with the pattern of paleoclimates.

An alternative interpretation of the data amassed from paleomagnetic measurements around the world has been presented by P.M.S. Blackett and his associates, who conclude that all the continents have drifted northward since the beginning of the Paleozoic and that most of them have also undergone some clockwise or counterclockwise rotation. Paleomagnetic measurements give no evidence as to longitudinal positions of samples and therefore cannot be used to trace eastward or westward motions if they occur. With respect to northward migration, the paleomagnetic data from India are of special interest for they definitely suggest that India lay in the southern hemisphere 150,000,000 years ago and has since drifted northward across the Equator. This interpretation of the data coincides with the stratigraphic and paleontological evidence that India was formerly part of the Gondwana landmass.

Those scientists who remain skeptical of geological interpretations based on paleomagnetism argue that the sampling is inadequate, the results are inconclusive, and the basic assumption of a permanently dipolar magnetic field is unjustified. However, more measurements are made each year and the method as a whole appears increasingly reliable.

Oceanic data. Evidence for continental drift has been derived mainly from the continents. During the past decade, however, new support has come from studies of the ocean basins. In comparison with the continents, which include rocks of all ages back to about 3,500,000,000 years, the ocean floors are geologically young.

Marine sediments and sea-floor spreading. The oldest marine sediments (*q.v.*) in dredge or core samples collected from any ocean are Jurassic, about 160,000,000 years old. The total thickness of deep-ocean sediments, red clay, and *Globigerina* ooze (largely composed of the remains of one-celled organisms) is everywhere surprisingly little, averaging about 400 metres in the Pacific and 500 metres in the Atlantic. At a mean rate of deposition of about 5 millimetres per 1,000 years, the sediments in both oceans could have accumulated in less than 200,000,000 years, beginning in the Triassic.

Marine sediment less than 200,000,000 years old

Deep-sea drilling near oceanic ridges (*q.v.*) has shown that the ages of sediments immediately overlying the bedrock in the South Atlantic increase linearly with distance away from either side of the mid-ocean ridge crest. Sediments deposited over the ridge when spreading began,

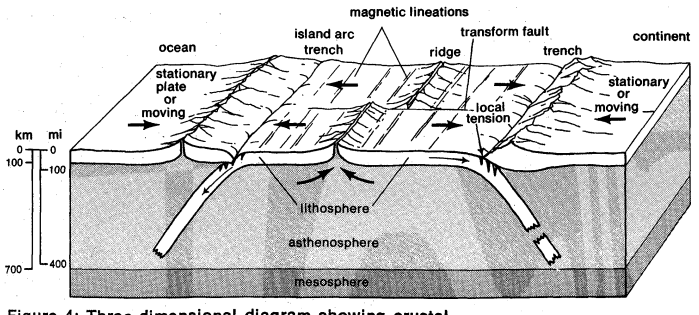


Figure 4: Three-dimensional diagram showing crustal generation and destruction according to the plate-tectonics principle.

and the Atlantic was a narrow rift, now lie near the continental shelves. The separation of these sediments, as dated by microfossils, suggests a sea-floor spreading rate of about two centimetres per year.

According to the sea-floor spreading hypothesis, the ocean floor is a thin rind on the upper mantle. The mechanism invokes the slow creep of hot crystalline materials: new rock rises in thin dikelike bodies up the ridges, cools, and then spreads horizontally away on both sides. A downward flow is assumed to occur where the ocean floor plunges into the mantle beneath continental blocks or along the linear trench systems that lie seaward from the island arcs. By this mechanism the oceanic bedrock is constantly renewed and the cover of sediments kept thin (Figure 4).

Given this type of movement in the upper mantle, the sialic continental blocks will presumably be floated passively away from the flanks of active ridges and centred over or between descending currents. The Red Sea is believed to be a new rift opening, with generation of ocean floor occurring at about three centimetres per year. The estimated rates of sea-floor spreading, based on the continental separation across the Atlantic and other ridges, range from 1 to 10 cm per year.

Magnetic anomalies. In 1963, linear magnetic anomalies symmetrically distributed in alternate strips along either side of oceanic ridge crests were reported by F.J. Vine and D.H. Matthews, English geophysicists, who suggested that the anomalies might record reversals of the Earth's geomagnetic field. Since that time it has been established that volcanic rocks on the continents record two main epochs of normal (north-seeking) polarity alternating with two of reversed polarity in the past 4,000,000 years. Oceanic rocks should respond to the same reversals, and if it is assumed that these rocks acquire their magnetism as they cool on the ridge crests, the widths of the anomaly belts indicate how far oceanic rocks have spread laterally during the past 1,000,000 to 4,000,000 years and, by extrapolation, over much longer periods. The rates of spreading, obtained by correlating oceanic and continental rocks of the same polarities and assumed ages, are identical to those that have been estimated from crustal separations: one to ten centimetres per year.

PLATE TECTONICS

The concept of sea-floor spreading led to a revolution in Earth science. Given this concept, many new data and old puzzles began to fall into place, and geologists were forced to review the entire framework of global tectonics in terms of large-scale horizontal crustal movements. Within a very short time, geological and geophysical evidence for these crustal movements had been incorporated into a comprehensive system called plate tectonics.

Plate tectonics refers crustal movements to a spherical surface and analyzes the motions due to sea-floor spreading and transform faulting on a globe of constant area. The Earth's surface (lithosphere) is conceived of as being divided into a few rigid plates, each of which is bounded by three kinds of surfaces: a ridge or rise (zone of divergence), where new crust is generated; a trench or folded mountain range (zone of convergence), where crust is shortened or destroyed; and transform fault planes, along which relative movement between plates occurs. The motion of each plate over the mantle and relative to adjacent plates is seen as a rotation about an axis. Plate margins are all zones of seismic activity, and with some exceptions, notably within the North American and Asiatic plates, they account for most of the earthquakes of the world. The rises and fault planes are characterized by the occurrence of shallow-focus earthquakes and are also delineated by their topography and magnetic anomalies. The trench systems and other lines of convergence are the loci of deep-focus as well as shallow seismicity.

Earth's surface conceived of as rigid plates

The present configuration of linear seismic zones divides the Earth's crust into at least six moving plates: two are wholly oceanic, and four consist of continental and oceanic crust coupled together (Figure 5). The assumption that continental rock possesses sufficient strength to permit the movement of massive blocks has always been fundamental to continental drift. Oceanic crust is too thin (only about five kilometres) to possess such strength. In the new global tectonics, therefore, the lithosphere is defined as a layer of significant strength, about 70 kilometres thick, including the crust and uppermost mantle. Beneath the lithosphere lies the asthenosphere, a solid layer close to the melting point and having no ef-

From "The Origin of the Oceans," by Sir Edward Bullard. Copyright © 1969 by Scientific American, Inc. All rights reserved

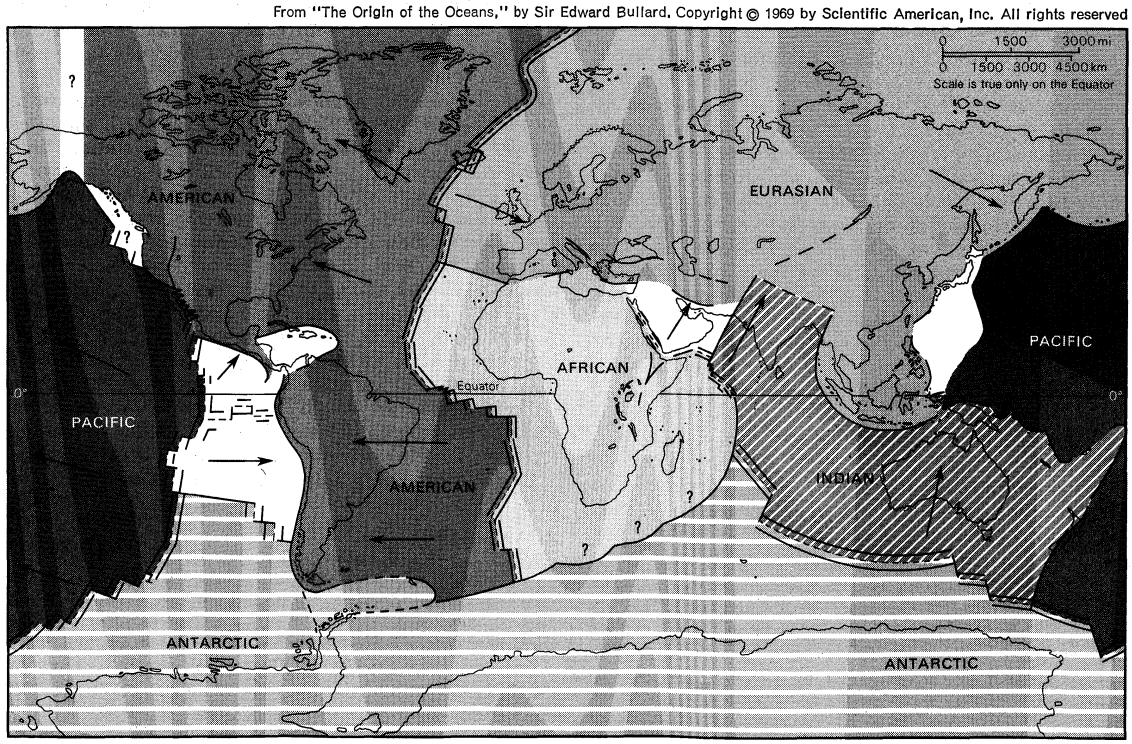


Figure 5: Relative motion of the six major plates or crustal segments today. The African plate is assumed to be stationary.

fective strength against long-term stress. The asthenosphere is not a homogeneous shell but includes at least three layers of increasing density and seismic velocity. It extends for some 700 kilometres down to the mesosphere, the inert lower mantle that is unaffected by tectonic movements.

The oceanic ridges have already been described as tensional zones where new crustal material is formed. Compensating for this crustal extension, the lithosphere must elsewhere be shortened by compression. Crustal shortening within continents is accomplished by folding and overthrusting in compressional mountain ranges. Within the ocean basins it occurs at plate margins where the oceanic lithosphere is thrust downward into the asthenosphere. This appears to be happening along the entire Pacific margin, where earthquake foci are distributed in planes dipping about 45° beneath the American continents, the Aleutian arc, and the system of trenches and island arcs rimming eastern Asia.

The underthrusting of thick slabs of cold lithosphere into a weaker, hotter substratum creates the Earth's zones of greatest tectonic instability. Shallow-focus earthquakes are frequent within the upper portions of the dipping slabs, and the Earth's only deep-focus quakes with foci down to 700 kilometres occur in the lower parts of the slabs. The landward side of the deeper-earthquake zones is marked by abnormally high heat flow and intense volcanism, of predominantly andesitic rather than basaltic petrology. Although the Pacific margins are fundamentally zones of compression, they bear superficial evidences of tension. Topographically the deep trenches resemble tensional rifts, but they are structurally controlled by the compressional furrow at the fault contact between two slabs of lithosphere.

Tension fractures are predicted along the bend in the oceanic slab where it begins to dip downward. These fractures may be filled with marine sediments that will be carried into the asthenosphere, heated, and mixed with mantle rock to produce silica-enriched hybrid magmas. Some petrologists ascribe the andesitic volcanics of the Pacific margin to this process; others believe andesites are primary differentiates from the mantle. Estimates of crustal shortening along the trenches, based on measurements of energy released in earthquakes, range from 1 to 15 centimetres per year.

THE PROBLEM OF MECHANISM

All variations of the concept of continental drift, from Wegener's hypothesis to plate tectonics, require a driving mechanism to generate the displacements indicated by geological and geophysical observations. No mechanism that is both theoretically and observationally sound, and is therefore generally acceptable to geoscientists, has been postulated to date. This lack of a sufficient mechanism was the reason for the wholesale rejection of the concept of continental drift by geophysicists in the 1920s and for the reluctance of some of them to accept the evidence for plate tectonics today. Suggested mechanisms that have been abandoned as inoperative or too ineffective to cause horizontal motions of the required sense or rates include global contraction, global expansion, tidal forces, gravity sliding, and a decrease in the gravitational constant.

Possibility of thermally driven mechanism

Today, proponents of plate tectonics are largely persuaded that the mechanism is some form of thermally driven motion in the upper mantle. Several lines of evidence, however, suggest that the mantle is too solid, elastic, and strong to allow for significant dynamic flow. The magnitude of Earth tides and the characteristic transmission of transverse seismic waves show that the effective rigidity of the mantle, at least in its deeper parts, is greater than that of steel. The figure of the Earth, out of equilibrium for a viscous medium with its present rate of rotation, is evidently maintained by the strength of the mantle.

There is increasing evidence that the upper mantle is markedly inhomogeneous and therefore not subject to mixing by convection. Gravity and heat-flow measurements are, on the average, equal over continents and ocean basins, despite the fact that continental rock is lighter and also at least six times richer in heat-producing radioactive elements than are the oceanic basalts. These determinations suggest that continents are balanced by deep sialic roots and, deeper still, are underlain by columns of mantle rock that are radically depleted in radioactivity as compared to oceanic mantle. All these indications militate against horizontal flow and work in favour of the formulation supported by some geophysicists, namely, that once rocky material is located on a given Earth radius, it will always remain in the vicinity of that radius.

Despite these problems the evidence for large-scale horizontal movements of crustal plates is so impressive that numerous geophysicists have attempted to construct plausible mechanisms of upper mantle movement within the framework of the solid Earth. Since advocates of plate tectonics regard the rigid, moving plates of lithosphere as 100 kilometres thick, they abandon the much shallower and markedly undulating Mohorovičić discontinuity (zone at the base of the crust marked by a change of velocity of seismic waves) as a site of relative motion. They have also abandoned the strict distinction between continental and oceanic lithosphere and so have reduced the contrast in bulk chemistry. As a result, Wegener's iceberg model of floating continents and the conveyor-belt analogy of sea-floor spreading are both replaced by the conception of thick plates, likened to paving blocks jostling one another at different rates along shifting boundaries.

Orderly convection cells cannot be delineated with reference to present plate boundaries; in any case, the concept of convection as known in liquids is evidently inapplicable to the mantle, some layers of which respond as elastic solids to short-term stress but as fluids to long-term stress. The ultimate driving mechanism must therefore be some variety of thermally produced solid-state creep for which the mathematical model has yet to be constructed.

Measuring crustal motion

The large-scale motions associated with plate tectonics are inferred from geological and geophysical phenomena such as the separation of matching coastlines, the separation of linear magnetic anomalies, and the release of energy in seismic zones. The estimated magnitude of these motions ranges from 1 to 15 centimetres per year, which correlates well with geodetic measurements of 5 centimetres per year along the San Andreas fault zone of California.

In the future it may become possible to measure crustal motion by measurements from artificial earth satellites. At present, the locations of fundamental stations in a worldwide network of optical satellite-tracking stations are known to an accuracy of ± 5 metres. When lasers replace cameras, the accuracy can theoretically be increased to an ultimate limit of 3 centimetres. Before measurements of this magnitude become practical, many problems must be solved with respect to the shape and effect of the Earth's gravity field and of the atmosphere. Nevertheless, within a decade or so it should be technologically feasible to detect changes in the relative positions of stations located on crustal plates that are moving ten centimetres per year. If measurements can be continued for a number of decades, it should finally be possible to derive a worldwide picture of relative crustal motion and, if it is occurring, to prove the reality of continental drift.

BIBLIOGRAPHY

Symposia: R.A. PHINNEY (ed.), *The History of the Earth's Crust* (1968), excellent, up-to-date volume with chapters by many scientists, including Vine, Stehli, Hurley and Rand, and Bullard, with sections on the upper mantle, sea-floor spreading, and the continental crust; P.M.S. BLACKETT (ed.), *A Symposium on Continental Drift* (1965), includes the computer reconstruction map by Bullard *et al.*, much data on paleomagnetism, oceanography, and transcurrent faults; work, however, was published before the development of our most significant evidence for crustal movements; A.E.M. NAIRN (ed.), *Problems in Paleoclimatology* (1964), essential background

on a wide variety of approaches to paleoclimatology; G.D. GARLAND (ed.), *Continental Drift* (1966), a short review of current ideas on drift followed by evidence from the Arctic and eastern seaboard of Canada; S.K. RUNCORN (ed.), *Continental Drift* (1962), includes much data on paleomagnetism and related subjects and an early article by Dietz on ocean-basin evolution by sea-floor spreading; T.F. GASKELL (ed.), *The Earth's Mantle* (1967), good collection of articles on the geophysics and geochemistry of the mantle.

Books (classics): A.L. WEGENER, *Die Entstehung der Kontinente und Ozeane*, 4th ed. (1929; Eng. trans., *The Origins of Continents and Oceans*, 1966), a clear exposition of ideas and principles, although geodetic data near end of book not to be taken seriously; A.L. DU TOIT, *Our Wandering Continents* (1937), excellent historical review of ideas on continental drift up to 1936, followed by much original work on Gondwanaland based in part on field experience; E. SUESS, *Das Antlitz der Erde*, 5 vol. (1883–1909), source book for Suess's ideas on Gondwanaland and foundering of ocean floors, plus monumental description of global geology; W.A.J.M. VAN WATERSCHOOT VAN DER GRACHT *et al.* (eds.), *Theory of Continental Drift* (1928), interesting chapters by leading geologists; of historical importance because this symposium killed the drift hypothesis for some 25 years.

(U.B.M.)

Continental Shelf and Slope

The most regularly developed major feature of the Earth's topography is the continental terrace, also known as the continental margin. The upper surface is formed by the continental shelf, a broad shallow strip of seabed that extends from the coast to depths of 100–200 metres (330–660 feet). Some authors include the coastal plain on the landward side, but the shore represents a reasonable limit. The breadth of the shelf and the depth of its outer margin vary greatly. Beyond the shelf is the continental slope, a much steeper zone that merges with the deep-sea floor at a depth of about 4,000 to 5,000 metres (13,000–16,000 feet). In addition to its great variation in steepness and smoothness, the slope is diversified in some areas by terraces or by basins and ridges. The gradient of the continental slope tends to decrease over its lower half. This part is called the continental rise and is considered to be part of the deep-ocean floor. In some places, the continental slopes extend down to deep-sea trenches that parallel the coast. The nomenclature associated with the shelf and slope and the general dimensions are given in Figure 1.

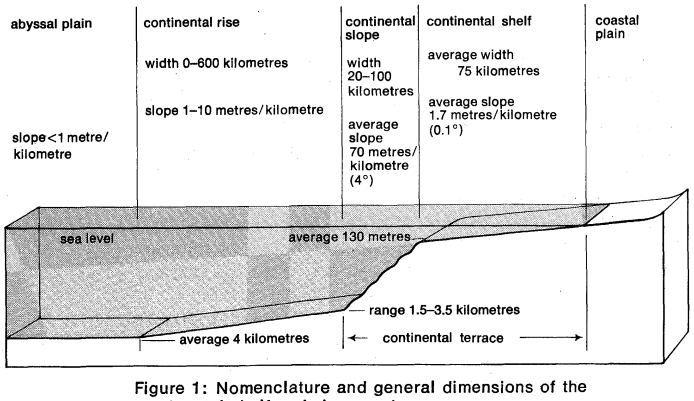


Figure 1: Nomenclature and general dimensions of the continental shelf and slope regions.

There are few reaches of coastline facing the ocean or deep inland seas where the continental shelf and slope is entirely absent. Even most islands are surrounded by similar, but smaller, flat-topped features called island terraces. The continental shelf and slope each cover approximately 8 percent of the total sea floor. That surrounding the U.S. has an area that is more than one-third that of the land surface of the country.

The most spectacular result of detailed echo sounding of the continental margins has been to reveal the existence of many huge, steep-walled valleys that gash the terrace, namely the submarine canyons (see CANYONS, SUBMARINE). Many canyons debouch on a delta-like feature, which is a gently sloping submarine fan. The outer margin of the shelf sustains islands or shallow banks. The most extreme examples occur where coral reefs have developed barrier reefs and atoll rims (see CORAL ISLANDS, CORAL REEFS, AND ATOLLS).

Economic value of the shelf

The economic value of the continental shelf is varied. Fisheries have vast potential and tin ore has been dredged for many years from submerged stream valleys of the Sunda Sea. Recently the search for oil and gas has intensified. Coal, salt, phosphate, sulfur, manganese, clay gravel and sand, diamonds, gold, silver, platinum, iron, titanium, chromium, and rare earths are other valuable products that the shelf can provide (see OCEANS AND SEAS).

The approaches to harbours traverse the shelf, and navigation always has been based on submarine topography and bottom sediment of this shallow area to assist in orientation and to help avoid dangerous shallows. Hydrography therefore has been almost entirely limited to the study of the shelves and coasts. Detailed charts now exist of practically all shelf areas. Although the advent of echo sounding has greatly speeded the survey of the continental slopes, these still are poorly explored.

CONTINENTAL MARGINS: DEFINITIONS AND BATHYMETRY

Underwater topography is revealed by sounding. Until half a century ago sounding was carried out with rope or wire and the horizontal location was determined from landmarks. Later, echo sounding provided a much faster method and one that was more accurate because the position of a ship can be more exactly determined when it remains underway than when it is stopped and left to drift each time a sounding is taken. A depth recording is relatively useless if the location is not known, and the accuracy of surveying is as dependent on precise positioning as on the reliability of the actual depth determinations. Since World War II, position finding has been greatly improved, first by various radar systems that are useful near shore and quite recently over the entire world ocean by satellite methods. Echo sounding now has attained the stage of being more accurate than wire sounding in all depths, and it provides a continuous record along the line of travel.

The continental shelf. The shelf begins at the low-water line, but there is uncertainty whether to include estuaries (*q.v.*), tidal flats, or shallow lagoons (*q.v.*). There is less reason to exclude barrier-reef lagoons, and it can be argued that atoll lagoons and banks of the same depth as shelves also are part of the system (*e.g.*, the Bahama Banks). A similar problem is whether shallow inland seas should also be incorporated. The difficulty is that there are many transitions between enclosed seas such as Hudson Bay, and shelves with a few minor islands such as the Farilhões off Portugal or Sable Island opposite Nova Scotia.

The outer margin of the shelf formerly was defined as the 100-fathom or 200-metre line. A much better limit now can be set by the break in slope, or shelf break, between the nearly horizontal platform and the much steeper continental slope, which nearly always starts abruptly. The depth at this break averages 130 metres (430 feet), but it may be less than 40 metres (130 feet) or as much as 500 metres (1,600 feet). There are areas, however, where no distinct break in slope exists and the 200-metre line must be adopted for a limit.

Shelf topography, valleys, and river deltas

The average slope is about one-tenth of a degree; near shore it normally is slightly steeper. The average width is 75 kilometres (45 miles). There is practically no shelf off Miami, Florida, or the French Riviera, but, in contrast, the Atlantic shelf off Patagonia is 500 kilometres (300 miles) wide. The surface of the shelf may be remarkably smooth or may consist of rocky or sandy shoals, depressions, or channels, one or two terrace-like features, and a gradual or abrupt longitudinal slope. In fact, the topography of continental shelves is at least as diversified as that of coastal plains.

Valleys on the shelf are of three distinct types. Some are broad with a fairly flat bottom and occur only in high latitudes. These evidently were scoured out by Pleistocene

glaciers. Some valleys actually form the continuation of fjords onto the shelf. Others are more isolated, like the huge Cabot Strait Trough between Nova Scotia and Newfoundland, which is 500 metres (1,600 feet) deep and almost 100 kilometres (60 miles) wide. There also are many instances of glacial troughs parallel to the coast, exemplified by the trench curving around the southern extremity of Norway, which is up to 800 metres (2,600 feet) deep, 800 kilometres (500 miles) long and 80 kilometres (50 miles) wide.

A different kind of valley results from the action of tidal scour. It occurs when restrictions force tidal currents to sweep with extra force over the seabed—as between the Frisian Islands of the southeastern North Sea, for example.

The third valley type results from the drowning of a river valley. The Hudson Channel, extending from New York harbour to the outer margin of the shelf, and the "Sunda River," a dendritic system of valleys between Sumatra, Malay Peninsula, and Borneo, are examples of this phenomenon. No sharp distinction can be made between valleys and oblong depressions of tectonic origin that may be modified by external processes.

River deltas (*q.v.*) form on the shelf. Some, like the Rhône Delta, appear to pinch out on top of the shelf. Other deltas, such as those of the Nile and Niger rivers, have built out the continental slope so that the terrace locally bulges seaward. The Mississippi Delta is a multiple feature. The present bird-foot set of distributaries has built out onto the shelf and has reached the edge of the continental slope at one point. Similar structures of greater age overlap each other and have disappeared by subsidence and marine attack. Thin sediments have contributed to the construction of the continental terrace off Louisiana.

The great South American rivers, notably the Paraná and Amazon, have large estuaries and do not appear to have extended the shelf seaward. The same lack of outbuilding characterizes the Orinoco Delta, which is situated partly behind the island of Trinidad. This may be due to strong tidal currents and the preponderance of fine sediment that is washed away by currents, partly in a lateral direction along the coast and partly into the deep sea. Many great rivers such as the Irrawaddy and Indus appear to have filled sinking trenches but have not extended the land onto the terrace. Others, including the Euphrates and Po, have built deltas with sufficient rapidity to prevent inundation by the sea due to land subsidence.

The actions of water waves and ocean currents (*qq.v.*) have combined to throw up sandy barriers on some shelves that are approximately parallel to the coast and separated from the mainland by a lagoon. One of the most spectacular examples is Pamlico Sound, where Cape Hatteras extends out into the Atlantic halfway across the shelf. Sedimentation and plant growth will reclaim these lagoons, thus adding large areas to the land. In tidal flats the struggle between destructive and constructive forces is in full swing.

The continental slope. The continental slope is the declivity beyond the outer edge of the shelf; it extends to great depths, to the ocean basins (*q.v.*). The gradient varies between 1° and near vertical, but the average is just over 4° for the first 2,000 metres (6,600 feet). Most slopes are steepest at the top but a few are precipitous in the deeper portion. The total length of the continental slope is 300,000 kilometres (200,000 miles), almost eight times the circumference of the Earth. The junction with the deep-sea floor or with a continental rise on the floor is usually indistinct.

In addition to being deeply indented by the numerous submarine canyons, the slope is diversified by the presence of broad plateaus and other features. The Blake Plateau, at a depth of about 800 metres (2,600 feet) off the southeastern corner of the U.S., is an example of a major submarine plateau. Elsewhere, mounds and dents cause irregularities in the topography. The most typical continental borderland occurs off southern California, where a dozen basins of about 1,000 square kilometres

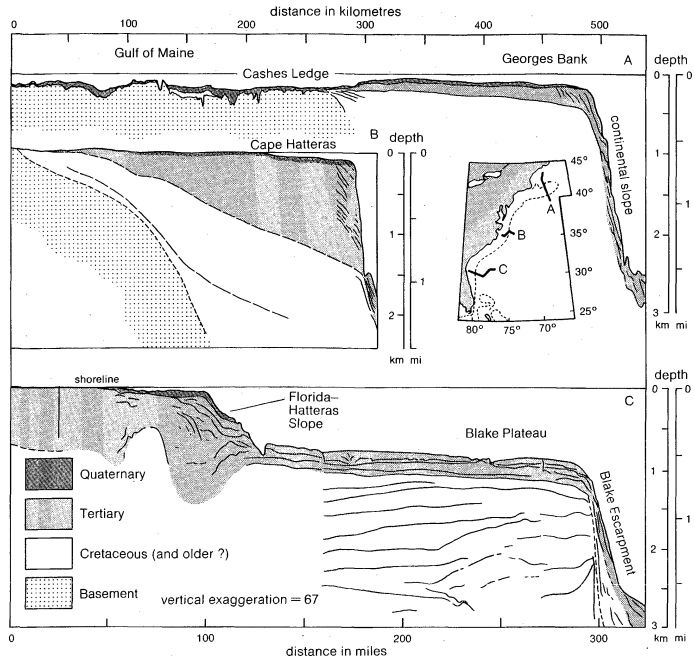


Figure 2: Three structure sections of the east coast terrace of the U.S. (inset), based on (A) dredging, (B) well data, and (C) seismic data.
By courtesy of Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

(400 square miles) each are separated by submarine ranges, some topped by islands.

COMPOSITION

No model of the Earth's crust and its development that does not account for the features of the continental margins between continents and ocean basins can be satisfactory. Many ideas on the subject have been offered, but until the middle of this century these remained speculative because of the lack of facts.

Bottom sampling and data collection

The two main sources of information about the terrace, aside from soundings, are bottom sampling and the application of geophysical techniques. The old method of sampling employed tallow on the sounding lead, which made it possible for the mariner to determine whether the bottom consisted of rock, mud, sand, gravel, or shells. For more sophisticated investigations, complete samples must be obtained for study in the laboratory. Core samples from the sea floor are even more desirable, but they are difficult to obtain except in fine-grained deposits. The available data, however, are now sufficient to indicate the nature of the principal types of deposits and environments.

A severe limitation inherent in bottom sampling concerns the thickness that is sampled. It can reveal only surficial aspects of the terrace and the nature of its composition. The obvious remedy is to drill into the terrace. This was actually done long before World War II at Cape Hatteras; scientifically, it was the most profitable deep well ever sampled (Figure 2). It showed that the terrace had been built up in shallow water on a continuously subsiding foundation of weathered granite that sank to a depth of more than 3,000 metres (10,000 feet) in the course of about 100,000,000 years. Based on this information and the geology of adjacent land areas, it was then possible to interpret the geophysical evidence with some confidence.

Geophysical testing with the use of sound waves forms an invaluable counterpart to all forms of sampling. Dependent on the method used, the penetration of the returned vibrations can go right through the Earth's crust to the mantle beneath or any lesser distance within the crust.

The two main objectives of this work are to determine the speed of wave propagation, which provides information on rock composition, and to determine reflecting

horizons in the crust, which reveal stratification and buried surfaces. Folding, crumpling, faulting, wedging, and complicated distortions and intrusions can be identified and accurately analyzed (Figure 2).

Early measurements were widely spaced, but today the interval is so short for some types of testing that continuous records are obtained. After processing, a relatively complete section of the seabed structure along the ship's course is obtained. If samples that provide the rock type also are available, knowledge of the continental terrace is considerably enhanced.

An additional source of information is underwater photography, which has provided valuable data on sediment types and current systems on the terrace. Some of these will now be discussed.

Beach, shelf, slope, and deep-sea deposits

Sediments and current systems. Depressions tend to collect fine material with a high percentage of organic matter, whereas topographic highs are coarser grained or consist of rocks and reefs. Near large rivers the sediment is fine, especially on the side to which dominant transport is directed. The northwest side of the Amazon, for example, receives so great a supply of mud near shore that serious trouble is experienced by shipping as far away as Guyana. The bulk of this mud is deposited along the coasts or fed into the Caribbean Basin. Calcareous ($CaCO_3$) matter is frequently found as a shelly admixture with other materials. Pure calcareous deposits are encountered mainly in low latitudes, especially around coral reefs. Off formerly glaciated coasts, gravel and boulders are common.

Most beaches are bare rock, cobbles, or sand, although in protected sites mud can accumulate (mangroves). With distance from the beach and deepening of the water, the grain size decreases to fine sand or mud. But beyond this point, often 10 to 20 kilometres (6 to 12 miles) from shore, medium to coarse sand or patches of gravel form the seabed over wide areas, and in many cases no mud is present.

Narrow shelves usually are underlain by hard rock, with or without a thin veneer of unconsolidated sediments. The more exposed the bottom is to currents and wave action, the coarser the deposits tend to be. Because longer waves reach deeper than short ones, ocean swell is more able to stir the bottom at depths of a few dozen metres than are steep storm waves. In tidal scour channels the bottom consists of bare rock or a washed-out deposit of coarse material.

Most of the continental slope is covered by fine mud, consisting of clay and silt with a variable admixture of volcanic ash or shelly materials (see MARINE SEDIMENTS). But in areas where the declivity exceeds about 10°, the underlying rock bottom normally is exposed. Either hard rock (granite, basalt, consolidated sediment, reef rock, etc.) or semiconsolidated terrace sediments are most common.

On the continental rise and particularly on subsea fans, large corers often cut through sandy beds that are interstratified with normal deep-sea deposits. The structures and components show that these beds have been deposited by turbidity currents. These density currents (*q.v.*) clean out the canyons and sweep the slopes at intervals of centuries or thousands of years. Within the confines of submarine canyons the deposits usually are coarse-grained. It is understandable that typical turbidity current deposits are missing there, because such flows are too mobile and competent to lose sediment in a steep gorge. The deposits that do occur represent the remainder of abortive turbidity currents—that is to say slides—and creeping beds modified by bottom currents that move independent of any sedimentary load.

Sea-level fluctuation

Influence of the Ice Age. During the Pleistocene Epoch (*q.v.*) sea level was lowered considerably during each major advance of the ice caps; the last such lowering amounted to about 100 metres (330 feet). Ice advances occurred about a half a dozen times in the course of the last 2,500,000 years. During intervening interglacial intervals sea level was equal to or higher than its present elevation.

The influence on the shelf has been dramatic. The breaker zone and its rough turbulence has passed down the shelf surface and back again, disturbing the deposits and winnowing out the finer sediments. Exposure of the shelf by the lowering of sea level resulted in weathering (*q.v.*) and overall degradation in some areas and deltaic deposition by minor rivers elsewhere. The larger rivers flowed down to the new coast on much steeper gradients than are usual in their unaffected lower courses. This tended to cause erosion of gorges on the exposed shelf. Vast amounts of sediment were brought to the outer edge of the shelf as a consequence of this process. The deposited material served partly to prograde the shelf; it also passed down submarine canyons and greatly stimulated the action of turbidity currents. Most estuaries (*q.v.*) are the result of postglacial submergence of river valleys scoured out during stages of lowered sea level. They serve as sediment traps today and hence reduce present deposition on the shelves.

Since the end of the last glacial interval sea level has risen 100 to 120 metres (330 to 390 feet) in about 12,000 years, or about one centimetre per year. Over the last 5,000 years sea level has fluctuated slightly around its present stand (elevation).

Sea level was depressed approximately by the same amount over the entire globe, and the lowering was at least of the same order of magnitude for the major ice advances. The growth and waning of the ice sheets and glaciers (*q.v.*) was irregular, and periods of stagnation or temporary reversal occurred. As a result, sea level sometimes remained almost stationary for longer periods, and this resulted in the cutting of terraces that are still preserved at various altitudes above and below sea level. But worldwide correlation is rendered difficult by local crustal movements.

Although the reality of all these effects of the Ice Age has been demonstrated, the configuration of the continental terraces before the Ice Age is not yet clear. At present, there is a thin veneer of shallow-water and reworked beach deposits between 150-metre (490-foot) depth and the shoreline (Figure 3). This cover rests on a lower veneer of earlier marine deposits or lowland formations (fluviatile deposits, marshes, dunes, boulder clay, etc.) of Pleistocene age that have been partly altered by weathering. This in turn is underlain by undisturbed preglacial shelf deposits. It is not unlikely that this picture is more complex where similar remnants of the earlier Pleistocene low levels have been partially preserved. A third veneer is now forming on the sea floor, consisting of sand (partly of organic origin) in very shallow water and mud farther offshore, out to a distance of a few dozen kilometres.

Beyond this recent covering deposit, the rock or coarse-grained Pleistocene formations have not been buried. These barren stretches of the shelf evidently are unsuitable for mud accumulation. The reason must be the action of incoming waves (long swell) and currents (tidal),

From Rhodes Fairbridge (ed.), *Encyclopedia of Oceanography;* © 1966 by Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company

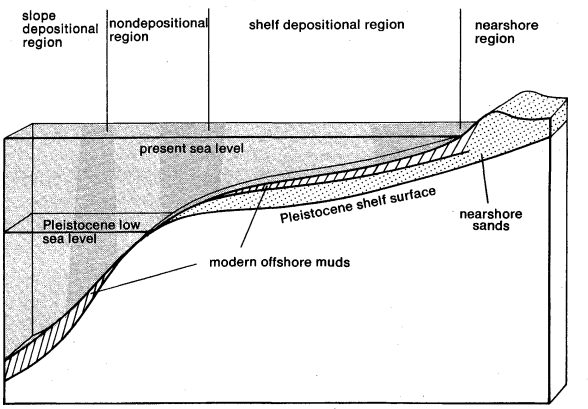


Figure 3: Idealized model of sediment deposition, showing effects of sea-level rise since the Pleistocene on distribution of surface sediments.

two factors that presumably act more energetically near the slope break but are damped off towards the coast. Not until the surf zone is approached does wave energy on the seabed increase again.

A complicating factor is the presence of bottom-dwelling animals—worms, bivalves, sea urchins, and others—that filter clay from the water and add it to the bottom as pellets. By burrowing and ploughing they also mix the sand with some fine-grained material deposited during periods of quiet.

STRUCTURE AND ORIGIN

It is obvious that the two main processes fashioning the shelf are deposition and erosion. On cliffed coasts wave action and currents are visibly making inroads on the terrestrial material. The retreat of an exposed coastline that is composed of unconsolidated sediments can be many metres per year, whereas in hard rock it is so slow that rates of retreat are very slight.

It once was assumed that waves eroded to depths of 100 metres, but it is now well established that they cannot attack hard rock at depths below a dozen metres. They can prevent sediment from accumulating at much greater depths, however. Because a wide, shallow platform of rock will effectively damp and absorb wave energy, marine erosion will have more influence when sea level changes with relation to the land than under stable conditions. The contribution from sedimentation is most readily demonstrated where the shelf bordering a delta has been built out into the deep-sea realm (*e.g.*, the Niger Delta).

Other possible origins are the subaerial development of a coastal plain, followed by subsidence beneath the sea. The elevation of a deeper zone to produce a shelf is less likely. It also has been postulated that the continental margin is due to down flexure beyond a hinge line that must evidently coincide with the break in slope. The precipitous slope south of the Riviera and its dendritic system of steep rocky valleys is considered to be an example. Most geologists assume that the location and shape of the break in slope was caused by the lowering of glacial sea level.

Terrace structures

Geophysical evidence. Geophysical investigations have demonstrated a variety of internal structures. Shelves bordered on the oceanic side by hard rock are thought to be infilled lagoons behind barrier reefs, rocky basins containing accumulated sediment, or an erosional platform that has been veneered with sediment except at its rim.

The first seismic reconnaissance of terrace structure suggested that bedding planes tended to bend upward under the slope of the large Atlantic terrace of the United States. This could be interpreted in terms of basin structure below the shelf combined with general subsidence. Later work has cast doubt on this view. The whole crust beneath the continental margin is in the process of warping down along the Gulf coast. Deep wells have demonstrated that Tertiary and Late Mesozoic sediments have accumulated to a thickness of at least 10,000 metres (33,000 feet), mainly by upbuilding and partly by outbuilding into the Gulf. The base of this great prism of sediment has been depressed at least three times as deep as the present depth of the Gulf.

The amount of waste produced by coastal erosion is estimated at only 1 percent of the fluviatile contribution to marine sedimentation. This accords very well with the finding that the dominant terrace structure is constructional and is the result of upbuilding combined with outbuilding. Details of the structure of many sections suggest the following development. Subsidence rapidly occurs over the area of the shelf and presumably part of the slope. Then, during the following period of crustal stability, sedimentation builds out from the landward side with a relatively steep dip slope, in some cases continuing beyond the break in slope and thus extending the slope into deep water. The cycle is then repeated by the onset of a new subsidence. The subsidence is attributable in part to the weight of the new sediment, which causes compaction of buried deposits and downwarping of the

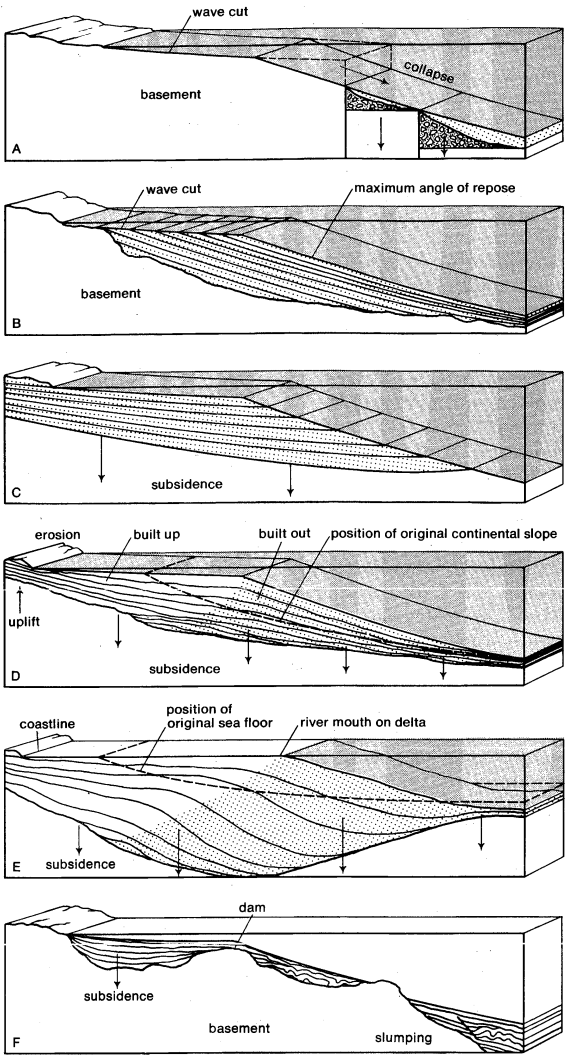


Figure 4: *Various types of terrace structures.* Structures resulting from (A) faulting and marine erosion, (B) outbuilding, (C) upbuilding in subsiding area, (D) combination of (B) and (C), (E) the same, but with great delta building by large river, and (F) tectonic dams and basins in a region of slumping.
From K.O. Emery, *Continental Shelf: Mineral Resources*

Earth's crust, and in part to internal causes such as consolidation of deeper layers, crustal thinning under tension, and loss of light material from below the crust by processes within the Earth's mantle.

Origin of the continental slope

Several origins are invoked for the continental slope. One possibility already mentioned is downwarping beyond a hinge line that would increase the slope. Another is subsidence of a subaerial slope together with the hinterland. This seems likely for the west coast of Corsica because the subaerial valleys continue all the way down to the deep floor of the Mediterranean without change in character. The likelihood of this subsidence is augmented by the presence of great volumes of quartz sand in Tertiary beds situated in southern France and the Apennines. This sand came from a former land area situated in the present Mediterranean, where the depth to the old land surface is now 3,000 metres (10,000 feet).

Some slopes have been built out, as shown by their internal structure. Others show outcrops of older rocks that evidently have been truncated on their seaward end. Erosion of the slope by turbidity currents, especially during low sea levels, and slumping of subaqueous landslides have both played a part. Where the beds thicken seaward from a point far inland of the coastline, there must be a slight tilting of the continental border toward the deep sea. Extremely precipitous slopes like the one to the west of Florida into the Gulf of Mexico are obviously huge fault scarps. The consequences of several possible pro-

cesses and events on terrace structure are shown in Figure 4.

A special case is formed by the steep slope to the east of Florida. Deepwater drilling has revealed that the adjoining Blake Plateau is underlain by shallow water sediments and subsidence evidently has occurred. But the Gulf Stream has prevented sediment accumulation and it still sweeps the bottom clean.

A recent investigation of the continental margin of Europe has shown that three major events have occurred. In the course of the Mesozoic Era (*q.v.*) the margin developed by faulting and subsidence of an old land surface. During the Cretaceous a thick mass of sediment accumulated. Upbuilding amounted to a maximum of 4 kilometres (2.5 miles) and outbuilding to many kilometres. An episode of erosion followed at the close of the Mesozoic, and this resulted in turbidite deposition in the deep ocean and an unconformity within the terrace with respect to the later Tertiary sediment. The northern and western margin of Iberia were produced by faulting, but there appears to have been little terrace building.

A second erosional episode of faulting, slumping, and canyon cutting occurred at the end of the Tertiary. Possibly it was induced by glacial lowering of sea level. Since then, accumulation has again predominated, but a renewed advance of the ice would tip the balance in the opposite direction again.

Relation to continental drift

Geological relationships. The continental margins obviously have a bearing on the theory of continental drift and sea-floor spreading (*qq.v.*). The strongest case for drift is evidence that the Atlantic Ocean constitutes a rift that opened in the course of the Mesozoic between the Americas and Europe–Africa. This dovetails neatly with the Early Cretaceous age of the continental shelves on both sides that must have been constructed since the break. The theory originated from the good fit between the opposite margins of the Atlantic. But the most convincing proof of the rift is now supplied by paleomagnetism, deep-ocean drilling, geological fitting, the relics of the Permian Ice Age, and paleo-biogeography. It has become obvious that the shelf break is hardly better than the coastline or the margin of the continental rise as a feature by which to test the claim of a fit between the opposite coasts. The shelf edge came into being long after the rift, and the original continental margins appear to have been spread out and attenuated by loss of lateral support. The split off of the Rockall and Porcupine banks and the rotation of Iberia have caused deformations. Alpine mountain building has distorted the blocks themselves. Volcanism and delta building have inserted new masses. In spite of all these alterations the fit, whatever system of comparison is taken, is remarkably good. It is still a strong point in favour of the drift theory.

The origin and age of the Atlantic terraces and of the other margins created by continental rifting thus can be reasonably well accounted for. The younger ones of deep inland seas can be explained along the same lines; they developed after subsidence of their floors. But there remain the Pacific margins and a few others that have fronted the presumably primordial oceanic basin for a dozen times as long. These terraces are not more voluminous as one would expect but tend to be of modest size. It has been suggested that the continent has actively overridden the ocean floor and terrace or passively has allowed the ocean crust to be carried obliquely beneath them by mantle currents. The coastal ranges of continents could be tectonically compressed former terraces, but this would imply a radically different origin for all the great mountain chains crossing the continents, a consequence that is most unsatisfactory.

Even in broad view, the question of how the continental terrace has come into being requires a complex answer. Some coastal erosion is involved, but upbuilding and outbuilding on a subsiding margin are more important, and faulting and submarine erosion are also implicated. Sea level basically has had a unifying influence. The shallow inland seas are not genetically related to the terrace and they can be considered as typical continental areas that happen to have been flooded. Deep inland seas are widely considered to be foundered continental areas, in which chasms opened by continental drift. But in spite of the insight obtained, the origin and obliteration of many terraces remain obscure, especially those of the Pacific.

BIBLIOGRAPHY. T.H. VAN ANDEL and G.G. SHOR (eds.), *Marine Geology of the Gulf of California* (1964), see esp. J.R. CURRAY and D.G. MOORE, "Pleistocene Deltaic Progradation of Continental Terrace, Costa de Nayarit, Mexico," pp. 193–215; T.H. VAN ANDEL and J.J. VEEVERS, "Morphology and Sediments of the Timor Sea," *Bull. Bur. Miner. Resour., Geol. Geophys. Aust. 83* (1967), an example of a recent detailed investigation; J.R. CURRAY and R.D. NASON, "San Andreas Fault North of Point Arena, California," *Bull. Geol. Soc. Am. 78*, pp. 413–418 (1967); G.W. DEAN, "A Pragmatic Look at the Ocean's Mineral Resources," *Trans. N.Y. Acad. Sci.*, ser. 2, 31:731–736 (1969); K.O. EMERY, *The Sea off Southern California* (1960), an authoritative text on the continental borderland off Southern California, *The Continental Shelf and its Mineral Resources* (1967), and "Shallow Structure of Continental Shelves and Slopes," *Seast. Geol.*, 9:173–194 (1968), an up-to-date review; J.R. CURRAY, "Continental Terrace," in *Encyclopedia of Oceanography*, pp. 207–214 (1966); M.N. HILL (ed.), *The Sea: Ideas and Observations on Progress in the Study of the Seas*, vol. 3, esp. pp. 233–311 (1963), a standard reference volume for marine geology and geophysics; G.E. MURRAY, *Geology of the Atlantic and Gulf Coastal Province of North America* (1961); F.P. SHEPARD, *Submarine Geology*, 2nd ed. (1963), an excellent detailed treatment by an outstanding investigator of marine geology, written before the internal structure had been elucidated, and (ed.) *et al.*, *Recent Sediments, Northwest Gulf of Mexico* (1960), a symposium summarizing the results of much intensive work on sedimentation on the shelf; A.H. STRIDE *et al.*, "Marine Geology of the Atlantic Continental Margin of Europe," *Phil. Trans. R. Soc.*, ser. A, 264:31–73 (1969), historical development admirably elucidated; W.F. WHITTARD and R. BRADSHAW (eds.), *Submarine Geology and Geophysics* (1965), several papers treating recent advances.

(P.H.K.)

Continents, Development of

Analyses of geological and geophysical data obtained from the ocean floors since the 1960s have greatly impressed most authorities and have given rise to the hypothesis of sea-floor spreading. Briefly stated, this hypothesis holds that molten material wells up along submarine mountain zones, called the midoceanic ridge system, and spreads laterally away from the ridges. This spreading creates a successively younger ocean floor, and the flow of material is thought to bring about the migration, or drifting apart, of the continents. The continents bordering the Atlantic Ocean, for example, are thought to be moving away from the Mid-Atlantic Ridge at a rate of about two centimetres (0.8 inch) per year, thus increasing the breadth of the ocean basin by twice that amount. Wherever continents are bordered by deep-sea trenches, as in the Pacific Ocean, the ocean floor is believed to plunge downward, underthrusting the continents and ultimately re-entering and dissolving in the deeper levels of the Earth from whence it originated (*i.e.*, in the mantle, the zone beneath the Earth's crust).

Concept of sea-floor spreading

A veritable legion of evidence supports this concept, including absolute age determinations of sea-floor material, patterns of rock magnetism that parallel oceanic ridges, heat flow measurements, and the age of marine sediments. The most ancient sediments of the ocean floor are not older than Jurassic in age—that is, they do not exceed 150,000,000 years in age. This suggests that today's ocean basins have gradually assumed their present configurations during that span of time and also that the continents have been drifting apart contemporaneously with sea-floor formation.

Information on the history of the ocean basins is therefore available principally for the last 150,000,000 years of Earth history. The continents, however, have existed for at least 3,500,000,000 years, or about 24 times as long, and thus have undergone development prior to the advent of sea-floor spreading. This point is emphasized here because many authorities would interpret the history of continents and ocean basins on the basis of the oceanic evidence described.

This article will treat the development of continents

from a somewhat different viewpoint. Specifically, an outline of present knowledge of the Earth's structure and composition will be followed by coverage of the processes of continental development and by treatment of the evolution of the several continental elements resulting from tectonic forces (movements of the Earth's crust) through time. It will not attempt to cover the development and configuration of the continents in relatively recent time, which are related to Tertiary mountain building (2,500,000 to 65,000,000 years ago) and Quarternary glaciations of the last 2,500,000 years. The interested reader should consult the several related articles suggested below for this and other supplementary information as well as for differing viewpoints on the subject. In this vein, see EARTH, GEOLOGICAL HISTORY OF; CONTINENTAL DRIFT; SEA-FLOOR SPREADING; MOUNTAIN-BUILDING PROCESSES; ISLAND ARCS; OCEANIC RIDGES; and ROCK MAGNETISM. See also EARTH, STRUCTURE AND COMPOSITION OF (particularly for details on the pyrolite model); PRECAMBRIAN TIME; and parallel articles on the history of the Earth's atmosphere and oceans—*i.e.*, ATMOSPHERE, DEVELOPMENT OF; and OCEANS, DEVELOPMENT OF.

THE EARTH'S CRUST AND UPPER MANTLE

Composition and structure of the crust. The composition of the Earth's crust is relatively restricted from both a chemical and mineralogical viewpoint. Approximately 75 percent of the crust consists of oxides of silicon and aluminum (Table 1); the bulk of the rock-forming minerals are, accordingly, silicate minerals. The latter combine in various associations to form rocks of different mineral composition and origin.

Some rocks are of magmatic origin and were formed by the cooling and crystallization of molten material that flowed onto the Earth's surface (extrusive igneous rocks, such as basalt) or penetrated the crust and formed at depths below the surface (intrusive rocks, such as granite). The extrusive rocks occur as flows or layers, whereas intrusive rocks occur as massifs or as tabular bodies such as dikes and sills.

Sedimentary rocks form as a result of the erosion of preexisting rocks and the transportation, deposition, and accumulation of the derived mineral materials on the surface of the Earth, mainly in marine basins. But sedimentation also occurs on the floors of lakes and rivers or, in fact, anywhere on the land surface.

Metamorphic rocks are formed by recrystallization of both sedimentary and magmatic rocks at high temperatures (300°–700° C [600°–1,300° F]) and pressures (from one to ten kilobars). In the process, the original composition can undergo change caused by the addition or removal of certain components.

Seismic characteristics and crystal layering

The lower boundary of the continental crust is the Mohorovičić Discontinuity, characterized by an abrupt increase in the rate of propagation of seismic waves. Above the discontinuity, velocities do not exceed 7.4 kilometres (4.6 miles) per second for longitudinal waves, whereas below it they are usually equal to or greater than 8.1 kilometres (5.0 miles) per second. The depth of the Mohorovičić Discontinuity, which is the thickness of the continental crust, varies from 20 to 70 kilometres (10 to 40 miles). The former figure characterizes the crust of the Hungarian lowland, the latter, the Pamirs in Central Asia. Greater thicknesses generally are encountered under high mountains and lesser thicknesses under lowlands. The continental crust has a mean thickness of 43.6 kilometres (27.1 miles).

Three general kinds of layers may be distinguished in the crust: the sedimentary, the granitic, and the basaltic. The first consists of various sedimentary rocks and ranges in thickness from zero to 20 kilometres (zero to 12 miles); the mean thickness is 3.4 kilometres (2.1 miles). The granitic, or intrusive, rock layer reaches the surface in many places; it consists of about one-half granites and one-half gneiss and other metamorphic and magmatic rocks and is characterized by longitudinal seismic waves with velocities of 5.5 to 6.5 kilometres (3.4 to 4.0 miles) per second. The thickness of this layer varies from eight to 40 kilometres (5 to 25 miles), with a mean thickness of 20 kilometres (10 miles). The basaltic layer is complicated in composition: apparently, it consists of metamorphic rocks that have been intensely altered (the granulite and eclogite facies of metamorphism) and which are penetrated by numerous intrusions of basic and ultrabasic magmatic rocks (those rich in iron and magnesium) that have risen from the Earth's mantle. The basaltic layer is 15 to 40 kilometres (9 to 25 miles) thick, with a mean thickness of 20 kilometres (12 miles). In many instances, the crust exhibits many more subdivisions, which differ in their elastic properties. Seismic data suggest that the crust is everywhere divided by vertical fractures into blocks with different internal structure.

The mean chemical composition of the sedimentary (excluding carbon dioxide), granitic, and basaltic layers of the crust is shown in Table 1.

Table 1: Mean Composition of the Earth's Crust and Its Layers
(in weight percent)

	average crust	sedimentary layer	granitic layer	basaltic layer
SiO_2	61.9	49.95	63.94	58.23
TiO_2	0.8	0.65	0.57	0.90
Al_2O_3	15.6	13.01	15.18	15.49
Fe_2O_3	2.6	2.98	2.00	2.86
FeO	3.9	2.82	2.86	4.78
MnO	0.1	0.11	0.10	0.19
MgO	3.1	3.10	2.21	3.85
CaO	5.7	11.67	3.98	6.05
Na_2O	3.1	1.57	3.06	3.10
K_2O	2.9	2.04	3.29	2.58
P_2O_5	0.3	0.17	0.20	0.30

The upper mantle. The composition of the upper mantle—a shell underlying the crust with a thickness of about 900 kilometres (550 miles)—is probably close to that of some meteorites. Alfred Edward Ringwood, an Australian geochemist, suggests a composition that he calls pyrolite (see Table 2).

Table 2: Chemical Composition of the Upper Mantle According to the Pyrolite Model
(in percent)

	weight		weight
SiO_2	45.16	CaO	3.08
TiO_2	0.71	Na_2O	0.57
Al_2O_3	3.54	K_2O	0.13
Fe_2O_3	0.46	P_2O_5	0.06
FeO	8.04	Cr_2O_3	0.43
MnO	0.14	NiO	0.20
MgO	37.49	CoO	0.01

Within the upper mantle, at a depth of about 100 kilometres (60 miles) beneath the continents, a layer begins within which the seismic velocities are approximately 0.3 kilometre (0.2 mile) per second lower than in the overlying and underlying media. The thickness of this layer is obscure, but its base may be at a depth of 250 kilometres (150 miles). The peculiar properties of the layer result from the fact that the combination of temperature and pressure within it favours a partial melting of pyrolite with the emergence of films and drops of liquid basalt between the solid crystals (mainly of olivine). This process reduces the viscosity of the material and suggests the term asthenosphere, meaning a zone without strength. The partial melting should also produce a reduction of density in the asthenosphere, relative to the overlying upper mantle. The Earth's crust together with the solid layer of the upper mantle overlying the asthenosphere is called the lithosphere. The lithosphere and asthenosphere together are sometimes called the tectonosphere.

Processes that govern crustal development. The division of the Earth into a series of shells (upper and lower crust, upper and lower mantle, and outer and inner core) has occurred by geochemical differentiation of an initial Earth of much greater homogeneity. In the process of

differentiation, the heavier components sank, and the lighter ones rose.

Table 3: Elements Released and Layer Thickness of Mantle

element	thickness of layer (km)
Si	60
Al	140
Mg	60
Ca	50
Na	180
K	1,300

The volume of the continental crust is equal to 6,500,000,000 cubic kilometres (1,500,000,000 cubic miles), and that of the transitional zones (continental margins) is 1,540,000,000 cubic kilometres (370,000,000 cubic miles). Considering the composition of the Earth's crust and mantle and the total volume of crust involved, the thickness of the mantle layer that would be required to yield the various elements of the crust by differentiation can be calculated. Table 3 gives this thickness for some elements in the crust. The figures—particularly those for potassium (K)—suggest that the material of the crust is intimately connected with considerably deeper layers of the Earth.

Exogenic and endogenic processes

The basic geological processes that are active on the surface of the continents and within the continental crust can be termed exogenic and endogenic, respectively. Exogenic processes include rock weathering and the formation of sedimentary rocks. Rock destruction ordinarily occurs in uplands and rock formation (sedimentary deposition) in lowlands. Because the location of uplands and basins changes during geological time, redistribution of sedimentary material is continuously occurring on the surface of the continents.

Endogenic processes include magmatism, metamorphism, and tectonism. Rising magma contributes new material to the crust; the entire crust was formed initially by magmatic outpourings. Because the mean composition of magma is much more acidic (richer in silica) than the composition of the mantle, the process of magmatism can be regarded as an expression of the geochemical differentiation of the Earth. Metamorphism leads to the recrystallization of rocks. A special form of metamorphism is granitization, in which partial melting and recrystallization of pre-existing sedimentary and metamorphic rocks gives rise to granitic rocks. This produces some important changes in the mechanical behaviour of the crust; it tends to "heal up" fractures and to promote greater homogeneity within the crust.

Tectonic movements are mechanical processes of several types that affect the entire crust. Slow vertical oscillations or undulations of general nature are a type that affects vast areas at the same time. These movements result in transgressions and regressions (advance and retreat) of the seas and to corresponding changes in the relationship between marine and nonmarine sediments.

Wavelike vertical oscillations of the crust are a second type of movement. They lead to a more or less stable division of the crust into adjacent zones of uplift and downwarp of long-term geological duration. The system of uplifts and depressions thus created may shift through time transforming absolute uplifts into relative ones and vice versa. Absolute uplifts have relief and undergo erosion, whereas depressions subside (sometimes over wide areas) and undergo filling. Depressions represent zones where sediments accumulate to thicknesses as great as the magnitude of the downwarping. They are called geosynclines and serve as loci for mountain-building.

Deep fractures that divide the crust into a mosaic of blocks produce structural zones and affect the location of boundaries between regions of uplift and subsidence; they constitute a third type of tectonism. Vertical displacements of individual blocks along abyssal fractures may measure as much as six to eight kilometres (four to five miles).

Large-scale transcurrent faults in the Earth's crust are also of fundamental importance. These faults are horizontal displacements of the crust bounded by vertical fractures. A prototype is the San Andreas Fault in California. During the earthquakes of the 19th and 20th centuries in that region, strike-slip movements (horizontal offsets along the trace of the fault) with amplitudes of several metres were observed.

Tectonic deformation includes folding (bending) of rock strata and attendant phenomena that occur in mountain ranges and belts. Such deformations do not embrace the entire crust but are always geographically confined.

ENDOGENIC REGIMES OF THE CONTINENTS

The continents are characterized by the effects of regular combinations of endogenic processes within the Earth. Such combinations constitute endogenic regimes, which appear in each interval of geologic time in definite zones that may be called tectonic zones.

Geosynclines and platforms. Geosyncline and ancient-platform regimes have been prominent on the continents during all of Phanerozoic time (the last 570,000,000 years); but they also existed earlier, in Precambrian time, although only certain aspects of these regimes are evident from those ancient rocks. The geosynclinal regime is characterized by tectonic undulations that divide the entire geosyncline into closely located zones of uplift and subsidence with a width of 50 to 100 kilometres (30 to 60 miles). Undulation amplitudes may attain 15 or 20 kilometres (9 to 12 miles), and regional metamorphism and granitization commonly occur. Strong linear folding and general crumpling of rock strata, together with block folding and injection folding give rise to a folded belt of mountains at the site of the geosyncline. Magmatic intrusions of several kinds may accompany and follow the folding and mountain building. The development of geosynclines is linked with intricate fracturing of the crust. A network of deep faults produces vertical movements of great amplitude and ensures high penetrability of the crust by magma.

Characteristics of folded zones

The ancient-platform regime is characterized by very slight contrasts in undulations. Broad and gently sloping uplifted and depressed areas have rounded irregular outlines; they are many hundreds of kilometres in diameter, and vertical movements have amplitudes of hundreds of metres to a few kilometres.

Regional metamorphism and granitization are absent, and only block folding and injection folding occur; there is no general crumpling of rock strata. Magmatism also is typically absent, and the regime characteristics all serve to indicate the presence of consolidated crust with low penetrability.

Intermediate regimes that comprise the entire spectrum from the typical geosyncline to the typical ancient platform also exist, but the rift regime and the regime of continental margins are quite distinctive.

Rifts and continental margins. In the rift regime, systems of large grabens (long tectonic depressions 30 to 60 kilometres [20 to 40 miles] in width and bounded by steep faults) are formed. The regime is confined to gentle archlike uplifts of the Earth's crust that are several kilometres in height and 1,000 to 2,000 kilometres in diameter. Rifts are invariably accompanied by intense volcanism (see further RIFT VALLEYS).

Faulting and volcanism

The regimes of continental margins are always characterized by subsidence of the crust. On Atlantic-type margins, the subsidence is quiescent, and the shelf gradually tilts toward the ocean. On Pacific-type margins, the subsidence occurs along fractures and is accompanied by intense volcanism and earthquakes.

Evolution of endogenic regimes. During the early part of Precambrian time (*q.v.*), regional metamorphism, granitization, and complicated deformations of rocks were widespread over the areas of the present continents (about 3,000,000,000 years ago). These phenomena point to the geosynclinal regime and suggest that it prevailed over the whole surface of the continents. During this stage of continental development the granite layer of the continental crust was formed; earlier, the upper part of

the crust may have been more basic (basaltic) in composition.

The crust later became divided: (1) into areas in which regional metamorphism, granitization, and rock deformation occurred, and (2) into areas in which such phenomena were absent during the same period of time. This stage marks the beginning of the division of the crust into geosynclines and platforms. Earlier than 1,600,000,000 to 1,800,000,000 years ago, however, the division into geosynclines and platforms was not stable, as the locations of these two regimes were constantly alternating. Thereafter, when a stable geosynclinal-platform regime did set in, the ancient platforms (shields) known today (North American, East European, Siberian, Chinese, South American, African, Indian, Australian, and Antarctic) began to form. They grew from the merger of a large number of small protoplatforms. Active endogenic processes ceased, and continental evolution became directed to the persistent growth of the area occupied by the platform regime. In this process the ancient platforms played a role of stabilization nuclei from which the platform regime was spreading over the continents.

Persistence of shield areas

As the area of the platforms expanded, the areas occupied by geosynclines diminished. As a result, the geologically young geosynclines, including those of Cenozoic age (the last 65,000,000 years), now form two well-defined linear zones—the Circum-Pacific and the Mediterranean (Tethyan) belts. The evolution of endogenic regimes toward lesser activity and more consolidation of the crust coincided, in certain places and times, with a reverse process of tectonic and magmatic activity. This reverse process was always subordinate, however.

The decrease in the area occupied by geosynclines indicates that the area where processes of differentiation, metamorphism, and granitization occurred also was decreasing, and that the formation of new continental crust was confined to gradually narrowing zones. The rift regime and the regime of continental margins left their marks only during the most recent geologic times.

Tectonic cycles. Determinations of the absolute age of metamorphic minerals suggest that Precambrian metamorphism was preferentially confined to certain intervals of time. These time intervals are (in 1,000 millions of years before the present): 3.0–2.8; 2.6–2.5; 2.2–2.0; 1.8–1.6; around 1.2; 1.0–0.9; and 0.6–0.5. These figures show that Precambrian metamorphism recurred at discrete intervals of 300,000,000 to 600,000,000 years. For the Phanerozoic Eon (the post-Precambrian record) it is possible to establish a recurrence of the entire complex of endogenic processes that proves to be almost synchronous for all continents, or for very large portions of them. Thus, it is possible to speak of endogenic cycles on a planetary scale; these are commonly called tectonic cycles.

Periodicity of uplift and magmatism

Each cycle begins with subsidence of the crust and ends in uplift of the crust, thus giving rise to marine transgressions and regressions. Undulations are evident on platforms, where they are only slightly obscured by other phenomena; but, in geosynclines, undulations have greater amplitudes. During subsidence, the geosyncline becomes a sea basin of some depth; after uplift at the end of a cycle, it is transformed into a zone of folded mountains.

Magmatism in geosynclines during the endogenic cycle is divided into three stages: the initial stage, coinciding with the subsidence of the crust, consists in the outflow of basic lavas onto the bottom. During this same stage, basic, ultrabasic, and often intrusive rocks are formed. The middle stage coincides with the formation of granitic massifs. This process is closely related to regional metamorphism and occurs almost simultaneously. At the same time, most intensive folding begins. The final stage, coinciding with the uplift of a geosyncline, is represented by the formation of magmatic veins and dikes (very different in composition from previous intrusions) and by volcanic eruptions, this time with the outflow of andesitic and basaltic lavas.

Compressive forces and folding. Folding occurs on a much smaller scale than do vertical movements of the crust, magmatism, and regional metamorphism. The basic types of folding are block folding, injection folding, and the folding associated with general crumpling. The latter type of folding is developed only in geosynclines. The other two types of folding occur both in geosynclines and on platforms. Only general crumpling, however, requires the direct action of horizontal compressive forces to explain it. Such forces may be generated within the geosyncline. The concept of external compression is in conflict with the shape of many folded zones, which often are sharply bent and break up into nearly closed ovals. This shape cannot be reconciled with external pressure, which would require that a single platform press on the geosyncline in different directions. The external-force idea also is contradicted by the history of folding in geosynclines. Folding always originates in the axial zone of the geosyncline, far away from the edges of the platform, and only later spreads toward the platform. If the folding were caused by pressure from the platform, the sequence would have to be reversed.

Origin of compressional forces

Moreover, the folding associated with general crumpling forms separate strips in the geosyncline, completely surrounded by zones of block folding and injection folding. The requirement that pressure applied from without "jumps over" the zones of block and injection folding to cause general crumpling in the strip cannot be explained. Horizontal compression of geosynclines should arise from within the geosyncline proper.

A model that accounts for the internal origin of the forces of compression stems from the assumption that the source of all deformations is the force of gravity. Three specific mechanisms have been suggested: the first is the slipping or flowing of layered rocks from the slopes of tectonic uplifts. This mechanism manifests itself in the formation of Alpine-type tectonic nappes (faulted and overturned folds). A second consists in a spreading of the upper part of uplifted crustal blocks as the vertical fractures bordering the blocks become inclined; the blocks then slip, exerting horizontal pressure on the surrounding rocks. A third mechanism consists of the formation of "deep diapirs" composed of granites, granite-gneisses, or metamorphic rocks that rise up through the overlying rocks of the Earth's crust. In the process, they undergo complicated deformations and also give rise to deformations in the surrounding rocks, which are pushed sideways.

ENDOGENIC REGIMES AND DEEP-SEATED PROCESSES

Nature of geophysical evidence. Information on the structure of deep-seated layers of the crust and the upper mantle is obtained by a variety of geophysical methods, principally seismic, gravimetric, and heat flow. These methods make it possible to trace the base of the Earth's crust and to establish its thickness at various sites.

Crustal thickness. It is known, for example, that the very high mountains, such as the Pamirs or Andes, are conspicuous by reason of the very thick crust (up to 70 kilometres [45 miles]) that underlies them. By contrast, beneath oceanic depressions the solid crust is very thin (five to seven kilometres [three to four miles]). In all other cases the relationship between the relief and the thickness of the crust varies greatly. Such ambiguous relations can be accounted for by several factors. First, there are great variations in mean density of the crust from place to place (horizontal inhomogeneities). In addition, the attainment of gravitational equilibrium (isostasy) depends not only on variations in the density and thickness of the crust but also on local variations in the density of the upper mantle. The situation is further complicated by the fact that structural peculiarities of the crust correspond not only to the more recent crustal regime but also to the previous history of the crust. Some deep structures have remained fixed in the crust for long periods of geologic time.

Density variations

Rift valleys serve as an example of correlation between deep seismic structures and a recent endogenic regime. They are characterized by reduced thickness of the crust (about 20 kilometres [12 miles]) and by the presence of a lens of material between the crust and the upper mantle

with seismic velocities that are intermediate between these two shells (7.4–7.8 kilometres [4.6–4.8 miles] per second for longitudinal waves). Such a lens has a thickness of 20 to 30 kilometres (12 to 20 miles) and it is believed to be a mixture of crustal material and material of the upper mantle.

The regimes of the continental margins are related to reduced thicknesses of the crust. In North America the crust thickness at the centre of the continent reaches an average of 40 kilometres (25 miles), whereas at the Atlantic margin it falls to 30 kilometres (20 miles) and along the Pacific Coast to 18 kilometres.

The asthenosphere and heat flow. There is a correlation between the endogenic regimes on the surface and the underlying asthenosphere. Under the more stable sections of the ancient platforms—the crystalline shields—the asthenosphere can hardly be identified. It is much more evident under regions of intense recent uplifts. Under rifts, the asthenosphere is even more evident. And under the volcanic zone of the Kamchatka Peninsula and the Kuril Islands, a projection of the asthenosphere actually reaches upward to the base of the crust.

Heat flow data

Important data have been derived from the study of the heat flow of the Earth. The average rate of heat flow on the continents is equal to the mean flow in the oceans and the mean for the whole Earth (1.47 microcalories per square centimetre per second). The lowest heat flow is observed on shields (the average is 0.98 in the same units). The mean value for platforms is close to the mean value for continents. But in the Tien Shan, in a Central Asian region of intense recent mountain uplift, the flow reaches 1.80. It is also high in rift valley areas, with the highest heat flows in volcanic regions. From these data it can be concluded that there is a correlation between heat flow and tectonic and magmatic activity.

The most active zones of continents are geosynclines. One of the processes confined to them is regional metamorphism. The study of the thermodynamic conditions of formation of various metamorphic minerals, with an evaluation of the most probable initial depth, has led to the conclusion that at the time of metamorphism the heat flow in geosynclines was three to five times higher than the average heat flow for continents. Because the metamorphic process occurs only during a particular stage of a tectonic cycle, the geosyncline apparently is a zone of high heat flow that undergoes periodic variations through time.

Vertical-force basis of continent building. All the processes described are connected with considerable depths of the Earth's interior, specifically with the asthenosphere, which differs beneath the various endogenic zones. But because the asthenosphere is dependent in its manifestations upon an influx of material and heat from deeper regions, the indicated relationships most obviously go still deeper. The repeated recurrence of a specific sequence of endogenic processes and the stable position of the ancient platforms, from which (as continental nuclei) the later regimes spread, point to a deep and long-term stability in the relationships between the continental surface and the interior. The mosaic character of the Earth's crust, in which regions with different regimes persist side by side for long periods of time, testifies to the fact that although vertical relations were long-term and deep, the horizontal relations and horizontal "mixing" of regimes were very restricted. This conclusion is in conflict with the hypothesis of sea-floor spreading, because it presupposes considerable horizontal movements both of the lithosphere of the continents and within the asthenosphere. Such movements would inevitably have disrupted all deep vertical relations.

Long-term continental stability

Heat and gravitational differentiation. The basic process in the interior of the Earth has always been heating provoked by radioactive processes. Heating causes the gravitational differentiation of the material of the Earth mainly by means of partial melting. The mechanism of this process is the most mighty source of internal planetary energy. It is not necessary to postulate the existence of any large loops of convection. The mosaic structure of the crust indicates that gravitational convection has been realized by the rising and sinking of relatively narrow streams, not surpassing the width of a separate tectonic zone in their horizontal dimension. The rising of relatively light material signifies a transfer of heat to the upper layers, within which the additional heating would stimulate differentiation. It is very probable that this mechanism of successive differentiation at different levels can account for the intricate complex of endogenic processes.

There is a possibility of differentiation within the Earth at depths of several hundred kilometres in accordance with the principle of zone melting. This process consists in the formation of a partly molten layer at a depth of 400–500 kilometres (250–300 miles) that subsequently moves upward and melts through the overlying material. In the process, the partly molten layer gradually cools off and becomes integrated with the surrounding material. Sometime thereafter, the second molten layer appears in the interior. The concept of a mobile and periodically renewed asthenosphere can be useful as an explanation of the tectonic cycles and large-scale undulations previously described. During this process of zone melting the volume of the mantle under the continents changes, leading to the periodical undulations on the surface relative to the level of the oceans. Upper layers of the mantle also are periodically enriched with the comparatively light material.

Melting within the asthenosphere

The next order of differentiation is probably located within the asthenosphere itself. Basalt is melted, and the material of the asthenosphere is altered in its mobility and density. If a relatively large volume of basalt is melted, the density of the asthenosphere decreases perceptibly and may become lower than the density of the overlying layers of the lithosphere. In this situation of density inversion on the surface of the asthenosphere, disturbances that could cause undulations in geosynclines may be expected.

Basaltic bodies (asthenoliths) that rise from the asthenosphere and penetrate the crust heat its material, thus stimulating metamorphism. The beginning of folding is closely connected in time with this event. The transition from a geosyncline to a platform may occur after repeated partial melting, when the composition of the asthenosphere becomes such that no further considerable melting is possible.

Crustal-sinking hypothesis. On the boundaries of continents and partly within them are seas and peripheral parts of oceans; the crust under both types of continental areas is of an oceanic type. The possibility exists that there has been a process of "basification" of the continental crust, a transformation of it into oceanic crust. This would occur as a result of a very intensive heating of the Earth's interior that would initiate a large-scale melting process and, in turn, would cause molten material to penetrate the continental crust in large quantities. There, the molten material would drive the contained water out of the crust and so alter its material that it would become heavier. After cooling, the crust would sink into the mantle in the form of separate blocks and there dissolve.

According to this concept the process of differentiation that leads to the formation of continental crust may be replaced in time by the process of basification, which tends to destroy the crust. Despite the fact that the continents have been under study for at least 200 years, many intricate mysteries about their development remain to be solved. An explanation of continental structure and history can only be set forth within the framework of plausible hypotheses, one of which has been presented in this article.

BIBLIOGRAPHY. A large and specialized literature treating the development of the continents and, in effect, the development of the Earth itself, is available. The interested reader can best consult two general works, however, which serve as a guide to many of the ideas mentioned or alluded to in this article and to additional references: P.J. HART (ed.), *The Earth's Crust and Upper Mantle* (1969); and V.V. BELOUSSOV, *Basic Problems in Geotectonics* (1962; Eng. trans. from the 2nd Russian ed. of 1962).

(V.Be.)

Contracts, Law of

The nature of contract law

A contract, in the simplest definition, is a promise enforceable by law. The promise may be to do something or to refrain from doing something. The making of a contract requires the mutual assent of two or more persons, one of them ordinarily making an offer and another accepting. If one of the parties fails to keep the promise, the other is entitled to legal recourse against him. The law of contracts has to do with such questions as whether a contract exists, what the meaning of it is, whether a contract has been broken, and what compensation is due the injured party.

THE EVOLUTION OF THE LAW OF CONTRACTS

Contract law is the product of a business civilization. It will not be found, in any significant degree, in precommercial societies. Most primitive societies have other ways of enforcing the commitments of individuals, through ties of kinship or by the authority of religion. In an economy based on barter, most transactions are self-enforcing as the transaction is complete on both sides at the same moment. Problems may arise if the goods exchanged are later found to be defective, but these problems will be handled in terms of property law —with its penalties for taking or spoiling the property of another—rather than in terms of contract law.

Even when transactions do not take the form of barter, primitive societies continue to work with notions of property rather than of promise. In early forms of credit transactions, kinship ties were relied upon to secure the debt, as when a tribe or a community gave hostages until the debt was paid. Other primitive forms of security took the form of pledging land or pawning an individual into "debt slavery." Some credit arrangements were essentially self-enforcing: livestock, for example, might be entrusted to a caretaker who received for his services a fixed percentage of the offspring. In other cases—constructing a hut, clearing a field, or building a boat—enforcement of the promise to pay was more difficult but still was based on concepts of property. In other words, the claim for payment was based not on the existence of a bargain or promise but on the unjust detention of another's money or goods. When a worker sought to obtain his wages, the tendency was to argue in terms of his right to the product of his labour.

A true law of contracts—that is to say, a law of enforceable promises—implies the development of a market economy. Where a commitment's value is not seen to vary as a function of time, ideas of property and injury are adequate and there will be no enforcement of an agreement if neither party has performed, since in property terms no wrong has been done. In a market economy, on the other hand, a person may seek a commitment today to guard against a change in value tomorrow; the person obtaining such a commitment feels harmed by a failure to honour it to the extent that the market value rises above the agreed price.

Roman law. The Roman law of contracts, as found in Justinian's law books of the 6th century AD, reflected a long economic, social, and legal evolution. It recognized various types of contracts and agreements, some of them enforceable, others not. A good deal of legal history turns upon the classifications and distinctions of the Roman law. Only at its final stage of development did Roman law enforce, in general terms, informal executory contracts—that is, agreements to be carried out after they were made. This stage of development was lost with the breakup of the empire. As western Europe declined from an urbanized, commercial society into a localized, agrarian society, the Roman courts and administrators were replaced by relatively weak and imperfect institutions.

The rebirth and development of contract law was a part of the economic, political, and intellectual renaissance of western Europe, including England. It was everywhere accompanied by a commercial revival and the rise of national authority. Both in England and on the Continent, the customary arrangements were found to be unsuited to the commercial and industrial societies that were emerging. The informal agreement, so necessary for trade and commerce in market economies, was not enforceable at law. The economic life of England and the Continent flowed, even after a trading economy began to develop, within the legal framework of the formal contract and of the half-executed transaction (that is, a transaction already fully performed on one side). Neither in continental Europe nor in England was the task of developing a law of contracts an easy one. Ultimately, both legal orders succeeded in producing what was needed: a body of contract doctrine by which ordinary business agreements, involving a future exchange of values, could be made enforceable.

The law of merchants

The new contract law began to grow up on the Continent and in England through the practices of merchants; these were at first outside the legal order and could not be upheld in courts of law. Merchants tended to develop informal and flexible practices appropriate for active commercial life. By the 13th century, merchants' courts had been established. The merchant courts provided expeditious procedures and prompt justice and were administered by men who were themselves merchants and thus fully aware of mercantile problems and customs.

In the 12th and 13th centuries, the development of the law of contracts on the Continent and in England began to diverge. In England the common law of contracts developed pragmatically through the courts. On the Continent the process was very different, with speculative and systematic thinkers playing a much larger role.

The common law. From perhaps the 13th century on, English common law dealt with contractual problems primarily through two actions: debt and covenant. When a fixed sum of money was owed, under an express or implied agreement, for a thing or a benefit given, the money was recoverable through a simple action at debt. Other debt action was available for breach of a promise, made in an instrument with a seal, to pay a fixed sum of money. A so-called action at covenant could also be brought for breach of a promise under seal. These actions did not, however, provide a remedy for the breach of an informal agreement to do something. In the 15th century, the common-law courts started to develop a form of action that would render such agreements enforceable, and by the middle of the 16th century, they had done so through the form of action known as assumpsit ("he has undertaken"). Originating as a form of recovery for the negligent performance of an undertaking, it came step by step to cover the many kinds of agreement called for by expanding commerce and technology. Having established in principle a comprehensive remedy, it was necessary for the courts to limit its scope. The courts found the limiting principle in the doctrine of "consideration," according to which a promise as a general rule is not binding unless something is presently given or promised in exchange. This consideration need not be of commensurate value, but it must be bargained for and cannot be simply a formality.

Civil law. On the Continent, the revived study of classical Roman law had an immense influence upon the developing law of contract. It stimulated men to rediscover or construct a general law concerning the validity of agreements. The Roman law, however, as crystallized in Justinian's law books, tended to confirm the notion that something more than an informal expression of agreement was required if a contract was to be upheld by a court. Another significant influence in the development of contract law on the Continent was the Roman Catholic Church. The church in its own law (canon law) strongly supported the proposition that a simple, informal promise should be binding (*pacta sunt servanda*). This attitude was to encourage the development of informal contracts. The natural-law philosophers took up such ideas as *pacta sunt servanda*, although they were slow to abandon the view that some contracts, especially contracts of exchange, should require part performance if they were to be held enforceable. By the time of the 18th century, the speculative and systematic thought of jurists and philosophers had finally and fully carried the

day. The legal writers and legislators of the period generally considered informal contracts as enforceable in the courts. Thus in the French Civil Code of 1804, contract was approached essentially in terms of agreement; obligations freely assumed were enforceable except when the welfare of society or the need to protect certain categories of persons such as minors dictated otherwise. With the generalization that contract rests ultimately on agreement, the civil-law systems achieved a foundation quite different from the common law's view that contract is basically a promise supported by a consideration.

Flexibility of modern contracts

All the Western systems of modern contract law provide mechanisms through which individuals can voluntarily assume, vis-à-vis others, legally binding obligations enforceable by the other person. Contract law strives to give legal expression to the endlessly varying desires and purposes that human beings seek to express and forward by assuming legal obligations. The resulting system is open-ended; in principle, no limits are set in modern contract law to the number of possible types of variations of contracts.

THE SETTING OF STANDARDS

In theory, contractual obligations should be concluded between parties of substantially equal awareness and bargaining power and for purposes fully approved by society. The law reflects this utopian idea in the sense that it tends to conceive of contract as an arrangement freely negotiated between two or more parties of relatively equal bargaining power. The manifestations of intention required to form a contract are accordingly thought of as indicating real willingness, although in fact they may simply represent acquiescence.

Fairness and social utility. Much of the law of contract is concerned with ensuring that agreements are arrived at in a way that meets at least minimum standards, respecting both parties' understanding of, and freedom to decide whether to enter into, the transactions. Such provisions include rules that void contracts made under duress or that are unconscionable bargains; protection for minors and incompetents; and formal requirements protecting against the ill-considered assumption of obligation. Thus, section 138 of the German Civil Code renders void any contract "whereby a person profiting from the distress, irresponsibility, or inexperience of another" obtains a disproportionately advantageous bargain. In addition, more general social requirements and views impinge upon contracts in a number of ways. Certain agreements are illegal, such as—in the United States—agreements in restraint of trade. Others, such as an agreement to commit a civil wrong, are held by the courts to be contrary to the public interest. Certain systems discourage some purposes, such as the assumption of a legally binding obligation to confer a gift of money or other gratuitous benefit upon another by various special requirements.

Legal systems often have recourse to interpretation in the interest of fairness and social utility. Many litigated cases in which a remedy is sought for breach of contract are concerned with the meaning to be attached to the verbal expressions and acts of the parties in their dealing with each other. Ambiguities, for example, may be resolved against the party thought to have the superior bargaining position. This decision is common in cases in which one party is able to set the terms of a contract without bargaining. Again, a written agreement may be interpreted against the party who drafts or chooses the language. Or the court may prefer an interpretation it finds to be in accord with the public interest.

Although all legal systems try to achieve a reasonable approach to freedom of contract, there are bound to be contractual obligations that depart in some degree from the ideal. No one seeking to enforce a contract is required to show affirmatively that it advances specific ends desired by society or that the contracting process is without blemish. Such a requirement would be administratively cumbersome and expensive. In addition, it would reduce the general usefulness of the contract as an economic and social instrument. Differences in the economic resources available to individuals are found in most societies; to the extent that these differences flow from general conditions and are reflected in, rather than produced by, individual contracts, it is usually not feasible to take remedial action through the law of contracts. A single contract, moreover, is often only one element in a complex of economic and legal relations. Thus, in times of severe inflation or deflation, it may simply not be feasible to seek to deal with the resulting inequities in terms of redoing individual contracts.

Big industry and the public interest

Contracts of adhesion. There are large areas of economic life in which the parties to contracts have such unequal bargaining positions that little real negotiation takes place. These contracts are often known as contracts of adhesion. Familiar examples of adhesion contracts are contracts for transportation or service concluded with public carriers and utilities and contracts of large corporations with their suppliers, dealers, and customers. In such circumstances a contract becomes a kind of private legislation, in the sense that the stronger party to a large extent assigns risks and allocates resources by its fiat rather than through a reciprocal process of bargaining. Enforcement of such standard contracts can be justified on the ground that they are economically necessary. The question then becomes whether these decisions are to be made by private enterprise or by other agencies of society—in particular, government—and to what extent the interest of those who deal with such economic enterprises can be represented and protected in the decision-making process.

Contract law in such cases provides only what can be called the legal relationship. The content of the relationship derives not from bargaining between the parties but from the fiat of the large enterprise often offset by the fiat of some government agency. In a sense, the socially regulated contract of adhesion seeks to eat the cake of bureaucratic rationality while having, as well, the cake of individual choice and decision. Doubtless both cakes are diminished in the process, but the result may well be more satisfying than if only one had to be chosen. At all events, the resulting legal-economic phenomenon is radically different from that envisaged by traditional contract law.

THE RULES OF DIFFERENT LEGAL SYSTEMS

Traditional contract law developed rules and principles controlling the voluntary assumption of obligations, regulating the performance of obligations so assumed, and providing sanctions for failure to perform.

The question of whether a contract exists

Offer and acceptance. Some of the rules respecting offer and acceptance are designed to operate only when a contrary intention has not been indicated. Thus, in German law an offeror cannot withdraw his offer until the time stipulated in the offer or, if no time is stipulated, until a reasonable time has passed; but this rule yields to a statement in the offer to the effect that it shall be revocable. In Anglo-American common law, when parties contract by correspondence the acceptance takes place on dispatch of the letter, but the offeror can stipulate that no contract will be formed until the acceptance has reached him. These rules serve to fill in points on which the parties in their negotiations have not, for one reason or another, been specific.

Another function of rules relating to offer and acceptance is to enable the parties to understand and to mark when their discussions pass from an exploratory stage to the stage of commitment. The concepts of offer and acceptance are somewhat formal; they assume that the negotiations pass through clearly distinguishable phases, which is often not the case. But they help the parties to distinguish negotiation from commitment. The two words offer and acceptance become firmly associated with the assumption of obligations.

Different legal systems frequently advance comparable policies in quite different ways. Several distinctly different patterns are found in the approach of modern legal systems to the problems of whether an offeror is free to revoke his offer before acceptance, and of when an acceptance is effective to form a contract. Perhaps the polar extremes are represented by the German law, on one

hand, and the Anglo-American common law on the other. In the German view, an offer binds the offeror for any stipulated period or, when the offer is silent as to time, for a reasonable period unless the offeror has expressly made the offer revocable. The common-law rule is the opposite: an offer is revocable until it has been accepted. The two systems also have sharply divergent rules with respect to the point at which, when the parties are contracting by correspondence, the acceptance takes effect to conclude the contract. In German law, the acceptance takes effect when it reaches the offeror, in the sense that he either knows or can learn of it. In the common law, on the other hand, if the offeree uses an appropriate means of communication, the acceptance is effective on dispatch unless the offeror stipulated the contrary in his offer. (A revocation by the offeror, however, does not take effect until received by the offeree.)

How are these divergencies in the rules respecting offer and acceptance to be explained? In particular, do they reflect fundamental policy differences or simply different techniques designed to forward quite similar purposes? An examination of a typical problem posed when parties contract by correspondence suggests the latter explanation. Upon receipt of an offer, the offeree frequently changes his position by, for example, refusing or ignoring other offers, neglecting to seek additional offers, or himself making propositions based on the offer made to him. For this reason the legal system sees a need to provide the offeree with a secure point of departure for his decision, in order both to protect him and to facilitate commerce generally. The German system provides this protection by making the offer in principle irrevocable. The common law, on the other hand, found this solution excluded by its doctrine of consideration; as the offeree does not give anything in exchange for the offer's irrevocability, consideration is lacking to support an obligation not to revoke. (On the other hand, the Uniform Commercial Code, which has been adopted everywhere in the United States except Louisiana, provides that a firm offer made by a merchant is irrevocable even though the other party has given no consideration. If legislation, rather than judicial development, were more usual in the common law, changes such as this one would presumably have occurred much earlier.) The common law is not entirely insensitive to the offeree's predicament. The rule that the acceptance is effective upon dispatch creates a situation in which the offeror who wishes to revoke his offer is uncertain whether or not he can still do so, since his revocation is not effective until receipt, whereas the offeree's acceptance, if one is made, takes effect on dispatch. This uncertainty makes the consequences of an attempted revocation unpredictable and thereby inhibits an offeror who might otherwise seek to revoke. In sum, the German and Anglo-American systems both try to achieve, and in a measure succeed in achieving, a fair balance between the offeror and the offeree.

The criteria of a legal contract

Unenforceable transactions. In all systems of contract law, certain classes of transactions are treated as unenforceable by the judicial process because they are thought to involve unusual hazards for a contracting party or to be of marginal social utility. There are, in both civil-law and common-law systems, four kinds of concern that lead the systems to treat certain types of transaction as unenforceable. These four kinds of concern may be called evidentiary, cautionary, channelling, and deterrent. The evidentiary concern springs from the desire to protect both the individual citizen and the courts against manufactured evidence and insufficient proof. The cautionary concern seeks to safeguard the individual against both his own rashness and the importuning of others. The channelling concern seeks to mark off or label obligations that may be enforceable and to direct attention to the problem of the extent and kind of the legal obligation, so that the individual will know the legal significance that his action may have. Finally, the deterrent concern refers to those types of transaction that are discouraged because they are felt to be of doubtful value to society.

Two quite different techniques are used to delineate types of transaction that are unenforceable in their natural, or normal, state. The first proceeds by describing the type in functional or economic terms. The common-law Statute of Frauds enacted by the British Parliament in 1677 provided that the following six kinds of contracts should be unenforceable unless expressed in writing: contracts to sell goods exceeding a certain value; contracts to sell any interest in land; agreements that are not to be performed within a year of their making; agreements upon consideration of marriage; suretyship agreements; and undertakings by an executor or administrator to be surety on a debt of the deceased for which the estate is liable. Civil-law systems typically describe as unenforceable in the absence of an appropriate formality noncommercial contractual obligations exceeding a certain value; mortgages created by contract; noncommercial compromise agreements; marriage contracts; agreements binding a party to transfer all, or a fractional part of, his property; leases to run for more than a year; assumptions of the obligation to stand as surety, at least when the operation is not a commercial one on the surety's part; promise of an annuity; and promises to make gifts.

Another less direct technique for delineating unenforceable types of transaction derives from the common law's doctrine of consideration. It holds transactions unenforceable in the absence of a bargained-for exchange. This class would include, for example, promises to make gifts. The approach tends to be too all-embracing, treating certain types of transaction as suspect when there is little or no practical justification for doing so. It is not clearly demonstrated, for example, that an option agreement made by two businessmen should be handled differently from many other kinds of commercial dealings. A strong argument exists that the common law's handling of commercial options, business compromises, and other business transactions lacking an element of exchange is more a logical deduction from the general doctrine of consideration than an expression of justifiable policy concerns.

Except in cases where the ground for unenforceability is radical, when a given transaction type is considered unenforceable the legal system should prescribe an extrinsic element the addition of which will cure the defect—for example, expressing the agreement in writing, performing it in part, or using a notarial contract, which involves the participation of a lawyer who holds a special appointment from the state and is charged with handling and recording various types of transactions.

Decline of the use of seal

A complex situation has arisen with respect to the two most generally available extrinsic elements, the seal and the payment of a nominal consideration. Various states of the United States no longer consider the seal as an effective extrinsic element. The seal's decline is rooted in its changed significance in the modern, literate, democratic world. The seal was originally an impression, usually in wax, of a device, or design, representing an individual or a family. In modern times, the courts, with legislative assistance in a fair number of the states of the United States, have recognized easy-going substitutes for the wax seal. The effect has been to render the seal progressively less effective, particularly from the cautionary perspective, and many courts now refuse to accept it as a satisfactory formality.

Nominal consideration is a subtle and ingenious formality. Its essence is the introduction of a contrived element of exchange into the transaction. Thus A, desiring to bind himself to give B $10,000, requests B to promise to give (or give) A a peppercorn in exchange. B's promise (or performance) is an element, extrinsic to a normal gift promise, introduced by the parties in an effort to render the transaction enforceable (since the law does not treat normal gift promises as enforceable). Common-law courts often accept nominal consideration when used in a business context, such as in an option arrangement or a compromise agreement; its effectiveness is understandably more doubtful in the context of a gift promise, since such a transaction involves greater dangers for one party and is socially more marginal.

Civil-law systems have less need than the common law

for a formality such as nominal consideration; they prescribe methods directly in their statutes. Interestingly enough, however, in some civil-law systems an analogous, judicially developed formality has emerged—the disguised donation (*donation déguisée*) of French law, in which the parties cast a gift promise in the form of an onerous transaction, such as a sale. It can be argued that both the nominal considerations and the disguised donation serve at least the cautionary and channelling functions of formalities mentioned above.

Another kind of extrinsic element recognized by some courts, especially in the common-law world, is one party's reliance upon the promise of the other. The fact of reliance argues in favour of enforcement because it indicates that an underlying understanding existed between the parties and because the relying party may suffer as a consequence of his change of position. Some courts will enforce initially suspect transactions when several extrinsic elements are present in combination. A common-law court, for example, may enforce a gift promise in which the element of reliance was present in addition to a seal or a nominal consideration. Other extrinsic elements, either alone or in combination with reliance, a seal, or a nominal consideration, may also render a transaction enforceable. Cases, for example, in which the promisor dies without attempting to revoke a gift promise could be enforced, as distinguished from cases in which the promisor seeks to revoke.

Performance. Contract law seeks to protect parties to an agreement not only by requiring formalities but in many other ways as well. Thus rules respecting deceit, fraud, and undue influence are designed to ensure that contractual obligations are assumed freely and without one party misleading the other. Other rules regulate the modification of ongoing contractual relations with a view to preventing a party with considerable bargaining power —a building contractor, for example—from unfairly imposing changes in the contract.

The law also allows contractual relations to be adjusted when they have been thrown out of balance by unforeseen circumstances. The task of adjustment is relatively easy in cases in which both parties made a mistake or in which one party laboured under a mistaken assumption that was, or plainly should have been, known to the other. The problem of mistake becomes more intractable when the error is chargeable to only one party. The solutions reached for such situations are complex and defy general statement.

Catastrophic events such as inflation, political upheaval, or natural disasters may upset the economy of a contract. In the case of natural catastrophes, relief is frequently available under theories of force majeure and "act of God." When the unsettling circumstances are economic in their nature, as with severe inflation or deflation, a solution is difficult to find. A party who benefits from inflation in one contractual or economic relation may suffer from it in another. A general readjustment in contracts would be enormously complicated and time-consuming and would interject an undesirable element of uncertainty into economic and business activity. Only under exceptional circumstances—and usually in the form of special legislation—does the law seek to adjust contractual relations to take into account the effects of severe economic dislocations.

Sanctions for non-performance

Failure to perform. Another branch of contract law deals with the sanctions that are made available to a contracting party when the other party fails to perform his contractual obligations. When these sanctions take the form of money damages—as they usually do in practice, even though some civil-law systems have a theoretical preference for specific relief—the system must decide whether the plaintiff is to be put in the same position economically that he would have been in had the contract been performed (so-called expectation damages) or simply reimbursed for the actual losses, if any, flowing from his reliance on the contract (reliance damages). Reliance damages can, of course, be very large. A subcontractor who fails to deliver parts required for the construction of an ocean liner (or delivers faulty parts) may be responsible for heavy reliance damages resulting from delay in the work or actual damage to the vessel. Legal systems utilize various techniques to limit both reliance and expectancy damages when otherwise they would be unreasonably large.

If a person has agreed to buy an article from a merchant, his refusing to take delivery will not ordinarily produce substantial reliance damages. Delivery costs will have been incurred, but the merchant will presumably not have lost sales elsewhere. In such circumstances, the merchant will seek to recover not his delivery costs but his lost profit—his expectation damages. The law allows relief on the basis that the expectancy created by an enforceable promise has a current economic value, measured by the economic gain that the party would derive if the particular agreement were performed.

In some circumstances, performance is not measureable in terms of market value—as, for example, when one relative has agreed to sell to another a family painting of sentimental value, but of little intrinsic worth. Many legal systems in such a case require specific performance. The availability of specific relief varies among contemporary legal systems, for reasons that seem more historical and doctrinal than practical.

Other problems of contract law. Many contracts involve more than two persons. The law of contracts provides special rules for regulating claims by multiparty plaintiffs or claims against multiparty defendants, or for determining rights among the parties. Multiparty problems arise in other contexts as well. There is the problem of whether the immediate parties to a contract can enter into an agreement that will confer rights upon a person not an original party to the contract. Probably because the dogmatic structure of contract law was largely formed on the model of the simpler two-party situation, and because the contract for the benefit of third parties did not have great practical importance until such relatively modern developments as the emergence of life insurance, many systems of contract law have encountered difficulty in working out the relationship between the third party and the underlying contract. English law took the view that, as a rule, a person cannot acquire a right on a contract to which he is not a party. Some of the problems posed are difficult of resolution: under what circumstances and to what extent should the third party control the underlying contract when, for example, the original parties desire to rescind or modify it?

Another variation of the party problem is presented by efforts to add or substitute parties to a contract. In the absence of an express regulation of the problem in the basic contract, the law works with the notion of the presumed intention of the contracting parties, based on considerations of fairness and practicality. A contracting party cannot, in principle, assign to another his right under a contract if the assignment would result in a significant change in the burden assumed by the other contracting party. A contractual right to receive money or goods is a different matter; it can ordinarily be assigned because the resulting burden on the person under obligation is not great, and because society as a whole benefits from having this flexible economic and legal instrument.

One pervasive problem of contract law that has been mentioned above deserves further consideration—the problem of interpretation. Many rules of contract law are simply presumptions, based on experience and tradition, as to what the parties ordinarily intend; if they clearly intend otherwise, the rules are not mandatory. Problems of interpretation frequently arise with respect to the particularities of a given agreement, and here the court seeks to determine what the parties actually had in mind. The effort to ascertain intention may encounter difficulties arising from the law of evidence. Many legal systems limit the use of testimonial evidence to explain a contract the essential elements of which have been stated by the parties in a written form.

CONTEMPORARY TENDENCIES

Arbitration. Modern commercial practice relies to a growing extent on arbitration to handle disputes, espe-

cially those that arise in international transactions. There are several reasons for the growing use of arbitration. The procedure is simple, it is more expeditious, and it may be less expensive than traditional litigation. The arbitrators are frequently selected by a trade association or business group for their expert understanding of the issues in the dispute. The proceedings are private, which is advantageous when the case involves trade or business secrets. In many legal systems, the parties can authorize arbitrators to base their decision on equitable considerations that the law excludes. Finally, when the parties are from different countries, an international panel of arbitrators may offer a greater guarantee of impartiality than would a national court. Despite these advantages of arbitration, the development of contract law may suffer considerably by a withdrawal from the courts of litigation involving some of the most significant and difficult problems of the present day, all the more so because the reasoning in arbitral awards is usually not made public.

Attempts at codification. Trade and commerce flow increasingly across national and state boundaries. In response to this there have been many efforts to unify the traditional legal systems. In the United States, the Uniform Commercial Code has replaced earlier uniform statutes such as the Sales Act and the Negotiable Instruments Law; by 1970 it had been adopted by every state except Louisiana. Internationally, the decade of the 1960s saw significant progress toward uniform regulation of the law of sales. The creation of a uniform body of substantive rules is, of course, easiest when the communities involved have roughly similar rules and principles. In addition, the greater the volume of multistate transactions, the greater is the pressure for uniform regulation. It is understandably easier to achieve a Uniform Commercial Code within the United States than to create such a system internationally.

When a transaction has a significant relationship with more than one legal order, difficult problems of private international law often arise with respect to which law shall govern. A kind of halfway house between legal diversity and unification—the creation of uniform rules for choice of law—is of some help, and in this area the Hague Conference on Private International Law has done significant work (see INTERNATIONAL LAW).

BIBLIOGRAPHY. A useful survey of the development of contract law from its beginnings in primitive societies is E.A. FARNSWORTH, "The Past of Promise: An Historical Introduction to Contract," *Columbia Law Review*, 69:576–607 (1969). An historical treatment of the common law of contract is C.H.S. FIFOOT, *History and Sources of the Common Law: Tort and Contract* (1949). Treatises on Anglo-American contract law include: G.C. CHESHIRE and C.H.S. FIFOOT, *The Law of Contract*, 7th ed. (1969); A.L. CORBIN, *On Contracts*, 8 vol. (1950–51); and S. WILLISTON, *A Treatise on the Law of Contracts*, 8 vol., rev. ed. by the author and G.J. THOMPSON (1936). An extensive discussion of French and German contract law may be found in A.T. VON MEHREN, *The Civil Law System* (1957). A classic discussion of the law of contract damages is L.L. FULLER and W.R. PERDUE, JR., "The Reliance Interest in Contract Damages," *Yale Law Journal*, 46:52–96, 373–420 (1936–37).

(A.T.v.M.)

Control Systems

A control system is a means by which a variable quantity or set of variable quantities is made to conform to a prescribed norm. In this broad sense, the bodies of living organisms contain numerous natural control systems, such as body temperature and acidity. Like these, the control systems in modern technology, to which this article is confined, may either hold the values of the controlled quantities constant or may cause them to vary in a prescribed way.

A control system may be operated by electricity, by mechanical means, by fluid pressure (liquid or gas), or by a combination of means. When a computer is involved in the control circuit, it is usually more convenient to operate all of the control systems electrically, although intermixtures are fairly common.

Control systems are intimately related to the concept of automation (*q.v.*), but the two fundamental types of control systems, feedforward and feedback, have classic ancestry. The loom invented by Joseph Jacquard of France in 1801 is an early example of feedforward; a set of punched cards programmed the patterns woven by the loom; no information from the process was used to correct the machine's operation. Similar feedforward control was incorporated in a number of machine tools invented in the 19th century, in which a cutting tool followed the shape of a model.

Feed forward and feedback

Feedback control, in which information from the process is used to correct a machine's operation, has an even older history. Roman engineers maintained water levels for their aqueduct system by means of floating valves, which opened and closed at appropriate levels. The Dutch windmill of the 17th century was kept facing the wind by the action of an auxiliary fan that moved the entire upper part of the mill. The most famous example from the Industrial Revolution is James Watt's flyball governor of 1769, a device that regulated steam flow to a steam engine to maintain constant engine speed despite a changing load (see also STEAM POWER).

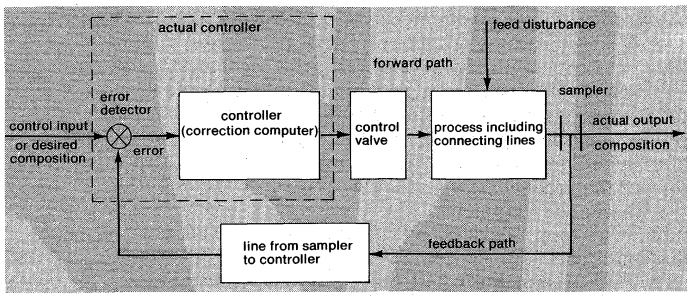
The first theoretical analysis of a control system, which presented a differential-equation model of the Watt governor, was published by James Clerk Maxwell, the Scottish physicist, in the 19th century. Maxwell's work was soon generalized and control theory developed by a number of contributions, including a notable study of the automatic steering system of the United States battleship "New Mexico," published in 1922. The 1930s saw the development of electrical feedback in long-distance telephone amplifiers and of the general theory of the servomechanism, by which a small amount of power controls a very large amount and makes automatic corrections. The pneumatic controller, basic to the development of early automated systems in the chemical and petroleum industries, and the analogue computer followed. All of these developments formed the basis for elaboration of control-system theory and applications during World War II, such as anti-aircraft batteries and fire-control systems.

Most of the theoretical studies as well as the practical systems up to World War II were single loop; that is, they involved merely feedback from a single point and correction from a single point. In the 1950s, the potential of multiple-loop systems came under investigation; in these systems feedback could be initiated at more than one point in a process and corrections made from more than one point. The introduction of analogue and digital-computing equipment opened the way for much greater complexity in automatic-control theory, an advance since labelled "modern control" to distiguish it from the older, simpler, "classical control."

BASIC PRINCIPLES

With few and relatively unimportant exceptions, all the modern control systems have two fundamental characteristics in common. These can be described, in somewhat oversimplified form, as follows: (1) The value of the controlled quantity is varied by a motor (this word being used in a generalized sense), which draws its power from a local source rather than from an incoming signal. Thus there is available a large amount of power to effect necessary variations of the controlled quantity and to ensure that the operations of varying the controlled quantity do not load and distort the signals on which the accuracy of the control depends. (2) The rate at which energy is fed to the motor, to effect variations in the value of the controlled quantity is determined more or less directly by some function of the difference between the actual and desired values of the controlled quantity. Thus, for example, in the case of a thermostatic heating system, the supply of fuel to the furnace is determined by whether the actual temperature is higher or lower than the desired temperature. A control system possessing these fundamental characteristics is called a closed-loop control system, or a servomechanism (see illustration). Open-loop control systems are feedforward systems.

Stability and transient response. The stability of a control system is determined to a large extent by its re-



Block diagram showing the essential components of a typical closed-loop control system.

From *McGraw-Hill Yearbook of Science and Technology*. Copyright 1963 by McGraw-Hill Book Company. Used with permission of McGraw-Hill Book Company.

sponse to a suddenly applied signal, or transient. If such a signal causes the system to overcorrect itself, a phenomenon called hunting may occur in which the system first overcorrects itself in one direction and then overcorrects itself in the opposite direction. Because hunting is undesirable, measures are usually taken to correct it.

Damping

The most common corrective measure is the addition of damping somewhere in the system. Damping slows down system response and avoids excessive overshoots or overcorrections. If the damping is such that the correction produces exactly the desired result without any overshooting or undershooting, the system is said to be critically damped. If damping is such that it takes the system an excessively long time to reach equilibrium, the system is overdamped. If the system oscillates a few cycles around the desired condition before equilibrium is reached, the system is underdamped. Damping can be in the form of electrical resistance in an electronic circuit, the application of a brake in a mechanical circuit, or forcing oil through a small orifice as in shock-absorber damping.

The critically damped system is the most desirable for the control engineer to obtain. Rather than risk obtaining an overdamped system with resulting system sluggishness, a slightly underdamped condition is usually designed into the system. A common criterion is that the first overshoot must not exceed the desired equilibrium value by more than 20 percent, and succeeding overshoots and undershoots must die out very rapidly.

Frequency response. Another method of ascertaining the stability of a control system is to determine its frequency response—that is, its response to a continuously varying input signal at various frequencies. The output of the control system is then compared to the input with respect to amplitude and to phase—that is, the degree with which the input and output signals are out of step.

Frequency response can either be determined experimentally—especially in electrical systems—or can be calculated mathematically if the constants of the system are known. Mathematical calculations are particularly useful for systems that can be described by ordinary linear differential equations. Graphic shortcuts also help greatly in the study of system responses.

Other factors. Several other techniques enter into the design of advanced control systems. Adaptive control is

Adaptive control

the capability of the system to modify its own operation to achieve the best possible mode of operation. A general definition of adaptive control implies that an adaptive system must be capable of performing the following functions: providing continuous information about the present state of the system or identifying the process; comparing present system performance to the desired or optimum performance and making a decision to change the system to achieve some previously defined optimum performance; and initiating a proper modification to drive the control system to the optimum. These three principles—identification, decision, and modification—are inherent in any adaptive system.

Dynamic-optimizing control requires the control system to operate in such a way that a specific performance criterion is satisfied. This criterion is usually formulated in such terms that the controlled system must move from the original to a new position in the minimum possible time or at minimum total cost.

Learning control implies that the control system contains sufficient computational ability so that it can develop representations of the mathematical model of the system being controlled and can modify its own operation to compensate for this newly developed knowledge. Thus, the learning control system is a further development of the adaptive controller.

Multivariable-noninteracting control involves large systems in which the size of internal variables is dependent upon the values of other related variables of the process. Thus the single-loop techniques of classical control theory will not suffice. More sophisticated techniques must be used to develop appropriate control systems for such processes.

CURRENT PRACTICE

There are many cases in industrial control practice in which theoretical automatic control methods are not yet sufficiently advanced to design an automatic control system or completely to predict its effects. This situation is true of the very large, highly interconnected systems such as occur in many industrial plants. In this case, operations research (*q.v.*), a mathematical technique for evaluating various possible procedures in a given situation, can be of value.

Conventional control. In determining the actual physical control system to be installed in an industrial plant, the instrumentation or control-system engineer has a wide range of possible equipment and methods to use. He may choose to use a set of analogue-type instruments, those that use a continuously varying physical representation of the signal involved—*i.e.*, a current, voltage, or an air pressure. Devices built to handle such signals, genrally called conventional devices, are usually capable of receiving only one input signal and one resulting output correction. Hence they are generally considered single-loop systems, and the total control system is built up of a collection of such devices. Analogue-type computers are available that can consider several variables at once for more complex control functions. These are usually very specific in their applications, however, and thus are not commonly used.

The number of control devices added to an industrial plant may vary widely from plant to plant. They may comprise only a very few instruments that are used mainly as indicators of plant-operating conditions. The operator is thus made aware of off-normal conditions and he himself manually adjusts such plant operational devices as valves and speed regulators in order to maintain control. On the other hand, there may be devices of sufficient quantity and complexity so that nearly all the possible occurrences may be covered by a control-system action ensuring automatic control of any forseeable failure or upset and thus making possible unattended control of the process. In the early 1970s the control systems of most industrial processes were somewhere between these two extremes. They are, however, tending toward the latter as more and more becomes known of the requirements for control of any particular process.

Computer control. With the development of very capable and very reliable models in the late 1960s, digital computers quickly became popular elements of industrial-plant-control systems. Computers (*q.v.*) are applied to industrial control problems in three ways: for supervisory or optimizing control; direct digital control; and hierarchy control.

Supervisory or optimizing control. In supervisory or optimizing control the computer operates in an external or secondary capacity, changing the set points in the primary plant-control system either directly or through manual intervention. A chemical process, for example, may take place in a vat the temperature of which is thermostatically regulated. For various reasons, the supervisory control system might intervene to reset the thermostat to a different level. The task of supervisory control is thus to "trim" the plant operation, thereby lowering costs or increasing production. Though the

overall potential for gain from supervisory control is sharply limited, a malfunction of the computer cannot adversely affect the plant.

Direct-digital control. In direct-digital control a single digital computer replaces a group of single-loop analogue controllers. Its greater computational ability makes the substitution possible, with obvious cost savings, and also permits the application of more complex advanced-control techniques.

Hierarchy control. Hierarchy control attempts to apply computers to all the plant-control situations simultaneously. As such, it requires the most advanced computers and most sophisticated automatic-control devices to integrate the plant operation at every level from top-management decision to the movement of a valve. Though most work in this area was still in the development stages, in the early 1970s several partial installations were already operating.

Advantages of digital computers

The advantage offered by the digital computer over the conventional control system described earlier, costs being equal, is that the computer can be programmed readily to carry out a wide variety of separate tasks. In addition, it is fairly easy to change the program so as to carry out a new or revised set of tasks should the nature of the process change or the previously proposed system prove to be inadequate for the proposed task. With digital computers, this can usually be done with no change to the physical equipment of the control system. For the conventional control case, at least, some of the physical hardware apparatus of the control system must be replaced in order to achieve new functions or new implementations of them.

Instrumentation and control in process industries. Two major problems impede adoption of the computer-controlled process: lack of knowledge about particular processes and the theoretical relationship among all the variables involved, and lack of instrumentation for continuously measuring some of the process variables.

Continuous quantitative measurements

Continuous quantitative measurement of several physical conditions can now be made—*e.g.*, temperature, pressure, level, weight, flow, and volume, together with viscosity, moisture content, conductivity, pH, liquid density, refractive index, specific gravity, and, in some cases, chemical composition (see also INSTRUMENTATION).

Much remains to be done. Not all the analyses can be undertaken continuously; few analytical techniques are fast enough to apply in a continuous control system, and reliability is still generally not high enough. Reliability is of prime importance because it is costly both to interrupt the process and to provide skilled maintenance. Another major problem preventing the widespread use of process analyzers is the difficulty of securing a representative sample of a process stream in various type of plants.

As plants become increasingly complex and more and more process variables are measured, the necessity for checking a large number of measured and controlled values makes some form of data reduction necessary and automatic scanning desirable. Much of the information gathered is of use only for research or during periods of disturbance or emergency. Data reduction and automatic monitoring and scanning of process variables are being used to an increasing extent. Such equipment in a power station will automatically and continuously scan and record at very high speeds the readings of instruments measuring flows, levels, pressures, and temperatures of boilers, furnaces, and turbine units. The system will print out the relevant data at regular intervals and will record and give warnings of any deviations from normal in the readings. It will also compute and record continuously the overall boiler efficiency.

Control in manufacturing. Control systems are also becoming a major component of the automation of production lines in modern factories. Automation (*q.v.*) began in the late 1940s with the development of the transfer machine, a mechanical device for moving and positioning large objects on a production line (such as partly finished automobile engine blocks). These early machines had no feedback control as described above. Instead, manual intervention was required for any final adjustment of position or other corrective action necessary. Because of their large size and cost, long production runs were necessary to justify the use of transfer machines.

The need to reduce the high labour content of manufactured goods, the requirement to handle much smaller production runs, the desire to gain increased accuracy of manufacture, combined with the need for sophisticated tests of the product during manufacture, have resulted in the recent development of computerized production monitors, testing devices, and feedback-controlled production robots.

The "programmability" of the computer to handle a wide range of tasks along with the capability of rapid change to a new program has made it invaluable for these purposes. Similarly, the need to compensate for the effect of tool wear and other variations in automatic machining operations has required the institution of a feedback control of tool positioning and cutting rate in place of the formerly used direct mechanical motion. Again, the result is a more accurately finished final product with less chance for tool or manufacturing machine damage.

BIBLIOGRAPHY. The first major textbook on digital computer control was that of E.S. SAVAS *et al.*, *Computer Control of Industrial Processes* (1965). The field of direct digital control is summarized in T.J. WILLIAMS and F.M. RYAN (eds.), *Progress in Direct Digital Control* (1969). The science of automatic control originated with the study of the design of feedback amplifiers and filters for telephone networks. Two basic and classic papers in this area are H. NYQUIST, "Regeneration Theory," *Bell Syst. Tech. J.*, 11:126–147 (1932); and H.W. BODE, "Relations Between Attenuation and Phase in Feedback Amplifier Design," *Bell Syst. Tech. J.*, 19:421–454 (1940). These contributions, along with W.R. EVANS, "Control System Analysis by Root Locus Method," *Trans. Am. Inst. Elec. Engrs.*, 69:66–69 (1950), formed the basis of the so-called classical design procedures—the "Nyquist Diagram," the "Bode Plot," and the "Root Locus Method." During the late 1940s and the 1950s, a number of excellent textbooks appeared. Important among these were those of G.S. BROWN and D.P. CAMPBELL, *Principles of Servomechanisms* (1948); HAROLD CHESTNUT and R.W. MAYER, *Servomechanisms and Regulating System Design*, vol. 1 (1951); and the classic by J.G. TRUXAL, *Automatic Feedback Control System Synthesis* (1955). The first textbook in industrial-process control was R.C. OLDENBOURG and H. SARTORIUS, *Dynamik selbsttätiger regelungen* (1944; Eng. trans., *The Dynamics of Automatic Control*, 1948). The "Era of Modern Control" can be traced to the writings of NORBERT WIENER, of which *Cybernetics* (1948), is especially well known; R.E. BELLMAN, *Dynamic Programming* (1957); L.S. PONTRYAGIN *et al.*, *The Mathematical Theory of Optical Processes* (Eng. trans., 1962); and especially in the United States to R.E. KALMAN. His "On the General Theory of Control Systems," in J.F. COALES (ed.), *Automatic and Remote Control: Proceedings, First International Congress of the International Federation of Automatic Control*, vol. 1, pp. 481–492 (1961), is one of the most important of his many papers. Much of the early interest in the subject of process dynamics can be traced to the work of DONALD P. CAMPBELL, *Process Dynamics* (1958); and to the Conference on Automatic Control, the papers of which were edited by ARNOLD TUSTIN in *Automatic and Manual Control* (1952).

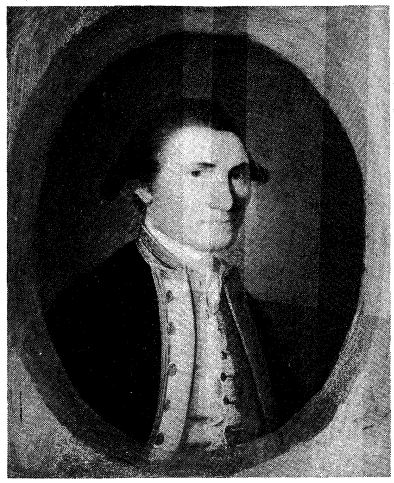
(T.J.W.)

Cook, James

As an outstanding sea captain, navigator, cartographer, and practical dietician, James Cook, on his three world-circling voyages, discovered more about the Pacific, the South Atlantic, the southern Indian, and the Arctic and Antarctic oceans than had been seen or imagined by all other navigators during the preceding two and one-half centuries. He defeated scurvy, established precise navigation as a common sea skill, and changed the face of the Pacific from the world's immense abode of myths and mystery to the map men know today.

Early life

James Cook was born October 27, 1728, at the village of Marton in Cleveland, Yorkshire, the son of a farm-hand migrant from Scotland. While Cook was still a child, his father became the foreman on a farm in a neighbouring village. Young James early showed signs of an inquiring and able mind, and his father's employer paid for his schooling in the village until he was 12 years old. His early teens were spent on the farm where his father worked, but a brief apprenticeship in a general



Cook, oil painting by J. Webber (1752–93). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

store in a coastal village north of Whitby brought him in contact with ships and the sea. At the age of 18, in 1746, he was apprenticed to a well-known Quaker shipowner, John Walker of Whitby, and at 21 was rated able seaman in the Walker collier-barks—stout, seaworthy, slow, 300- and 400-tonners mainly in the North Sea trade. When the ships were laid up for refitting (done by the apprentices and crews) at Whitby during the worst months of winter, Cook lived ashore and studied mathematics by night. The Whitby barks, constantly working North Sea waters off a dangerous and ill-marked lee shore, offered Cook splendid practical training: the young man who learned his seamanship there had little to fear from any other sea.

Promoted to mate in 1752, Cook was offered command of a bark three years later, after eight years at sea. Advancement of this nature opened up a career that would have satisfied most working seamen, but instead Cook volunteered as able seaman in the Royal Navy. The navy, he was sure, offered a more interesting career for the competent professional seaman, and greater opportunity than in the North Sea barks. Tall, of striking appearance, Cook almost immediately caught the attention of his superiors, and with excellent power of command, he was marked for rapid advancement.

Actions in the Seven Years' War. After advancing to master's mate, and boatswain, both noncommissioned ranks, he was made master of HMS "Pembroke" at the age of 29. During the Seven Years' War between Great Britain and France (1756–63) he saw action in the Bay of Biscay, was given command of a captured ship, and took part in the siege of Louisburg in Nova Scotia and in the successful amphibious assault against Quebec. His charting and marking of the more difficult reaches of the St. Lawrence River contributed to the success of General Wolfe's landing there. Based at Halifax during the winters, he mastered surveying with the plane table. Between 1763 and 1768, after the war had ended, he commanded the schooner "Grenville" while surveying the coasts of Newfoundland, sailing most of the year and working on his charts at his base in England during the winters. In 1766 he observed an eclipse of the sun and sent the details to the Royal Society in London—an unusual activity for a noncommissioned officer, for Cook still rated only as master.

Charting the St. Lawrence

Voyages and discoveries. In 1768 the Royal Society, in conjunction with the Admiralty, was organizing the first scientific expedition to the Pacific, and the rather obscure 40-year-old James Cook was appointed commander of the expedition. Hurriedly commissioned as lieutenant, he was given a homely looking but extremely sturdy Whitby coal-hauling bark renamed HMS "Endeavour," then four years old, of just 368 tons, and less than 98 feet long. Cook's orders were to convey gentlemen of the Royal Society and their assistants to Tahiti to observe the transit of the planet Venus across the sun. That done, on June 3, 1769, he was to find the southern continent, the so-called Terra Australis, which philosophers argued must exist to balance the landmasses of the Northern Hemisphere. The leader of the scientists was the rich and able Joseph Banks, aged 26, who was assisted by Daniel Solander, a Swedish botanist, as well as astronomers (Cook rating as one) and artists. Cook carried an early nautical almanac and brass sextants, but no chronometer on the first voyage.

His success is history. Striking south and southwest from Tahiti, where his predecessors had sailed west and west-northwest with the favouring trade winds, Cook found and charted all of New Zealand, a difficult job that took six months. After that, instead of turning before the west winds for the homeward run around Cape Horn, he crossed the Tasman Sea westward and, on April 19, 1770, came on the southeast coast of Australia. Running north along its 2,000-mile eastern coast, surveying as he went, Cook successfully navigated Queensland's Great Barrier Reef—since reckoned as one of the greatest navigational hazards in the world—taking the Coral Sea and the Torres Strait in his stride. Once the bark touched on a coral spur by night, but it withstood the impact and was refloated. After the "Endeavour" was grounded on the nearby Queensland coast and repaired, Cook sailed it back to England. He stopped briefly at Batavia (modern Djakarta) for supplies, and although the crew had been remarkably healthy until then, 30 died of fever and dysentery contracted while on land. None of the crew, however, died of scurvy (a dietary disease caused by a lack of ascorbic acid and that habitually decimated the crews of ships on lengthy voyages in the 18th century). This was because, in addition to ensuring cleanliness and ventilation in the crew's quarters, Cook insisted on an appropriate diet that included cress, sauerkraut, and a kind of orange extract. The health in which he maintained his sailors in consequence made his name a naval byword.

Coasting the Great Barrier Reef

Back in England, he was promoted to commander, presented to King George III, and soon began to organize another and even more ambitious voyage. The success of the expedition of Joseph Banks and his scientists (which established the useful principle of sending scientists on naval voyages—*e.g.*, Charles Darwin in the "Beagle," T.H. Huxley in the "Rattlesnake," and J.D. Hooker with Sir James Ross to the Ross Sea in the Antarctic) stimulated interest not only in the discovery of new lands but in the new knowledge in many other scientific subjects. The wealth of scientifically collected material from the "Endeavour" voyage was unique. Cook was now sent out with two ships to make the first circumnavigation of and penetration into the Antarctic.

Second expedition, 1772–75

Between July 1772 and July 1775 Cook made what ranks as one of the greatest sailing ship voyages, again with a small former Whitby ship, the "Resolution," and a consort ship, the "Adventure." He found no trace of Terra Australis, though he sailed beyond latitude 70° S in the Antarctic, but he successfully completed the first west–east circumnavigation in high latitudes, charted Tonga and Easter Island during the winters, discovered New Caledonia in the Pacific, and the South Sandwich Islands and South Georgia Island in the Atlantic. He showed that a real Terra Australis existed only in the landmasses of Australia, New Zealand, and whatever land might remain frozen beyond the ice rim of Antarctica. And, once again, not one of his crew died of scurvy. Back in England, he was promoted to captain at last, elected a fellow of the Royal Society, and awarded one of its highest honours, the gold Copley Medal, for a paper he prepared on his work against scurvy.

The last voyage. There was yet one secret of the Pacific to be discovered: whether there existed a northwest passage around Canada and Alaska or a northeast one around Siberia, between the Atlantic and Pacific. As the passages had long been sought in vain from Europe, it was thought that the search from the North Pacific might be successful. The man to undertake the search obviously was Cook, and in July 1776 he went off again on the "Resolution" with another Whitby ship, the "Discovery."



The voyages of Captain Cook

This search was unsuccessful, for neither a northwest nor northeast passage usable by sailing ships existed, and the voyage led to Cook's death. On February 14, 1779, at the age of 50, in a brief fracas with Hawaiians over the stealing of a cutter, Cook was slain on the beach at Kealakekua by the Polynesian natives.

Cook's voyaging left him comparatively little time for family life. Although Cook had married Elizabeth Batts in 1762, when he was 34 years old, he was at sea for more than half of their married life. The couple had six children, three of whom died in infancy. The three surviving sons, two of whom entered the navy, had all died by 1794.

Cook had set new standards of thoroughness in discovery and seamanship, in navigation, cartography, and the sea care of men, in relations with natives both hostile and docile, in the application of science at sea: and he had peacefully changed the map of the world more than any other single man in history.

BIBLIOGRAPHY. *The Journals of Captain Cook*, ed. by J.C. BEAGLEHOLE, 4 vol. (1955–67), the standard work on the three voyages, presenting Cook's own text for the first time, thoroughly annotated and explained, and containing a separate portfolio with reproductions of charts and views drawn on these voyages; J.C. BEAGLEHOLE, *The Exploration of the Pacific*, 3rd ed. (1966), provides the background to Cook's voyages; JAMES A. WILLIAMSON, *Cook and the Opening of the Pacific* (1946), useful for background; CHARLES DE BROSSES, *Histoire des navigations aux terres australes* (1756); JAMES BURNEY, *A Chronological History of the Discoveries in the South Sea or Pacific Ocean*, 5 vol. (1803–17); J.C. BEAGLEHOLE (ed.), *The Endeavour Journal of Joseph Banks*, 2nd ed. (1963), useful for Banks's contribution and his influence and life; ARTHUR KITSON, *Captain James Cook* (1907); HUGH CARRINGTON, *Life of Captain Cook* (1939); ALAN J. VILLIERS, *Captain James Cook* (British title, *Captain Cook, the Seamen's Seaman*, 1967); G.M. BADGER (ed.), *Captain Cook, Navigator and Scientist* (1969), papers presented at the Cook Bicentenary Symposium, Australian Academy of Science, Canberra.

(A.J.V.)

Cooper, James Fenimore

The first major American novelist, James Fenimore Cooper established a permanent place for himself in literary history as the virtual inventor of the sea romance and the frontier adventure novel.

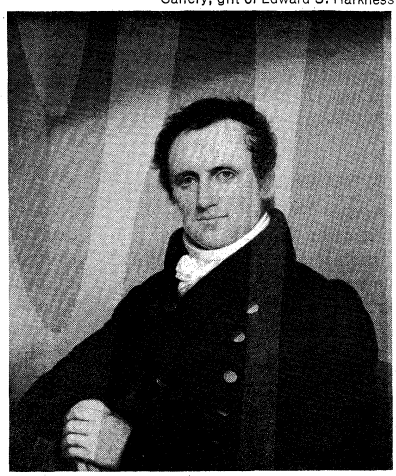
Cooper was born on September 15, 1789, at Burlington, New Jersey. His mother, Elizabeth Fenimore, was a member of a respectable New Jersey Quaker family; and his father, William, later founded a frontier settlement at the source of the Susquehanna River (now Cooperstown, New York) and served as a Federalist congressman during the administrations of George Washington and John Adams. It was a most appropriate family background for a writer who, by the time of his death, was generally considered America's "national novelist."

Early years. James was but a year old when William Cooper moved his family to the primitive settlement in upstate New York. He was doubtless fortunate to be the 11th of 12 children, for he was spared the worst hardships of frontier life while he was able to benefit educationally both from the rich oral traditions of his family and a material prosperity that afforded him a gentleman's education. After private schooling in Albany, he attended Yale from 1803 to 1805. Little is known of his college career other than that he was the best Latin scholar of his class and was expelled in his junior year because of a prank. Since high spirits seemed to fit him for an active life, his family allowed him to join the Navy as a midshipman. But prolonged shore duty at several New York

stations merely substituted naval for academic discipline. His father's death in 1809 left him financially independent, and in 1811 he married Susan De Lancy and resigned from the Navy.

For ten years after his marriage he led the active but unproductive life of a dilettante, dabbling in agriculture, politics, the American Bible Society, and the Westchester militia. It was in this amateur spirit that he wrote and published his first fiction, reputedly on a challenge from his wife. *Precaution* (1820) was a plodding imitation of Jane Austen's novels of English gentry manners. It is mainly interesting today as a document in the history of American cultural colonialism and as an example of a clumsy attempt to imitate Jane Austen's investigation of the ironic discrepancy between illusion and reality. His second novel, *The Spy* (1821), was based on another British model, Sir Walter Scott's "Waverley" novels, stories of adventure and romance set in 17th- and 18th-century Scotland. But in *The Spy* Cooper broke new ground by using an American Revolutionary War setting (based partly on the experiences of his wife's British Loyalist family) and by introducing several distinctively American character types. Like Scott's novels of Scotland, *The Spy* is a drama of conflicting loyalties and interests in which the action mirrors and expresses more subtle internal psychological tensions. *The Spy* soon brought him international fame and a certain amount of wealth. The latter was very welcome, indeed necessary, since his father's estate had proved less ample than anybody had thought, and, with the death of his elder brothers, he had found himself responsible for the debts and widows of the entire Cooper family.

By courtesy of Yale University Art Gallery, gift of Edward S. Harkness



James Fenimore Cooper, oil painting by John Wesley Jarvis (1780–1840). In Yale University Art Gallery.

The "Leatherstocking" tales

Novels. The first of the famous "Leatherstocking" tales, *The Pioneers* (1823), followed and adhered to the successful formula of *The Spy*, reproducing its basic thematic conflicts and utilizing family traditions once again. In *The Pioneers*, however, the traditions were those of William Cooper of Cooperstown, who appears as Judge Temple of Templeton, along with many other lightly disguised inhabitants of James's boyhood village. No known prototype exists, however, for the novel's principal character—the former wilderness scout Natty Bumppo, *alias* Leatherstocking. The Leatherstocking of *The Pioneers* is an aged man, of rough but sterling character, who ineffectually opposes "the march of progress," namely, the agricultural frontier and its chief spokesman, Judge Temple. Fundamentally, the conflict is between rival versions of the American Eden: the "God's Wilderness" of Leatherstocking and the cultivated garden of Judge Temple. Since Cooper himself was deeply attracted to both ideals, he was able to create a powerful and moving story of frontier life. Indeed, *The Pioneers* is both the first and finest detailed portrait of frontier life in American literature; it is also the first truly original American novel.

Both Cooper and his public were fascinated by the Leatherstocking character. He was encouraged to write a series of sequels in which the entire life of the frontier scout was gradually unfolded. *The Last of the Mohicans* (1826) takes the reader back to the French and Indian wars of Natty's middle age, when he is at the height of his powers. That work was succeeded by *The Prairie* (1827) in which, now very old and philosophical, Leatherstocking dies, facing the westering sun he has so long followed. (The five novels of the series were not written in their narrative order.) Identified from the start with the vanishing wilderness and its natives, Leatherstocking was an unalterably elegiac figure, wifeless and childless, hauntingly loyal to a lost cause. This conception of the character was not fully realized in *The Pioneers*, however, because Cooper's main concern with depicting frontier life led him to endow Leatherstocking with some comic traits and make his laments, at times, little more than whines or grumbles. But in these sequels Cooper retreated stylistically from a realistic picture of the frontier in order to portray a more idyllic and romantic wilderness; by doing so he could exploit the parallels between the American Indians and the forlorn Celtic heroes of James Macpherson's pseudo-epic *Ossian*, leaving Leatherstocking intact but slightly idealized and making extensive use of Macpherson's imagery and rhetoric.

Cooper intended to bury Leatherstocking in *The Prairie*, but many years later he resuscitated the character and portrayed his early maturity in *The Pathfinder* (1840) and his youth in *The Deerslayer* (1841). These novels, in which Natty becomes the centre of romantic interest for the first time, carry the idealization process further. In *The Pathfinder* he is explicitly described as an American Adam, while in *The Deerslayer* he demonstrates his fitness as a warrior-saint by passing a series of moral trials and revealing a keen, though untutored, aesthetic sensibility.

The "Leatherstocking" tales are Cooper's great imperfect masterpiece; but he continued to write many other volumes of fiction and nonfication. His fourth novel, *The Pilot* (1823), inaugurated a series of sea novels, which were at once as popular and influential as the "Leatherstocking" tales. And they were more authentic: such Westerners as General Lewis Cass, governor of Michigan Territory, and Mark Twain might ridicule Cooper's woodcraft, but old salts like Herman Melville and Joseph Conrad rightly admired and learned from his sea stories, in particular, *The Red Rover* (1827) and *The Sea Lions* (1849). Never before in prose fiction had the sea become, not merely a theatre for but the principal actor in, moral drama that celebrated man's courage and skill at the same time that it revealed him humbled by the forces of God's nature. As developed by Cooper, and later by Melville, the sea novel became a powerful vehicle for spiritual as well as moral exploration. Not satisfied with mere fictional treatment of life at sea, Cooper also wrote a meticulously researched, highly readable *History of the Navy of the United States of America* (1839).

The Pilot and the sea-novel series

Cultural and political involvement. Though most famous as a prolific novelist, he did not simply retire to his study after the success of *The Spy*. Between 1822 and 1826 he lived in New York City and participated in its intellectual life, founding the Bread and Cheese Club, which included such members as the poets Fitz-Greene Halleck and William Cullen Bryant, the painter and inventor Samuel F.B. Morse, and the great Federalist judge James Kent. Like Cooper himself, these were men active both in cultural and political affairs.

Cooper's own increasing liberalism was confirmed by a lengthy stay (1826–33) in Europe, where he moved for the education of his son and four daughters. Those years coincided with a period of revolutionary ferment in Europe; and because of a close friendship that he developed with the old American Revolutionary War hero Lafayette, he was kept well-informed about Europe's political developments. Through his novels, most notably *The Bravo* (1831), and other more openly polemical writings, he attacked the corruption and tyranny of oligarchical regimes in Europe. His active championship of the prin-

ciples of political democracy (though never of social egalitarianism) coincided with a steep decline in his literary popularity in America, which he attributed to a decline in democratic feeling among the reading—*i.e.* the propertied—classes to which he himself belonged.

Return to America

When he returned to America, he settled first in New York City and then for the remainder of his life in Cooperstown. In the gentlemanly tradition of Jefferson and Lafayette he attacked the oligarchical party of his day, in this case the Whig Party, which opposed Pres. Andrew Jackson, the exponent of a more egalitarian form of democracy. The Whigs, however, were soon able to turn the tables on Cooper and other leading Jacksonians by employing Jackson's egalitarian rhetoric against them. Squire Cooper had made himself especially vulnerable to popular feeling when, in 1837, he refused to let local citizens picnic on a family property known as Three Mile Point. This incident led to a whole series of charges of libel, and suits and countersuits by both the Whigs and Cooper. At this time, too, agrarian riots on the estates of his old New York friends shattered his simple Jeffersonian faith in the virtue of the American farmer. All of this conflict and unrest was hard to bear, and harder still because he was writing more and earning less as the years went by. The public, which had revelled in his early forest and sea romances, was not interested in his acute political treatise, *The American Democrat* (1838), or even in such political satires as *The Monikins* (1835) or *Home As Found* (1838). And though he wrote some of his best romances—particularly the later "Leatherstocking" tales and *Satanstoe* (1845)—during the last decade of his life, profits from publishing so diminished that he gained little benefit from improved popularity. Though his circumstances were never straitened, he had to go on writing; and some of the later novels, such as *Mercedes of Castile* (1840) or *Jack Tier* (1846–48), were mere hack work. His buoyant political optimism had largely given way to calm Christian faith, though he never lost his troubled concern for the well-being of his country. He died on September 14, 1851.

MAJOR WORKS

NOVELS: *Precaution* (1820); *The Spy* (1821); *The Pioneers* (1823); *The Pilot* (1823); *Lionel Lincoln* (1824); *The Last of the Mohicans* (1826); *The Prairie* (1827); *The Red Rover* (1827); *The Wept of Wish-ton-Wish* (1829); *The Water Witch; or, The Skimmer of the Seas* (1830); *The Bravo* (1831); *The Monikins* (1835); *Homeward Bound* (1838); *Home As Found* (1838); *The Pathfinder* (1840); *The Deerslayer* (1841); *The Two Admirals* (1842); *The Wing-and-Wing;* or *Le Feu-follet* (1842); *Wyandotté;* or, *The Hutted Knoll* (1843); *Ned Myers* (1843); *Afloat and Ashore* (1844); *Satanstoe;* or, *The Littlepage Manuscripts* (1845); *The Chain-Bearer* (1845); *The Redskins: Or, Indian and Injin* (1846); *The Crater: Or, Vulcan's Peak* (1847); *The Sea Lions* (1849); *The Ways of the Hour* (1850).

OTHER WORKS: *Notions of the Americans: Picked Up by a Travelling Bachelor* (1828); *A Letter to His Countrymen* (1834); *Sketches of Switzerland* (1836); *Gleanings in Europe* (1837); *The American Democrat* (1838); *The History of the Navy of the United States of America* (1839).

BIBLIOGRAPHY. R.E. SPILLER and P.C. BLACKBURN, *A Descriptive Bibliography of the Writings of James Fenimore Cooper* (1934), is now somewhat out of date but still standard. The major depository for Cooper's papers is the Beinecke Rare Book and Manuscript Library, Yale University. JAMES F. BEARD (ed.), *Letters and Journals of James Fenimore Cooper,* 6 vol. (1960–68), is a monument of precision and learning. Cooper's works have otherwise not been edited definitively. Important collected editions of his novels are *J. Fenimore Cooper's Works,* Household Edition, 32 vol. (1876–84), with introductions by Cooper's daughter Susan; and *The Works of James Fenimore Cooper,* Mohawk Edition (1895–1900). No collected edition is now in print, though his most popular individual works are available in paperback.

Until James F. Beard's promised critical biography of Cooper is published, the fullest and most accurate account is that contained in Beard's prefaces and notes to the *Letters and Journals.* JAMES GROSSMAN, *James Fenimore Cooper* (1949), is a sound and readable critical biography.

Important critical studies include: MARK TWAIN, "Fenimore Cooper's Literary Offences," *North American Review,* vol. 161 (1895), unfair, but shrewd and funny; D.H. LAWRENCE, "Fenimore Cooper's Leatherstocking Novels," *Studies in Classic American Literature* (1923), factually inaccurate but brilliantly suggestive; YVOR WINTERS, "Fenimore Cooper, or the Ruins of Time," *In Defense of Reason* (1947), first published in 1938, a masterly reappraisal of Cooper's life and writings; HENRY NASH SMITH, *Virgin Land* (1950), establishes Leatherstocking's place in the development of the mythology of the American West; MARIUS BEWLEY, *The Eccentric Design* (1959), includes a useful study of Cooper's social values and literary methods; THOMAS PHILBRICK, *James Fenimore Cooper and the Development of American Sea Fiction* (1961), a definitive study, notable for exhaustive scholarship and penetrating criticism; DONALD A. RINGE, *James Fenimore Cooper* (1962), an able critical survey of Cooper's fiction; KAY S. HOUSE, *Cooper's Americans* (1966), a pioneering study of Cooper's character types and methods of characterization; and GEORGE DEKKER, *James Fenimore Cooper* (1967), examines Cooper's achievement as a historical novelist.

(G.De.)

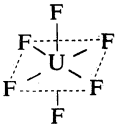
Coordination Compounds

Coordination compounds are substances with characteristic chemical structures in which a central metal atom is surrounded by nonmetallic atoms (or groups of atoms), called ligands, joined to it by chemical bonds. The class includes a number of important biological materials, such as vitamin B_{12} and hemoglobin, the red colouring matter of blood. It also includes a number of industrially important materials used as dyestuffs and pigments; as agents for extracting, purifying, and analyzing metals; and as catalysts for preparing such useful organic substances as the polyethylene plastics. Coordination compounds have been much studied because of what they reveal about molecular structure and chemical bonding, as well as because of the unusual chemical nature and the useful properties of certain coordination compounds.

The general class of coordination compounds—or complexes, as they are sometimes called—is extensive and diverse. The substances in the class may be composed of electrically neutral molecules or of positively or negatively charged species (ions). (The formula for a compound lists the kinds of atoms in the molecule and, in subscripts, the number of each atom. Water, for example, is composed of molecules that consist of two atoms of hydrogen and one of oxygen: H_2O. The electrical charge on an ion is represented by a superscript following the symbol of the element—*e.g.*, K^+ for the potassium ion—or following a group of atoms enclosed in brackets—*e.g.*, $[NH_4]^+$ for the ammonium ion.)

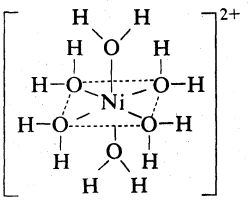
Neutral coordination compounds

Among the many coordination compounds having neutral molecules is uranium hexafluoride. The molecular formula of this compound, UF_6, indicates that six atoms of fluorine (F) are joined to one atom of uranium (U). The structural formula of the compound represents the actual arrangement of atoms in the molecules:



In this formula the solid lines, which represent bonds between atoms, show that four of the fluorine atoms are bonded to the single atom of uranium and lie in a plane with it, the plane being indicated by dotted lines (which do not represent bonds); whereas the remaining two fluorine atoms (also bonded to the uranium atom) lie above and below the plane, respectively.

An example of an ionic coordination complex is the hydrated (water-containing, with formula H_2O) ion of nickel, symbol Ni, molecular formula $[Ni(H_2O)_6]^{2+}$, the structure of which is

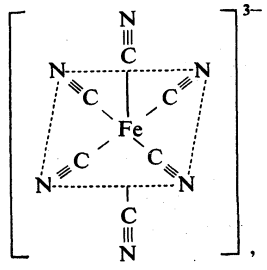


in which the symbols and lines are used as above and the brackets and the "two plus" (2+) sign show that the double positive charge goes with the whole unit.

The central metal atom in a coordination compound itself may be neutral or charged (ionic). The coordinated groups—or ligands, as they are usually called—may be neutral molecules such as water (in the above example), ammonia, or carbon monoxide; negatively charged ions (anions) such as fluoride (in the first example above) or cyanide ion; or, occasionally, positively charged ions (cations) such as hydrazinium or nitrosonium ion.

Complex ions

Complex ions—that is, the ionic members of the family of coordination substances—may exist as free ions in solution, or they may be incorporated into crystalline materials (salts) with other ions of opposite charge. In such salts, the complex ion may be either the cationic (positively charged) or the anionic (negatively charged) component (or, on occasion, both). The hydrated nickel ion (above) is an example of a cationic complex. An anionic complex is the hexacyanide (composed of carbon, C, and nitrogen, N) of ferric (symbol Fe) ion, $[Fe(CN)_6]^{3-}$, or



in which the symbols have the same general meaning as in the above examples. Crystalline salts containing these ions are potassium (symbol K) ferricyanide (with iron, Fe; carbon, C; and nitrogen, N), formula $K_3[Fe(CN)_6]$, and the hexahydrate of nickel chloride, formula

$$[Ni(H_2O)_6]Cl_2,$$

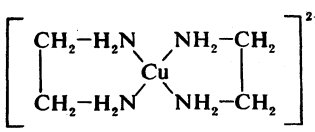
in each case the charge on the complex ion being neutralized by appropriate numbers of ions of opposite charge. In the case of potassium ferricyanide, three positively charged potassium ions, K^+, neutralize the charge on the complex, and in the nickel complex the charges are neutralized by negative chloride ions, Cl^-.

Each molecule or ion of a coordination compound includes a number of ligands, and, in any given substance, the ligands may be all alike or they may be different. Attachment of the ligands to the metal atom may be through only one bond or it may be through several bonds. When only one bond is involved, the ligand is said to be unidentate; when two are involved, it is bidentate; and so on. In general, ligands utilizing more than one bond are said to be multidentate. Because a multidentate ligand is joined to the metal atom in more than one place, the resulting complex is said to be cyclic—*i.e.*, to contain a ring of atoms. Coordination compounds containing multidentate ligands are called chelates (Greek *chele*, "claw"), and their formation is termed chelation. Chelates are particularly stable and useful (see below *Multidentate*). An example of a typical chelate is the complex formed between cupric ion (Cu^{2+}) and the organic compound ethylenediamine (formula $NH_2CH_2CH_2NH_2$). The formula of the complex is

Chelates and chelation

$$[Cu(NH_2CH_2CH_2NH_2)_2]^{2+},$$

the structural formula is



The total number of bonds from the ligands to the metal atom is called the coordination number; generally, coordination numbers range between 2 and 12, with 4 and 6 being the most common. The ligands surrounding the central atom often are referred to collectively as the coordination sphere.

The distinction between coordination compounds and other substances is, in fact, somewhat arbitrary. The designation coordination compound, however, is generally restricted to substances the molecules or ions of which are discrete entities and in which the central atom is metallic. Accordingly, molecules such as sulfur hexafluoride and carbon tetrafluoride are not normally considered to be coordination compounds because sulfur and carbon are nonmetallic elements. Yet there is no great difference between these compounds and, say, uranium hexafluoride. Furthermore, such simple ionic salts as sodium chloride or nickel difluoride are not considered to be coordination compounds because they consist of continuous ionic lattices rather than discrete molecules. Nevertheless, the arrangement (and bonding) of the anions surrounding the metal ions in these salts is similar to that in coordination compounds. Coordination compounds generally are characterized by a variety of distinctive physical and chemical properties, such as colour, magnetic susceptibility, solubility and volatility, an ability to undergo oxidation-reduction reactions, and catalytic activity (see below).

GENERAL ASPECTS

History. Among the earliest recorded examples of a coordination compound is the substance Prussian blue, containing iron (Fe), carbon (C), and nitrogen (N), with formula $Fe_4[Fe(CN)_6]_3$, which has been used as an artist's pigment since the beginning of the 18th century. Another early example of the preparation of a coordination compound is the use of a sparingly soluble compound, potassium hexachloroplatinate, formula $K_2[PtCl_6]$, in 1760 to refine the element platinum.

The sustained and systematic development of modern coordination chemistry, however, usually is considered to have begun with the discovery in 1798 that ammoniacal solutions of a cobalt chloride, $CoCl_3$, on standing overnight, deposited orange crystals with the composition $CoCl_3 \cdot 6NH_3$ (cobalt, Co; chlorine, Cl; nitrogen, N; ammonia, NH_3), the correct formulation of which is now recognized to be $[Co(NH_3)_6]Cl_3$; this shows that the six ammonia groups are associated with the cobalt atom and the whole is neutralized by three chlorine atoms. The feature of this observation that was particularly significant was the recognition that two independently stable compounds (*i.e.*, cobalt chloride and ammonia) could combine to form a new chemical compound with properties quite different from those of the constituent compounds.

Discoveries of further cobaltammine complexes and many other coordination compounds were made in the 19th century, but it was not until 1893 that Alfred Werner, a Swiss chemist, proposed the first essentially correct formulation of such compounds and provided the first insight into the nature of their bonding. Werner reformulated a series of platinum complexes and showed that their dissociation into ions in solution took place as shown in Table 1. In each case, the number of ions

Formulations of Werner

Table 1: Platinum Complexes as Formulated by Werner

old formulation	Werner formulation		dissociation
$PtCl_4 \cdot 6NH_3$	$[Pt(NH_3)_6]Cl_4$	→	$[Pt(NH_3)_6]^{4+} + 4Cl^-$
$PtCl_4 \cdot 5NH_3$	$[Pt(NH_3)_5Cl]Cl_3$	→	$[Pt(NH_3)_5Cl]^{3+} + 3Cl^-$
$PtCl_4 \cdot 4NH_3$	$[Pt(NH_3)_4Cl_2]Cl_2$	→	$[Pt(NH_3)_4Cl_2]^{2+} + 2Cl^-$
$PtCl_4 \cdot 3NH_3$	$[Pt(NH_3)_3Cl_3]Cl$	→	$[Pt(NH_3)_3Cl_3]^+ + Cl^-$
$PtCl_4 \cdot 2NH_3$	$[Pt(NH_3)_2Cl_4]$	→	none
$PtCl_4 \cdot NH_3 \cdot KCl$	$K[Pt(NH_3)Cl_5]$	→	$K^+ + [Pt(NH_3)Cl_5]^-$
$PtCl_4 \cdot 2KCl$	$K_2[PtCl_6]$	→	$2K^+ + [PtCl_6]^{2-}$

formed was determined by measurements of the electrical conductivity of the solutions and by analytical determinations of the free chloride ions (where these were formed). It is noteworthy that in Werner's formulations the coordination number of platinum retains the constant value of 6. This constancy of coordination number of

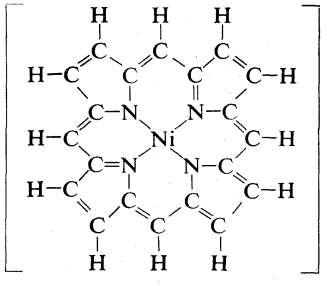
metal atoms occurs frequently, though not invariably.

Classification. A coordination compound is characterized by the nature of the central metal atom or ion; the oxidation state of the latter (that is, the gain or loss of electrons in passing from the neutral atom to the charged ion); and by the number, kind, and arrangement of the ligands. Because virtually all metallic elements form coordination compounds, sometimes in several oxidation states and usually with many different ligands, a large number of coordination compounds are known.

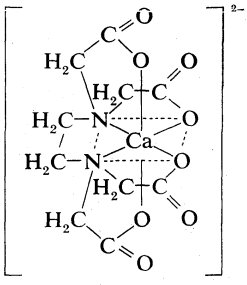
Mononuclear, unidentate. The simplest types of coordination compounds are those containing a single metal atom or ion (mononuclear compounds), surrounded by unidentate ligands. Most of the coordination compounds already cited, such as the platinum complexes of Werner, belong to this class. Among the ligands forming such complexes are a wide variety of neutral molecules, such as ammonia, water, carbon monoxide, and nitrogen, as well as monoatomic and polyatomic anions, such as the hydride, fluoride, chloride, oxide, hydroxide, nitrite, thiocyanate, carbonate, sulfate, and phosphate ions. Coordination of such ligands to the metal virtually always occurs through an atom possessing an unshared pair of electrons, which it donates to the metal to form a coordinate bond with the latter. Among the atoms that are known to coordinate to metals are those of virtually all the nonmetallic elements (such as hydrogen, carbon, oxygen, nitrogen, and sulfur), with the exception of the noble gases (helium, neon, argon, krypton, and xenon).

Coordination via unshared electron pair

Multidentate. The chelate complex of a copper ion and ethylenediamine mentioned above is an example of a compound formed between a metal ion and a bidentate ligand. Two further examples of chelate complexes are shown below. These are a nickel complex with a tetradentate large-ring ligand, known as a porphyrin, and a calcium complex with a hexadentate ligand, ethylenediaminetetraacetate. Because metal-ligand attachment in such chelate complexes is through several bonds, such complexes tend to be very stable.



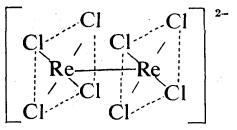
nickel-porphyrin complex



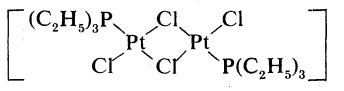
calcium-ethylenediaminetetraacetate complex

Certain ligands may be either unidentate or multidentate, depending on the particular compound in which they occur. The carbonate ion, formula $(CO_3)^{2-}$, for example, is coordinated to the cobalt ions in two cobalt compounds, $[Co(NH_3)_5(CO_3)]^+$ and $[Co(NH_3)_4(CO_3)]^+$, through one and two oxygen atoms, respectively.

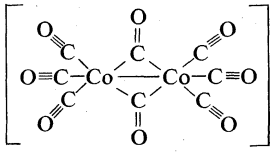
Polynuclear. Polynuclear complexes are coordination compounds containing two or more metal atoms, or ions, in a single coordination sphere. The two atoms may be held together through direct metal–metal bonds, through bridging ligands, or both. Examples of each are shown below, along with a unique metal-cluster complex having six metal atoms in its nucleus.



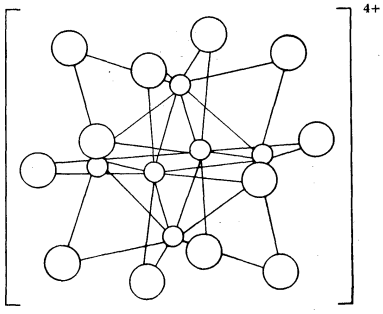
complex with metal–metal bond (Re is rhenium)



complex with bridging ligands (C_2H_5 is ethyl radical, Pt is platinum, P is phosphorus, and Cl is chlorine)



complex with metal–metal bond and bridging ligands



complex with metal-cluster nucleus

○ = molybdenum atoms

◯ = chlorine atoms

Nomenclature. Further discussion of coordination compounds requires a knowledge of electronic structure, chemical reactions, and vocabulary that cannot be provided in this article. The necessary background, however, can be found in detail in the articles ATOMIC STRUCTURE; CHEMICAL ELEMENTS; PERIODIC LAW; CHEMICAL COMPOUNDS, ORGANIC; and CHEMICAL REACTIONS.

Generally, the systematic naming of coordination compounds is carried out by rules recommended by the International Union of Pure and Applied Chemistry (IUPAC). Among the more important of these are the following:

IUPAC rules of nomenclature

1. Neutral and cationic complexes are named by first identifying the ligands, followed by the metal and its oxidation number (in Roman numerals enclosed within parentheses). Anionic ligands are cited ahead of neutral ones, and inorganic ligands ahead of organic ones. When the complex contains more than one ligand of a given kind, the number of such ligands is designated by one of the prefixes di-, tri-, tetra-, penta-, and so on, or, in the case of complex ligands, bis-, tris-, tetrakis-, pentakis-, and so on. The oxidation number of the metal is defined in the customary way as the residual charge on the metal if all the ligands were removed together with the electron pairs involved in coordination to the metal. The following examples are illustrative:

$[Co(NH_3)_5Cl]^{2+}$	chloropentaamminecobalt(III)
$[Ni(NH_2CH_2CH_2NH_2)_3]^{2+}$	tris(ethylenediamine)nickel(II)
$[Fe(CO)_5]$	pentacarbonyliron(0)
$[Al(H_2O)_5(OH)]^{2+}$	hydroxopentaaquoaluminum(III)
$[RhH(CO)(P[C_6H_5]_3)_3]$	hydridocarbonyltris(triphenylphosphine)rhodium(I).

2. Anionic complexes are similarly named, except that the name is terminated by the suffix -ate; *e.g.*,

$[PtCl_4]^{2-}$	tetrachloroplatinate(II)
$[Ag(S_2O_3)_2]^{3-}$	bis(thiosulfato)argentate(I)
$[OsNCl_5]^{2-}$	nitridopentachloroosmate(VI)
$[Cr(SCN)_4(NH_3)_2]^-$	tetrathiocyanatodiamminechromate(III).

3. In the case of salts, the cation is named first, then the anion; *e.g.*,

$[Pt(NH_3)_3Cl]Cl$	chlorotriammineplatinum(II) chloride
$K_3[Fe(CN)_6]$	potassium hexacyanoferrate(III)
$[Cr(NH_3)_6][Co(CN)_6]$	hexaamminechromium(III) hexacyanocobaltate(III).

4. Polynuclear complexes are named as follows, bridging ligands being identified by a prefix consisting of the Greek letter mu (μ-):

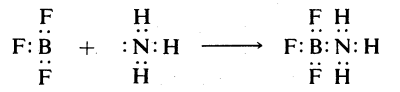
$[(CO)_5Mn-Mn(CO)_5]$	decacarbonyldimanganese (0) *or* bis(pentacarbonylmanganese)
$[(NH_3)_5Cr-OH-Cr(NH_3)_5]Cl_5$	μ-hydroxo-bis[pentaamminechromium(III)] chloride

In addition to their systematic designations, many coordination compounds are also known by names reflecting their discoverers or colours. Examples are:

$K[PtCl_3(C_2H_4)]$	Zeise's salt
$NH_4[Cr(NH_3)_2(SCN)_4] \cdot H_2O$	Reinecke's salt
$[Co(NH_3)_5(H_2O)]Cl_3$	roseocobaltic chloride (red)
$Fe_4[Fe(CN)_6]_3$	Prussian blue

Coordinate bonding. Werner originally postulated that coordination compounds could be formed because the central atoms carried the capacity to form secondary or coordinate bonds, in addition to the normal or valence bonds. A more complete description of coordinate bonding, in terms of electron pairs, became possible in the 1920s following the introduction of the concept that all covalent (nonionic) bonds consist of electron pairs shared between atoms, a concept advanced chiefly by the U.S. chemist Gilbert N. Lewis. In Lewis' formulation, when both electrons are contributed by one of the atoms, as in the boron–nitrogen bond, which is formed when the substance boron trifluoride combines with ammonia, the bond is called a coordinate bond:

Electron pair theory of bonding



boron trifluoride — ammonia — compound with coordinate bond

In these formulas, the valence (or bonding) electrons are indicated by dots, each pair of dots between two atomic symbols representing a bond between the corresponding atoms.

Following Lewis's ideas, the suggestion was made that the bonds between metals and ligands were of this same type, with the ligands acting as electron donors and the metal ions as electron acceptors. This suggestion provided the first electronic interpretation of bonding in coordination compounds. The coordination reaction between silver ions and ammonia illustrates the resemblance of coordination compounds to the situation in the boron–nitrogen compound. According to this view, the metal ion can be regarded as a so-called Lewis acid and the ligands as Lewis bases (for further information on Lewis acids and bases see ACID–BASE REACTIONS AND EQUILIBRIA):



silver ion — ammonia — coordination complex

In the above formulas, the dots again represent valence electrons.

Considerable success in understanding certain coordination compounds also has been achieved by treating them as examples of simple ionic or electrostatic bonding. This view attributes the bonding in coordination compounds to electrostatic forces between the positively charged metal ions and negatively charged ligands or, alternatively, in the case of neutral ligands (*e.g.*, water and ammonia), charge separations (dipoles) that appear within the molecules. Although this approach meets with considerable success for complexes of metal ions with small electronegative ligands, such as fluoride or chloride ions or water molecules, it breaks down for ligands of low polarity (charge separation), such as carbon monoxide. It also requires modification to explain why the spectral (light absorption) and magnetic properties of coordinated metal ions generally differ from those of the free ions and, for a given metal ion, depend on the nature of the ligands. Thus, the complex ion hexafluoroferrate(III) (FeF_6^{3-}), has magnetic properties to be expected from a substance with five unpaired electrons, as does the free iron(III) ion (Fe^{3+}), whereas the magnetic properties of the closely related hexacyanoferrate(III) $[Fe(CN)_6]^{3-}$ correspond to only one unpaired electron.

Since 1950 it has been apparent that a more complete theory, which incorporates contributions from both ionic and covalent bonding, is necessary to give an adequate account of the properties of coordination compounds. Such a theory is the so-called ligand-field theory, which has its origin in the more general theory of chemical bonding called the molecular-orbital theory (molecular orbitals describing the spatial distributions of electrons in molecules, just as atomic orbitals describe the distributions in atoms). This theory accounts with remarkable success for most properties of coordination compounds (see also CHEMICAL BONDING).

The ligand-field theory

An important conclusion from ligand field theory is that two types of bonds, called sigma bonds and pi bonds, occur in coordination compounds just as they do in ordinary covalent (organic) compounds. Sigma bonds are the more usual of the two, and they are symmetrical about the axis of the bond; pi bonds, which are less common, are unsymmetrical with regard to the bond axis. In coordination compounds, pi bonding may result from donation of electrons from ligands, such as fluorine or oxygen atoms, to empty *d* orbitals of the metal atoms (the designation *d* merely being a way of indicating the particular orbitals involved). An example of this type of bonding is the chromate ion, $(CrO_4)^{2-}$, in which the oxygen atoms donate electrons to the central chromium ion. Alternatively, electrons from *d* orbitals of the metal atom may be donated to empty orbitals of the ligand. This is the case in the compound tetracarbonylnickel, in which empty pi orbitals in the carbon monoxide molecules accept *d*-orbital electrons from the nickel atom.

Coordination number and geometry. Among the essential properties of coordination compounds are the number and arrangement of the ligands attached to the central metal atom or ion—that is, the coordination number and the coordination geometry, respectively. The coordination number of a particular complex is determined by the relative sizes of the metal atom and the ligands, by spatial (steric) constraints governing the shapes (conformations) of multidentate ligands, and by electronic factors, most notably the electronic configuration of the metal ion. The influence of the latter is illustrated by the examples in Table 2. The numbers labelled

Table 2: Coordination Numbers and Geometries of Metal Cyanide Complexes

electron configuration*	metal ion	cyanide complex	geometry	total number of valence electrons
d^2	Mo^{4+}	$[Mo(CN)_8]^{4-}$	dodecahedral	18
d^3	Cr^{3+}	$[Cr(CN)_6]^{3-}$	octahedral	15
d^4	Mn^{3+}	$[Mn(CN)_6]^{3-}$	octahedral	16
d^5	Fe^{3+}	$[Fe(CN)_6]^{3-}$	octahedral	17
d^6	Co^{3+}	$[Co(CN)_6]^{3-}$	octahedral	18
d^7	Co^{2+}	$[Co(CN)_5]^{3-}$	square pyramidal	17
d^8	Ni^{2+}	$[Ni(CN)_4]^{2-}$	square planar	16
d^8	Ni^{2+}	$[Ni(CN)_5]^{3-}$	square pyramidal or trigonal bipyramidal	18
d^{10}	Cd^{2+}	$[Cd(CN)_4]^{2-}$	tetrahedral	18
d^{10}	Ag^+	$[Ag(CN)_2]^-$	linear	14

*Number of *d* electrons indicated by superscript.

"total number of valence electrons" in this table comprise the *d* electrons of the metal ion together with the pair of electrons donated by each of the ligands.

Contributing to the pattern of coordination numbers in Table 2 is the 18-electron rule (sometimes called the noble gas rule), which states that coordination compounds in which the total number of valence electrons approaches but does not exceed 18 (the number of electrons in the valence shells of the noble gases) are most stable. The stabilities of 18-electron valence shells are also reflected in the coordination numbers of the stable mononuclear carbonyls of different metals—*e.g.*, tetracarbonylnickel, pentacarbonyliron, and hexacarbonylchromium (each of which has a valence shell of 18).

The 18-electron rule

The 18-electron rule applies particularly to covalent

complexes, such as the cyanides, carbonyls, and phosphines. For more ionic (also called outer-orbital) complexes, such as fluoro or aquo complexes, electronic factors are less important in determining coordination numbers, and configurations corresponding to more than 18 valence shell electrons are not uncommon. Several nickel complexes, for example, including the hexafluoro, hexaaquo, and hexaammine complexes, each have 20 valence shell electrons.

Any one metal ion tends to have the same coordination number in different complexes—*e.g.*, generally 6 for chromium(III)—but this is not invariably so. Differences in coordination number, for example, may result from difference in the sizes of the ligands. The iron(III) ion, for example, is able to accomodate six fluoride ions in the hexafluoride complex, but only four of the larger chloride ions in the tetrachloride. In some cases, a metal ion and ligand form two or more complexes with different coordination numbers—*e.g.*, tetracyanonickelate(II) and pentacyanonickelate(II).

Isomerism. Coordination compounds often exist as isomers—*i.e.*, as compounds with the same chemical composition but different structural formulas (see ISOMERISM). Many different kinds of isomerism occur among coordination compounds. The following are some of the more common types.

Ionization isomerism. Certain isomeric pairs occur that differ only in that two ionic groups exchange positions within (and without) the primary coordination sphere. These are called ionization isomers and are exemplified by the two compounds $[Co(NH_3)_5Br]SO_4$ and $[Co(NH_3)_5(SO_4)]Br$, in the first of which the bromide ion is coordinated to the cobalt ion, and the sulfate ion is outside the coordination sphere; and in the second of which the sulfate ion occurs within the coordination sphere, and the bromide ion is outside it.

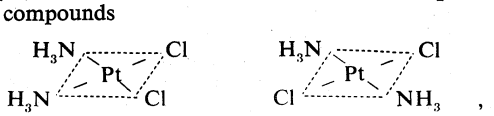
Coordination isomerism. Ionic coordination compounds in which both the positive and negative ions are ionic can exist as isomers if the ligands associated with the two metal atoms are exchanged, as in the pair of compounds shown by the formulas $[Co(NH_3)_6][Cr(CN)_6]$ and $[Cr(NH_3)_6][Co(CN)_6]$. Such compounds are called coordination isomers, as are the isomeric pairs obtained by redistributing the ligands between the two metal atoms, as in the doubly coordinated pair, $[Pt(NH_3)_4][PtCl_6]$ and $[Pt(NH_3)_4Cl_2][PtCl_4]$.

Ligand isomerism. Isomeric coordination compounds are known in which the overall isomerism results from isomerism solely within the ligand groups. An example of such isomerism is shown by the ions

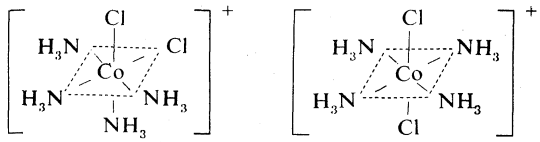
$$[Pt(NH_2CH_2CH_2CH_2NH_2)_2]^{2+} \text{ and } [Pt(NH_2\underset{\displaystyle CH_3}{\underset{|}{C}}HCH_2NH_2)_2]^{2+}.$$

Linkage isomerism. Isomerism also results when a given ligand is joined to the central atom through different atoms of the ligand. Such isomerism is called linkage isomerism. A pair of linkage isomers are the ions $[Co(NH_3)_5(NO_2)]^{2+}$ and $[Co(NH_3)_5(ONO)]^{2+}$, in which the nitro group is joined to the cobalt atom through nitrogen and oxygen atoms, as shown by designating the nitro group by the formulas NO_2^- and ONO^-, respectively. Another example of this variety of isomerism is given by the pair of ions $[Co(CN)_5(NCS)]^{3-}$ and $[Co(CN)_5(SCN)]^{3-}$, in which an isothiocyanate $(NCS)^-$, and a thiocyanate group $(SCN)^-$, are bound to the cobalt ion through a nitrogen and a sulfur atom, respectively.

Geometrical isomerism. Geometrical isomers of coordination compounds differ from one another only in the manner in which the ligands are distributed; for example, in the isomeric pair of dichlorodiammineplatinum (II) compounds



the two ammonia molecules and the two chlorine atoms are situated next to another in one isomer, called the *cis* (Latin: "on this side") isomer, and across from one another in the other, the *trans* (Latin: "on the other side") isomer. A similar relationship exists between the *cis* and *trans* forms of the dichlorotetraamminecobalt(III) ion:

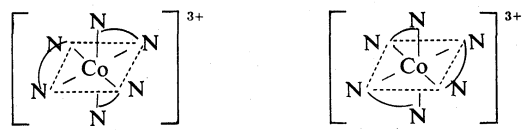


cis *trans*

Optical isomerism. Optical isomers possess the ability to rotate plane-polarized light in opposite directions (a property determined with an instrument called a polarimeter). Optical isomers exist when the molecules of the substances are mirror images but are not superimposable upon one other. In coordination compounds, optical isomerism can arise either from the presence of an asymmetric ligand, such as one isomer of an amino acid,

$$NH_2\underset{\displaystyle CH_3}{\underset{|}{C}}HCOOH\ ,$$

or from an asymmetric arrangement of the ligands. Familiar examples of the latter variety are octahedral complexes carrying three bidentate ligands, such as ethylenediamine, $NH_2CH_2CH_2NH_2$. The two optical isomers corresponding to such a complex are depicted by the structures below:



For convenience, the ethylenediamine ligands above are indicated merely by a curved line between the symbols for the nitrogen (N) atoms.

Optical isomers differ from the other isomeric coordination compounds in that their physical and chemical properties are identical.

PRINCIPAL CLASSES

The tendency for complexes to form between a metal ion and a particular combination of ligands and the properties of the resulting complexes depend on a variety of properties of both the metal ion and the ligands. Among the pertinent properties of the metal ion are its size, charge, and electron configuration. Relevant properties of the ligand include its size and charge, the number and kinds of atoms available for coordination, the sizes of the resulting chelate rings formed (if any), and a variety of other geometrical (steric) and electronic factors.

Stabilization of oxidation states

Many elements, notably certain of the metals, exhibit a range of oxidation states—that is, they are able to gain or lose varying numbers of electrons. The relative stabilities of these oxidation states are markedly affected by coordination of different ligands. The highest oxidation states correspond to empty or nearly empty *d* subshells (as the patterns of *d* orbitals are called). These states are generally stabilized most effectively by small negative ligands, such as fluorine and oxygen atoms, which possess unshared electron pairs. Such stabilization reflects, in part, the contribution of pi bonding caused by electron donation from the ligands to empty *d* orbitals of the metal ions in the complexes. Conversely, neutral ligands, such as carbon monoxide and unsaturated hydrocarbons, which are relatively poor electron donors but which can accept pi electrons from filled *d* orbitals of the metal, tend to stabilize the lowest oxidation states of metals. Intermediate oxidation states are most effectively stabilized by ligands such as water, ammonia, and cyanide ion, which are moderately good sigma electron donors but relatively poor pi electron donors or acceptors (see above *Coordinate bonding*). These trends are illustrated in Table 3, in which the complexes formed by chromium with various ligands are related to the oxidation states and electron configurations.

Table 3: Chromium Complexes of Various Oxidation States

oxidation state	electron configuration*	coordination complex
+6	d^0	$[CrF_6]$, $[CrO_4]^{2-}$
+5	d^1	$[CrO_4]^{3-}$
+4	d^2	$[CrO_4]^{4-}$, $[Cr(OR)_4]$†
+3	d^3	$[Cr(H_2O)_6]^{3+}$, $[Cr(NH_3)_6]^{3+}$
+2	d^4	$[Cr(H_2O)_6]^{2+}$
0	d^6	$[Cr(CO)_6]$, $[Cr(C_6H_6)_2]$

*Number of *d* electrons indicated by superscript.
†Capital R symbolizes an organic alkyl radical.

Aquo complexes. There are few ligands that equal water in respect to the number and variety of metal ions with which they form complexes. Nearly all metallic elements form aquo complexes. frequently in more than one oxidation state. Such aquo complexes include hydrated ions in aqueous solution as well as hydrated salts such as hexaaquochromium(III) chloride, $[Cr(H_2O)_6]Cl_3$. For metal ions with partially filled *d* subshells (*i.e.*, transition metals), the coordination numbers and geometries of the hydrated ions in solution can be inferred from their light-absorption spectra, which are generally consistent with octahedral coordination by six water molecules. Higher coordination numbers probably occur for the hydrated rare-earth ions such as lanthanum(III).

Complex ions in aqueous solution

When other ligands are added to an aqueous solution of a metal ion, replacement of water molecules in the coordination shell may occur, with the resultant formation of other complexes. Such replacement is generally a stepwise process, as illustrated by the following series of reactions that results from the progressive addition of ammonia to an aqueous solution of nickel(II) salt:

$$[Ni(H_2O)_6]^{2+} + NH_3 \rightleftharpoons [Ni(H_2O)_5(NH_3)]^{2+} + H_2O$$

$$[Ni(H_2O)_5(NH_3)]^{2+} + NH_3 \rightleftharpoons [Ni(H_2O)_4(NH_3)_2]^{2+} + H_2O$$

$$[Ni(H_2O)_4(NH_3)_2]^{2+} + NH_3 \rightleftharpoons [Ni(H_2O)_3(NH_3)_3]^{2+} + H_2O$$

$$[Ni(H_2O)_3(NH_3)_3]^{2+} + NH_3 \rightleftharpoons [Ni(H_2O)_2(NH_3)_4]^{2+} + H_2O$$

$$[Ni(H_2O)_2(NH_3)_4]^{2+} + NH_3 \rightleftharpoons [Ni(H_2O)(NH_3)_5]^{2+} + H_2O$$

$$[Ni(H_2O)(NH_3)_5]^{2+} + NH_3 \rightleftharpoons [Ni(NH_3)_6]^{2+} + H_2O$$

With increasing additions of ammonia, the equilibria are shifted toward the higher ammine complexes (those with more ammonia and less water) until ultimately the hexaamminenickel(II) ion predominates.

The tendency of metal ions in aqueous solution to form complexes with ammonia as well as with organic amines (derivatives of ammonia, carrying chains of carbon atoms attached to the nitrogen atom) is widespread. The stabilities of such complexes exhibit a considerable range of dependence on the nature of the metal ion as well as on that of the amine. The marked enhancement of stability that results from chelation is reflected in the equilibrium constants of the reactions—values that indicate the relative proportions of the starting materials and the products at equilibrium. Table 4, for example, shows the equilibrium constants for the formation of complexes of hexaaquonickel(II) ions with a series of polyamines—*i.e.*, for a series of reactions of the type

Equilibrium constants

$$[Ni(H_2O)_6]^{2+} + nL \rightleftharpoons [Ni(H_2O)_{6-n}(L)_n]^{2+} + nH_2O,$$

in which *L* is the ligand and *n* the number of water molecules displaced from the complex. In the table it can be seen that the equilibrium constants, K_L, increase dramatically as the possibilities for chelation increase—that is, as the number of nitrogen atoms (N) available for bonding to the metal atom increase.

Table 4: Equilibrium Constants for the Formation of Various Nickel Amine Complexes

n	amine (*L*)	equilibrium constant K_L, M^{-1}*
1	NH_3	5×10^2
2	$\underline{N}H_2CH_2CH_2\underline{N}H_2$	4×10^7
3	$\underline{N}H_2CH_2CH_2\underline{N}HCH_2CH_2\underline{N}H_2$	5×10^{10}
4	$\underline{N}H_2CH_2CH_2\underline{N}HCH_2CH_2\underline{N}HCH_2CH_2\underline{N}H_2$	1×10^{14}

*M is molar concentrations.

It should be noted that in the particular examples cited above the coordination number of the metal ion is invariant throughout the substitution process, but this is not always the case. Thus, the ultimate products of the addition of the cyanide ion to an aqueous solution of hexaaquonickel(II) ion are tetracyanonickelate(II) and pentacyanonickelate(II). Similarly, addition of the chloride ion to a solution of hexaaquoiron(III) yields tetrachloroferrate(III).

Halide complexes. Probably the most widespread class of complexes involving anionic ligands is that of the complexes of the halide ions—*i.e.*, the fluoride, chloride, bromide, and iodide ions. In addition to forming simple halide salts, such as sodium chloride and nickel difluoride (in which the metal ions are surrounded by halide ions, these in a sense being regarded as coordinated to them), many metals form complex halide salts, such as potassium tetrachloroplatinate(II), $K_2[PtCl_4]$, that contain discrete complex ions. Most metal ions also form halide complexes in aqueous solution. The stabilities of such complexes span an enormous range—from the alkali-metal ions (lithium, sodium, potassium, and so on), the formation of whose halide complexes in aqueous solution can barely be detected, to extremely stable halide complexes, such as the tetraiodomercurate(II), tetrachlorothallate(III), and tetrachloropalladate(II) ions, the extent of whose dissociation is extremely small.

The stabilities of halide complexes reflect a pattern in which metal ions can be divided into two general classes, designated as *A* and *B* or, alternatively, as hard and soft, respectively. (Generally, the electrons in the atoms of the hard elements are considered to form a compact and not easily deformable group, whereas those in the atoms of the soft elements form a looser group—that is, one more easily deformed.) For the former class, which includes beryllium, magnesium, scandium, chromium, iron, nickel, copper, indium, and tin, the order of increasing stability of the halide complexes in aqueous solution is iodides, bromides, chlorides, and fluorides. Conversely, for the class *B* (or soft) ions, such as platinum, silver, cadmium, mercury, thallium, and lead, the order of increasing stability of the halide complexes is fluorides, chlorides, bromides, and iodides. In contrast to class *A* metals, those of class *B* also tend to form more stable complexes with sulfur-containing ligands than with oxygen-containing ligands and more stable complexes with phosphorus ligands than with nitrogen ligands.

Hard and soft ions

Metal carbonyls. Following the discovery of the first metal carbonyl complex, tetracarbonylnickel, in 1890, many compounds containing carbon monoxide coordinated to transition metals have been prepared and characterized. For reasons already discussed, such compounds generally contain metal atoms or ions in low oxidation states. The following are some of the more common types of metal carbonyl compounds: (1) simple mononuclear carbonyls of zero-valent metals, such as tetracarbonylnickel, pentacarbonyliron, and hexacarbonylchromium—highly toxic, volatile compounds, the most stable of which have filled valence shells of 18 electrons; (2) salts of anionic and cationic carbonyls, such as tetracarbonylcobaltate(I) and hexacarbonylmanganese(I); (3) binuclear and polynuclear carbonyls, such as bis(tetracarbonylcobalt), the structural formula of which was shown earlier (see above *Classification: Polynuclear*); and (4) mixed complexes containing other ligands additional to CO: chloropentacarbonylmanganese, hydridotetracarbonylcobalt (I), and tricarbonylnitrosylcobalt.

Although molecular nitrogen, N_2, is isoelectronic with carbon monoxide (that is, has the same arrangement of electrons), its tendency to form complexes with metals is much smaller. The first complex containing molecular nitrogen as a ligand—*i.e.*, pentaamminenitrogenruthenium(II), $[Ru(NH_3)_5(N_2)]^{2+}$—was prepared in 1965, and many others have been discovered subsequently. Such

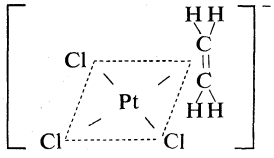
complexes have attracted considerable interest because of their possible roles in the chemical and biological fixation of nitrogen.

Complexes of molecular nitrogen

Organometallic complexes. Organometallic complexes, complexes formed between organic (carbon-containing) groups and metal atoms, can be divided into two general classes: (1) complexes containing metal–carbon sigma bonds and (2) pi-bonded metal complexes of unsaturated hydrocarbons—that is, compounds with multiple bonds between carbon atoms.

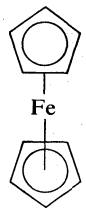
The first of these classes is exemplified by dimethylmercury(II) and iodotrimethylplatinum(IV). Such complexes of transition metals tend to be unstable, and only since 1950 have a large number of them been prepared and characterized. Most of these contain, in addition to the organic ligands, other ligands, such as carbon monoxide, cyanide, or phosphines, that appear to exert a stabilizing influence. Examples of organometallic complexes stabilized by other ligands are pentacyanomethylcobaltate(III), chlorobis(triethylphosphine)methylplatinum (II), and pentacarbonylmethylmanganese(I).

Zeise's salt, $K[PtCl_3(C_2H_4)]$, discovered about 1830, is the earliest recorded example of a metal complex of an unsaturated hydrocarbon. This salt contains the trichloroethyleneplatinate(II) anion:

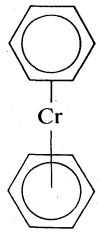


Analogous olefin complexes are formed by a number of other metal ions, including silver, copper, palladium, and ruthenium. The scope of this general field of chemistry was greatly expanded in 1951 by the discovery of the first "sandwich" compound, bis(cyclopentadienyl)-iron(II), whose structure is depicted below. The term sandwich refers to the fact that the two flat cyclopentadienyl rings lie, respectively, above and below the iron atom (which is, in effect, sandwiched between them). A great variety of such compounds are now known, not only of the cyclopentadienyl type but also containing other unsaturated cyclic ligands; *e.g.*, dibenzenechromium, benzenetricarbonylchromium, and cyclobutadienetricarbonyliron, all of which are also shown below. This class of compounds is characterized by great stability. Their bonding is closely related to that in the metal carbonyls and involves donation of pi electrons from the hydrocarbon ligands to unfilled *d* orbitals of the metal, together with "back donation" of metal *d* electrons to empty pi orbitals of the ligand (see ORGANOMETALLIC COMPOUNDS):

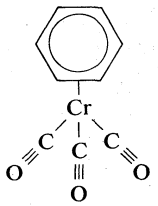
"Sandwich" compounds



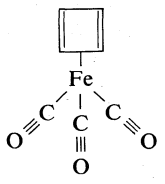
bis(cyclopentadienyl)iron



dibenzenechromium



benzenetricarbonyl chromium



cyclobutadienetricarbonyliron

In these diagrams the polygons represent hydrocarbon rings, and the circles indicate rings of pi electrons.

GENERAL METHODS OF PREPARATION

The great variety of coordination compounds is matched by the diversity of methods through which such compounds can be synthesized. Complex halides, for example, may be prepared by direct combination of two halide salts (either in the molten state or in a suitable solvent). Palladium chloride and potassium chloride, for example, react to give the complex potassium tetrachloropalladate(II), as shown in the following equation:

$$PdCl_2 + 2KCl \longrightarrow K_2[PdCl_4].$$

Another widely used route to coordination compounds is through the direct combination of a metal ion and appropriate ligands in solution. Thus, the addition of a sufficiently high concentration of ammonia to an aqueous solution of a nickel(II) salt leads, through a series of reactions (see above *Aquo complexes*) to the formation of the hexaamminenickel(II) ion, which can be precipitated, for example, as the sulfate salt, $[Ni(NH_3)_6]SO_4$.

Complexes of metal ions in high oxidation states are sometimes more readily formed by adding the ligands to a solution of the metal ion in a lower oxidation state in the presence of an oxidizing agent. Thus, addition of ammonia to an aqueous solution of a cobalt(II) salt in the presence of air or oxygen leads to the formation of cobalt(III) ammine complexes such as hexaamminecobalt(III), $[Co(NH_3)_6]^{3+}$, and aquopentaamminecobalt(III), $[Co(NH_3)_5(H_2O)]^{3+}$, ions.

Complexes of metals in low oxidation states, such as the carbonyls of zero-valent metals, can sometimes be prepared by direct combination of the metal with the ligand as, for example, in the reaction of nickel metal with carbon monoxide:

$$Ni(solid) + 4CO \xrightarrow[1\ atm]{25^\circ} Ni(CO)_4.$$

More commonly, a salt of the metal is reduced in the presence of the ligand. An example of this type of synthesis is the reduction of cobalt carbonate with hydrogen in the presence of carbon monoxide to give bis(tetracarbonylcobalt):

$$2CoCO_3 + 2H_2 + 8CO \xrightarrow[250-300\ atm]{120-200^\circ} Co_2(CO)_8 + 2CO_2 + 2H_2O.$$

Similar procedures are applicable to the synthesis of metal sandwich compounds containing cyclopentadienyl and benzene ligands. Dibenzenechromium, for example, can be prepared from chromic chloride, benzene, and aluminum, as shown in the following equation:

$$CrCl_3 + 2C_6H_6 + Al \longrightarrow Cr(C_6H_6)_2 + AlCl_3.$$

Hydrido complexes of transition metals can be prepared by reactions of suitable precursors either with molecular hydrogen or with suitable reducing agents such as hydrazine or sodium borohydride; for example,

$$2[Co(CN)_5]^{3-} + H_2 \longrightarrow 2[CoH(CN)_5]^{3-}$$

$$PtCl_2[P(CH_3CH_2)_3]_2 \xrightarrow{N_2H_4 \text{ or } Na[BH_4]} PtHCl[P(CH_3CH_2)_3]_2$$

Transition metal complexes containing metal–carbon bonds can be prepared by a variety of routes, some of the more important of which are illustrated by the following examples:

$$Mn(CO)_5Br + LiCH_3 \longrightarrow Mn(CO)_5(CH_3) + LiBr,$$

$$[Mn(CO)_5]^- + CH_3Br \longrightarrow Mn(CO)_5(CH_3) + Br^-,$$

$$PtHCl[P(CH_3CH_2)_3]_2 + CH_2{=}CH_2 \longrightarrow Pt(CH_2CH_3)Cl[P(CH_3CH_2)_3]_2, \text{ and}$$

$$2[Co(CN)_5]^{3-} + CH_3I \longrightarrow [Co(CN)_5(CH_3)]^{3-} + [Co(CN)_5I]^{3-}.$$

PRINCIPAL REACTIONS

Acid–base. Coordination to a positive metal ion usually enhances the acidity (*i.e.*, the tendency to lose protons) of hydrogen-containing ligands, such as water and ammonia. Thus, metal ions in aqueous solution commonly exhibit acidic behaviour. Such behaviour is exemplified by hydrolysis reactions of the type shown in the following equilibrium:

$$[M(H_2O)_x]^{n+} \rightleftharpoons [M(H_2O)_{x-1}(OH)]^{(n-1)+} + H^+,$$

in which M represents the metal ion, n its charge, and x the number of coordinated water molecules.

The acidities of such aquo ions depend upon the charge, size, and electronic configuration of the metal ion. This dependence is reflected in the values of acid dissociation constants, which range from about 10^{-14} (a value only slightly larger than for pure water, for which the dissociation constant $= 10^{-15.7}$) for the hydrated lithium ion, to about 10^{-2} (a value equivalent to that of a fairly strong acid) for the hydrated uranium(IV) ion. Acid–base equilibria, such as shown in the above equation, are rapidly established in solution, generally within a fraction of a second (see ACID–BASE REACTIONS AND EQUILIBRIA).

In some cases, hydrolysis of a metal ion may be accompanied by polymerization to form dinuclear or polynuclear hydroxo- or oxygen-bridged complexes:

$$2[Fe(H_2O)_6]^{3+} \rightleftharpoons \left[(H_2O)_4Fe\begin{matrix}H\\O\\O\\H\end{matrix}Fe(H_2O)_4\right]^{4+} + 2H^+ + 2H_2O.$$

Even very weakly acidic ligands, such as ammonia, can acquire appreciable acidity through coordination to a metal ion. Thus, the hexaammineplatinum(IV) ion dissociates according to the equilibrium

$$[Pt(NH_3)_6]^{4+} \rightleftharpoons [Pt(NH_3)_5(NH_2)]^{3+} + H^+.$$

Substitution. One of the most general reactions exhibited by coordination compounds is that of substitution, or replacement, of one ligand by another. This process is depicted in a generalized manner by the equation $ML_{x-1}Y + Z \rightarrow ML_{x-1}Z + Y$ for a metal complex of coordination number x. The ligands L, Y, and Z may be chemically similar or different. (Charges have been omitted here for simplicity.)

A class of substitution reactions that affords the widest possible comparison of different metal ions is the replacement of water in the coordination shells of metal-aquo complexes in aqueous solution. The substitution may be by another water molecule (which can be labelled with the isotope oxygen-18 to permit the reaction to be followed) or by a different ligand, such as the chloride ion. Reactions of both sorts occur as shown below (oxygen-18 is indicated by the symbol $\overset{*}{O}$):

Replacement of water

$$[M(H_2O)_x]^{n+} + H_2\overset{*}{O} \longrightarrow [M(H_2O)_{x-1}(H_2\overset{*}{O})]^{n+} + H_2O$$

$$[M(H_2O)_x]^{n+} + Cl^- \longrightarrow [M(H_2O)_{x-1}Cl]^{(n-1)+} + H_2O.$$

Many such reactions are extremely fast, and it has been only since 1950, following the development of appropriate experimental methods (including stopped flow, nuclear magnetic resonance, and relaxation spectrometry), that the kinetics and mechanisms of this class of reactions have been extensively investigated. Rates of substitution of metal-aquo ions have been found to span a wide range, the characteristic times required for substitution ranging from less than 10^{-9} second for monopositive ions, such as hydrated potassium ions, to several days for certain more highly charged ions, such as hexaaquochromium(III) and hexaaquorhodium(III). The rate of substitution parallels the ease of loss of a water molecule from the coordination shell of the aquo complex and thus increases with increasing size and with decreasing charge of the metal ion. For transition metal ions, electronic factors also have an important influence on rates of substitution.

There are two limiting mechanisms (or pathways) through which substitution may occur; namely, dissociative and associative mechanisms. In the dissociative mechanism, a ligand is lost from the complex to give an intermediate compound of lower coordination number. This type of reaction path is typical of octahedral complexes, many aquo complexes, and metal carbonyls such as tetracarbonylnickel. An example of a dissociative reaction pathway for an octahedral complex of cobalt is as follows:

Dissociative and associative mechanisms

$$[Co(CN)_5(H_2O)]^{2-} \xrightarrow{-H_2O} [Co(CN)_5]^{2-} \xrightarrow{+I^-} [Co(CN)_5I]^{3-}.$$

The associative mechanism for substitution reactions, on the other hand, involves association of an extra ligand with the complex to give an intermediate of higher coordination number; then one of the original ligands is lost to restore the initial coordination number. Substitution reactions of square planar complexes, such as those of the nickel(II), palladium(II), and platinum(II) ions, usually proceed through associative pathways involving intermediates with coordination number 5. An example of a reaction following such a pathway is

$$[Pt(NH_3)_4]^{2+} \xrightarrow{Cl^-} \left[\begin{matrix}NH_3\\(NH_3)_3Pt\\Cl\end{matrix}\right]^+ \xrightarrow{-NH_3} [Pt(NH_3)_3Cl]^+.$$

A characteristic feature of this class of reactions is the sensitivity of the rate of substitution of a given ligand to the nature of the ligand in the *trans* position (that is, on the opposite side of the metal atom). The *trans* ligand activates a ligand for replacement as follows, in decreasing order: carbon monoxide, cyanide ion, or ethylene, greater than phosphines, or hydride, nitrite, iodide, or thiocyanate ion, greater than bromide or chloride ion, greater than ammonia or water. The *trans* effect may be utilized for synthetic purposes; thus, the reaction of the tetrachloroplatinate(II) ion with ammonia yields *cis*-dichlorodiammineplatinum(II); whereas the reaction of the tetraammineplatinum(II) ion with the chloride ion gives the *trans* isomer, *trans*-dichlorodiammineplatinum(II). The reactions are shown below. In both reactions the *trans* effect causes introduction of the ligand *trans* to chloride rather than *trans* to ammonia:

$$\left[\begin{matrix}Cl\\Cl-Pt-Cl\\Cl\end{matrix}\right]^{2-} \xrightarrow[-Cl^-]{+NH_3} \left[\begin{matrix}Cl\\Cl-Pt-NH_3\\Cl\end{matrix}\right] \xrightarrow[-Cl^-]{+NH_3} \left[\begin{matrix}Cl\\Cl-Pt-NH_3\\NH_3\end{matrix}\right]$$

$$\left[\begin{matrix}NH_3\\H_3N-Pt-NH_3\\NH_3\end{matrix}\right]^{2+} \xrightarrow[-NH_3]{+Cl^-} \left[\begin{matrix}NH_3\\H_3N-Pt-Cl\\NH_3\end{matrix}\right]^{+} \xrightarrow[-NH_3]{+Cl^-} \left[\begin{matrix}NH_3\\Cl-Pt-Cl\\NH_3\end{matrix}\right]$$

Isomerization. Coordination compounds that exist in two or more isomeric forms (see above *Isomerism*) may undergo reactions that convert one isomer to another. Examples are the linkage isomerization and *cis-trans* isomerization reactions depicted below. The first of these has been shown to proceed intramolecularly (*i.e.*, without dissociation of the nitrite ligand), whereas the second probably occurs through dissociation of one of the water-molecule ligands:

$$[Co(NH_3)_5(ONO)]^{2+} \longrightarrow [Co(NH_3)_5(NO_2)]^{2+}$$

$$cis\text{-}[Co(NH_2CH_2CH_2NH_2)_2(H_2O)_2]^{3+} \longrightarrow trans\text{-}[Co(NH_2CH_2CH_2NH_2)_2(H_2O)_2]^{3+}.$$

Oxidation–reduction. Transition metals commonly exhibit two or more stable oxidation states, and their complexes accordingly are able to undergo oxidation–reduction reactions. The simplest such reactions involve electron transfer between two complexes, with little if

any accompanying rearrangement or chemical change. An example is shown below:

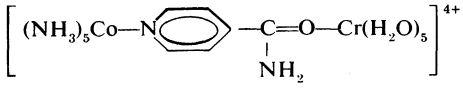
$$[Fe(CN)_6]^{4-} + [IrCl_6]^{2-} \longrightarrow [Fe(CN)_6]^{3-} + [IrCl_6]^{3-}.$$

In other cases, oxidation–reduction is accompanied by significant chemical rearrangement. An example is

$$[Co(NH_3)_5Cl]^{2+} + [Cr(H_2O)_6]^{2+} + 5[H_3O]^+ \longrightarrow [Co(H_2O)_6]^{2+} + [Cr(H_2O)_5Cl]^{2+} + 5[NH_4]^+.$$

Outer-sphere and inner-sphere mechanisms

Two limiting mechanisms of electron transfer, commonly designated outer-sphere and inner-sphere mechanisms, have been recognized. Outer-sphere electron transfer occurs without dissociation or disruption of the coordination shell of either complex; *i.e.*, through both intact coordination shells. The first reaction above is of this type. On the other hand, inner-sphere electron transfer—*e.g.*, the second reaction above—proceeds by formation of a binuclear complex in which the two metal ions are joined by a common bridging ligand (in this case the chloride ion) through which the electron is transferred. Such electron transfer also may occur through polyatomic bridging ligands to which the two metal ions are attached at different sites separated by several atoms; for example, the reduction of pentaammine(isonicotinamide)cobalt(III) by chromium(II) ion through a bridged intermediate,



Strikingly large differences in rates of electron transfer are observed even among closely related reactions. Thus, the rate of reduction of the bromopentaamminecobalt-(III) ion by the hexaaquochromium(II) ion is about 10^7 times higher than that of the acetatopentaamminecobalt-(III) ion by the same chromium ion.

Oxidative addition. The oxidations of certain complexes, notably those of metal ions with nearly filled *d* shells, are accompanied by increases in their preferred coordination numbers in accord with the patterns shown in Table 2. Such complexes are particularly effective in reducing saturated molecules, such as chlorine, hydrogen peroxide, methyl iodide, and hydrogen, with accompanying incorporation of the fragments of reductive cleavage (the chloride ion, hydroxide ion, methyl carbanion, and hydride ion, respectively) into the coordination shells to achieve the necessary expansions of the coordination numbers. Three such classes of reactions, namely those characteristic of five-coordinate (d^7), four-coordinate (d^8), and two-coordinate (d^{10}) complexes, are illustrated by the equations below:

$$2[Co(CN)_5]^{3-} + H_2O_2 \longrightarrow 2[Co(CN)_5(OH)]^{3-}$$

$$IrCl(CO)[P(C_6H_5)_3]_2 + H_2 \longrightarrow IrH_2Cl(CO)[P(C_6H_5)_3]_2$$

$$Pt[P(C_6H_5)_3]_2 + CH_3I \longrightarrow PtI(CH_3)[P(C_6H_5)_3]_2$$

Interest in the synthetic and catalytic applications of such oxidative addition reactions has prompted intensive study of these reactions since about 1960, which has resulted in the discovery of new reactions of this type, and in the elucidation of certain aspects of the mechanisms by which they occur.

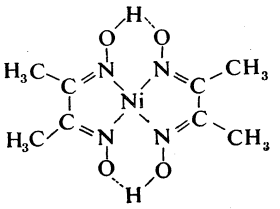
SIGNIFICANCE AND APPLICATIONS

The applications of coordination compounds in chemistry and technology are many and diverse. Naturally occurring coordination compounds are vitally important to living organisms.

Dyes and pigments. The brilliant and intense colours of many coordination compounds render them of great value as dyes and pigments. The early use of Prussian blue as a pigment has already been mentioned. Phthalocyanine complexes—for example, copper phthalocyanine—containing large-ring ligands closely related to the porphyrins, constitute an important class of dyes for fabrics.

Extraction and separation of metals. Several important hydrometallurgical processes utilize metal complexes. Nickel, cobalt, and copper can be extracted from their ores as ammine complexes using aqueous ammonia. Differences in the stabilities and solubilities of the ammine complexes can be utilized in selective precipitation procedures that bring about separation of the metals. The purification of nickel can be effected by reaction with carbon monoxide to form the volatile tetracarbonylnickel complex, which can be distilled and thermally decomposed to deposit the pure metal. Aqueous cyanide solutions usually are employed to separate gold from its ores in the form of the very stable dicyanoaurate(I) complex. Cyanide complexes also find application in electroplating.

Chemical analysis. There are a number of ways in which coordination compounds are used in the analysis of various substances. These include (1) the selective precipitation of metal ions as complexes, for example, nickel(II) ion as the dimethylglyoxime complex (shown below); (2) the formation of coloured complexes, such as the tetrachlorocobaltate(II) ion, that can be determined spectrophotometrically—that is, by means of their light absorption properties; and (3) the preparation of complexes, such as metal acetylacetonates, that can be separated from aqueous solution by extraction with organic solvents:



Sequestering. In certain circumstances the presence of metal ions is undesirable, as for example in water, in which calcium and magnesium ions cause hardness. In such cases the undesirable effects of the metal ions frequently can be eliminated by "sequestering" the ions as harmless complexes through the addition of an appropriate complexing reagent. Ethylenediaminetetraacetic acid (EDTA) forms very stable complexes and is widely used for this purpose. Its applications include water softening (by tying up calcium and magnesium ions) and the preservation of organic substances, such as vegetable oils and rubber, in which case it combines with traces of transition metal ions that would catalyze oxidation of the organic substances.

Catalysis. A technological and scientific development of major significance was the discovery in 1954 that certain complex metal catalysts, viz., a combination of titanium trichloride and triethylaluminum bring about the polymerizations (or joining together of the molecules) of organic compounds with carbon–carbon double bonds under mild conditions to form polymers of high molecular weight and highly ordered (stereoregular) structures. Certain of these polymers are of great commercial importance because they are used to make fibres, films, and plastic articles of many kinds. Other technologically important processes based on metal complex catalysts include the catalysis by metal carbonyls, such as hydridotetracarbonylcobalt(I), of the so-called hydroformylation of olefins—*i.e.*, of their reactions with hydrogen and carbon monoxide to form aldehydes—and the catalysis by tetrachloropalladate(II) ions of the oxidation of ethylene in aqueous solution to acetaldehyde (see also CATALYSIS).

Biology. Metal complexes play a variety of important roles in biological systems. Many enzymes, the naturally occurring catalysts that regulate biological processes, are metal complexes (metalloenzymes); for example, a hydrolytic enzyme important in digestion, carboxypeptidase, contains a zinc ion coordinated to several amino acid residues of the protein. Another enzyme, catalase, which is a very efficient catalyst for the decomposition of hydrogen peroxide, contains iron–porphyrin complexes. In both cases the coordinated metal ions are probably

the sites of catalytic activity. Hemoglobin also contains iron–porphyrin complexes, its role as an oxygen carrier being related to the ability of the iron atoms to coordinate oxygen molecules reversibly. Other biologically important coordination compounds include chlorophyll (a magnesium–porphyrin complex) and vitamin B_{12}, a complex of cobalt with a macrocyclic ligand known as corrin.

BIBLIOGRAPHY. J.C. BAILAR (ed.), *Chemistry of the Coordination Compounds* (1956), a comprehensive monograph containing a historical account and detailed chapters on various specialized topics, such as stereochemistry and applications; F. BASOLO and R.C. JOHNSON, *Coordination Chemistry* (1964), an introductory account with emphasis on reactions of coordination compounds; M.M. JONES, *Elementary Coordination Chemistry* (1964), the most complete elementary account of the subject including nomenclature, descriptive chemistry, bonding, structures, stabilities, and applications; L.E. ORGEL, *An Introduction to Transition-Metal Chemistry: Ligand-Field Theory*, 2nd ed. (1966), a qualitative, highly readable account of coordination and organometallic compounds of transition metals; D.P. GRADDON, *An Introduction to Co-ordination Chemistry*, 2nd ed. (1968), an excellent concise account of the most important aspects of coordination compounds including stereochemistry and applications; J. HALPERN, "Homogeneous Catalysis by Coordination Compounds," *Adv. Chem. Ser.* 70:1–24 (1968), a review of catalysis of hydrogenation and related reactions of olefins; "Coordination Compounds in Homogeneous Catalysis," *Pure Appl. Chem.*, 20:59–75 (1969), a brief general review of the catalytic properties of coordination compounds.

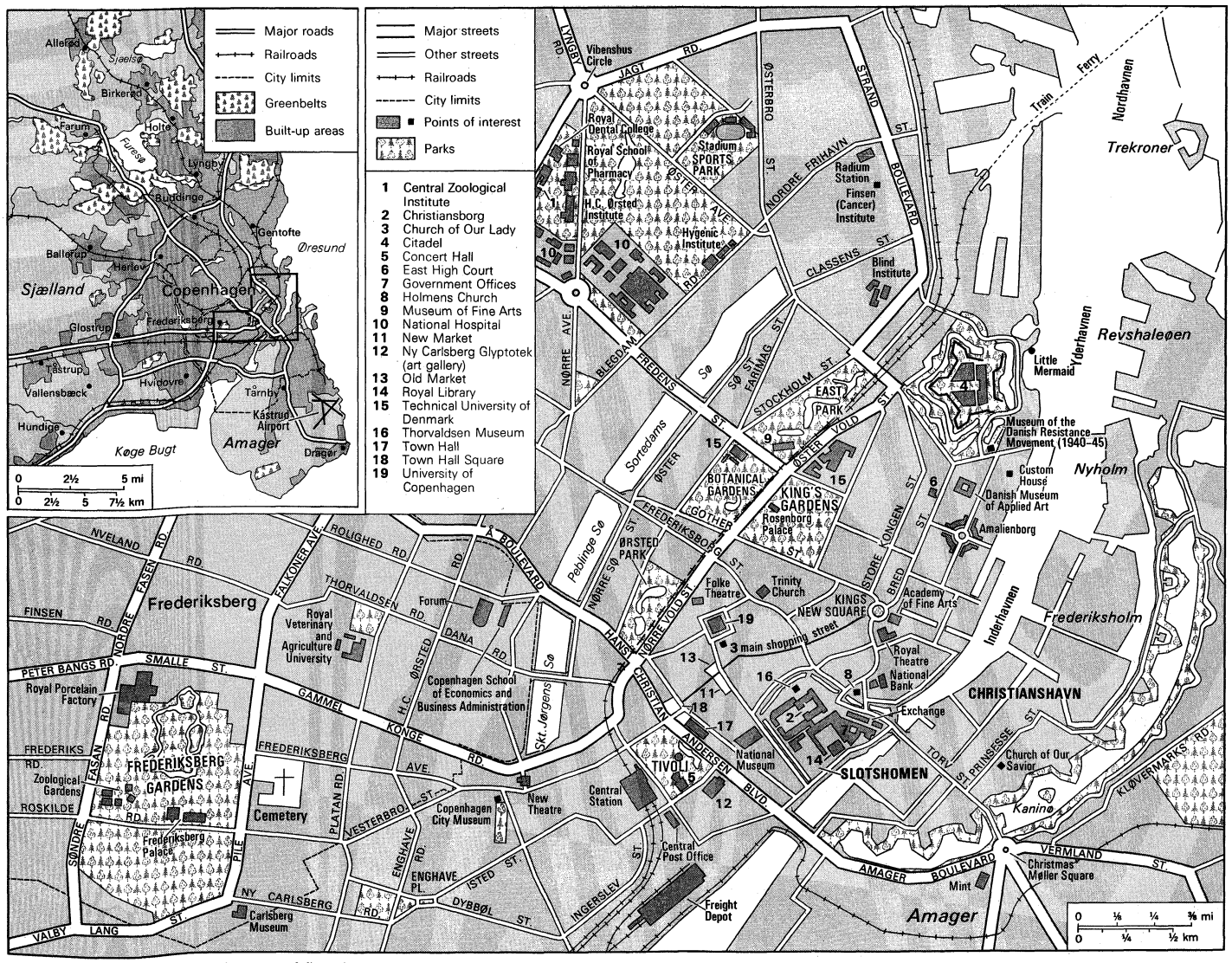
(J.Ha.)

Copenhagen

Copenhagen (Danish København, or Merchant's Harbour), the capital of Denmark, lies on the eastern coast of the island of Zealand (Sjælland) and on the northern part of the island of Amager at the southern end of the Sound (Øresund) opposite Malmö, Sweden. It is the largest and most important city of Denmark, the seat of a number of cultural and social institutions, and a centre of trade and industry. The city proper had a population of about 813,000 (including the municipalities of Frederiksberg and Gentofte) in the early 1970s, but Greater Copenhagen had about 1,400,000 inhabitants, almost one-third of the population of Denmark.

The city's geographical position gives it a relatively mild climate for its latitude, particularly in wintertime when it draws warmth from the Gulf Stream. Temperatures go down to about freezing in January and February; the average in July and August is about 60° F (16° C). Average annual precipitation is 28 inches, more than in London, Paris, or Berlin. The rainiest months are July and August; winter and spring are relatively dry.

Until 1856 Copenhagen was a fortress town with ramparts and moats, surrounded by a ring of suburbs and villages that were later absorbed by the expanding city. The old town retains its narrow streets and Baroque and neoclassical buildings. Outside the former ramparts the buildings vary from five- or six-story apartment houses to high-rise apartment blocks and single-family houses with gardens.



Central Copenhagen and (inset) its metropolitan area.

The fortress town

A small village existed on the site of the present city as early as the year 900. In 1167 Bishop Absalon of Roskilde built a castle on an islet off the coast, the remains of which can still be seen underneath the present castle of Christiansborg. The Bishop marked out the boundaries of the town and surrounded it with ramparts and a moat. In the Middle Ages it had four churches, three convents, and a number of clerical establishments; a town hall and a university were inaugurated in 1479. In the civil and religious conflicts of the age of the Reformation the town was often attacked; the churches were plundered, and the convents were abolished.

During the latter part of the 16th century, Copenhagen's trade began to flourish and the city itself to expand, adding Christianshavn on the south and Ny København (New Copenhagen) on the northeast. Among the new buildings erected were the Børsen (Exchange), the Holmens Kirke (Holmens Church), Trinitatis (Trinity) Kirke, with the adjacent famous Round Tower, and the palace of Rosenborg (now the museum of the royal family).

During the wars with Sweden in 1658–60, the city was besieged for two years. In 1661 the King permitted Copenhagen to have its own City Council of 32 men elected by the leading citizens. In 1728 a fourth of the town was burned down; the rebuilding was done in a Baroque style that still characterizes some of the streets in the centre of the city. Another great fire occurred in 1795, and the houses built after that were all in neoclassic style. In 1807, during the war against England, the university quarter around Vor Frue Kirke (Church of Our Lady) was destroyed. The church, which dates from the 12th century, was rebuilt and adorned with marble works by Bertel Thorvaldsen, the most famous of which is the colossal statue of Christ.

The contemporary city

The nucleus of the city is the area between the sea and the old ramparts and moats. West of the ramparts is a series of small freshwater lakes, Sankt Jørgens Sø, Peblinge Sø, and Sortedams Sø. Around the original city lie the densely populated districts built after the removal of the ramparts in 1856. In the early 20th century the city expanded still further, incorporating what are now the districts of Valby, Vigerslev, Vanløse, Brønshøj, and Husum on the Zealand side and Sundby on the Amager side, thus trebling its area and completely encircling the independent municipality of Frederiksberg.

The heart of the city is the Raadhuspladsen (Town Hall Square). From the square, an old crooked shopping street leads northeast to the former centre of the city, Kongens Nytorv (King's New Square), laid out in the 17th century. Buildings there include the Thott Palace (now the French Embassy) and the Charlottenborg Palace (now the Royal Academy of Arts), both of the 17th century, and the Royal Theatre, built in 1874. Kongens Nytorv touches on a narrow arm of the harbour lined with picturesque old gabled houses. From the southern end of the square a street called Holmens Kanal winds past the National Bank to the Holmens Kirke (the naval church), which has a chapel containing the tombs of the great admirals Niels Juel and Peder Tordenskiold.

Holmens Kanal crosses a bridge to Slotsholmen (Castle Islet), where stands the Christiansborg Palace, built on the site of the old castle founded by Bishop Absalon in 1167. Since 1928 the palace has been occupied by Parliament, the Supreme Court, and the Foreign Office. Nearby buildings house other government offices. Slotsholmen also contains the Bertel Thorvaldsen Museum, the Arsenal, the state archives, and the Royal Library.

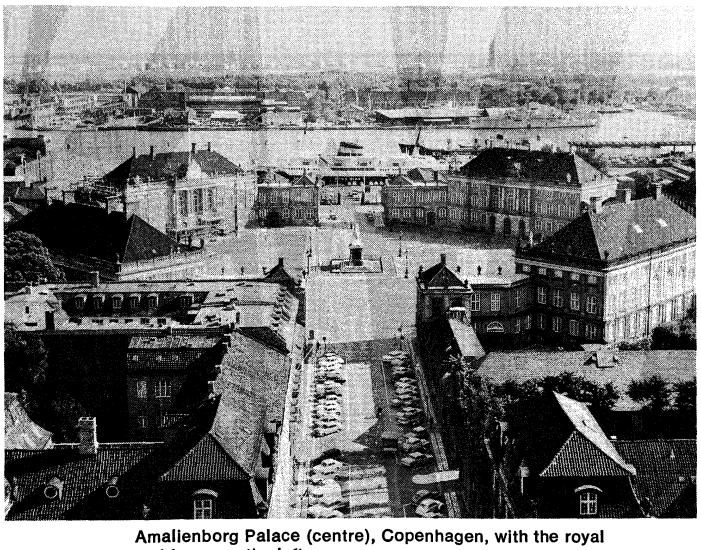
On the mainland to the west of Slotsholmen is the Prinsens Palace, now the National Museum. In Nytorv (New Market) is the former town hall built in 1815, now the city court, to the north of which is Gammeltorv (Old Market), the main square of the medieval town. From there an old street leads to the Vor Frue Kirke, the cathedral church of Copenhagen. Just north is the university, founded in 1479, and opposite it the Petri Kirke, in Gothic style, used after 1585 as a parish church for the German residents of the city. The area from Kongens Nytorv northeast to the 17th-century citadel includes the palace of Amalienborg and the castle and gardens of Rosenborg.

Southwest from the citadel a line of parks extends along the site of the former fortifications, skirted by wide boulevards. The botanical gardens laid out in 1874 have an observatory with a statue of the Danish astronomer Tycho Brahe. H.C. Andersens Boulevard runs past the Raadhuspladsen and crosses the harbour on the Langebro (Long Bridge) to residential districts on the island of Amager. It also skirts the Tivoli pleasure garden and passes the Ny Carlsberg Glyptotek, which has a fine collection of old and new art.

The old quarter of Christianshavn is on the harbour to the south. It contains the 17th-century Vor Frelsers Kirke (Church of Our Saviour). The western quarter contains the Frederiksberg Park, with its palace and a zoological garden.

Formerly a centre of trade and shipping, Copenhagen has also become a manufacturing city. In 1969 it had over 155,000 industrial workers, about 40 percent of the national total. One-third of all Danish factories are situated in the Greater Copenhagen area, and nearly half the working population is engaged in manufacturing and handicrafts, including shipbuilding, machinery production, and canning and brewing. Some of the internationally known firms include the East Asiatic Company, with its many subsidiaries; Burmeister & Wain's machinery and shipbuilding company; large engineering contractors such as F.L. Smidth & Co., Christiani & Nielsen, and Kampsax; the Carlsberg and Tuborg breweries; Peter F. Heering (Cherry Heering); the Royal Copenhagen Porcelain Factory and Bing and Grøndahl's porcelain factory; and the George Jensen silverworks.

Sven Samelius



Amalienborg Palace (centre), Copenhagen, with the royal residence on the left.

Municipal government

The constitution of the municipality of Copenhagen dates back to 1857. Power resides in the Town Council of 55 members, with an executive council consisting of a chief burgomeister, five other burgomeisters, and five aldermen, all elected by the Town Council. The suburban municipalities of Frederiksberg and Gentofte, as well as the 19 other suburbs, all have similar governmental structures.

Like other contemporary cities, Copenhagen has seen a shift of population and industry to the suburbs. The population of the city declined from about 975,000 in 1950 to less than 875,000 in 1965. At the same time there has been an increase of automobile traffic in all parts of the city. A number of arterial streets carry traffic toward the centre, across the three harbour bridges. There are electric railways (*S-baner*) for commuters and a network of city bus lines. The last streetcars disappeared in 1972. Copenhagen's airport is at Kastrup, about five miles southeast of the city.

The University of Copenhagen had over 20,000 students in 1971–72. In addition, Copenhagen has a number

of other institutions of higher education: the Technical University of Denmark, the Engineering Academy of Denmark, the Royal Danish Academy of Music, the Royal Veterinary and Agricultural College, the Royal Danish Academy of Fine Arts (established in 1754), a college of pharmacy, a dental college, several training colleges for the education of primary school teachers, the Copenhagen School of Economics and Business Administration, and several schools of commerce.

BIBLIOGRAPHY. Historical works (in Danish) include: OLUF NIELSEN, *Kjøbenhavns Historie og Beskrivelse*, 6 vol. (1877–92), a history and description of Copenhagen from its beginnings to about 1730; CARL BRUNN, *Kjøbenhavn: En illustreret Skildring af dets Historie, Mindesmaerker of Institutioner*, 3 pt. (1887–1901), on Copenhagen's history, memorials, and institutions; SVEND CEDERGREIN BECH, *Københavns historie gennem 800 år. København* (1967), a modern history of Copenhagen covering a period of 800 years; and H.U. RAMSING, *Københavns Historie og Topografi i Middelalderen*, 3 vol. (1940), a history of Copenhagen in the Middle Ages. There are very few books in English devoted wholly to this city. Among them are DANSKE SELSKAB, *Capital of a Democracy*, 3rd ed. (1964); OLIVER WARNER, *A Journey to the Northern Capitals* (1967); SYLVIA FURNESS, *Copenhagen* in the "Great Centers of Art Series" (1970); and MICHAEL BUSELLE, *Copenhagen* (1967), mainly photographs. Statistical data may be found in the *Statistical Yearbook* for the municipality of Copenhagen (1919–).

(S.L.)

Copernicus

A major contribution to Western thought was the publication in 1543 of *De revolutionibus orbium coelestium, libri VI* (Eng. trans., *On the Revolutions of the Celestial Spheres*, 1952; Latin reprint, 1965) by Nicolaus Copernicus, Polish astronomer, who is noted for the Copernican theory of the heavens. By attributing to the Earth a daily motion around its own axis and a yearly motion around the stationary Sun, Copernicus developed an idea that had far-reaching implications for the rise of modern science. Henceforth, the Earth could no longer be considered the centre of the cosmos; rather, as one celestial body among many, it became subject to mathematical description.

By courtesy of the Museum of Jagiellonian University, Cracow, Poland



Copernicus, 17th-century copy by an unknown artist of a 16th-century self-portrait. In the Museum of Jagiellonian University, Cracow, Poland

Education

Copernicus was born on February 19, 1473, at Toruń, near the Vistula River in eastern Poland, where his father was a merchant of social standing. In 1491 Copernicus entered the University of Cracow, where he became interested in the study of astronomy; he probably returned home in 1494 (or 1496). His maternal uncle, Lucas Waczenrode, newly elected bishop of Ermeland, wanted him to enter the canonry of Frauenburg in order to secure lifelong financial independence. While waiting for a vacancy to occur, he was sent by his uncle in 1497 for further training to the University of Bologna, where he associated himself with the German students.

For three and a half years Copernicus studied the Greek language, mathematics, and the writings of Plato; he also became further acquainted with the astronomical thought of the day. In Bologna he also made his first recorded observation of the heavens, an occultation (overlapping, or eclipse) of the star Aldebaran by the Moon on March 9, 1497; the light of the former was shut off by the Moon. The same year he was elected (by proxy) a canon of Frauenburg. He travelled to Rome in 1500 for the great jubilee celebration and may have given informal lectures in mathematics there. In 1501 he briefly visited Frauenburg to claim his post on the cathedral staff, returning promptly to Italy under special leave of absence to continue his studies at the University of Padua. There, enrolled with other Polish students, he studied both law and medicine. Except for a short interruption in 1503, when he was granted the degree of doctor of canon law by the University of Ferrara, he spent almost four years in Padua.

On returning to Poland in 1503, he visited Cracow and later acted as adviser to his uncle until the latter's death in 1512. Copernicus settled permanently at Frauenburg, where he acted as representative of the cathedral chapter, his medical skill being used particularly in aid of the indigent.

As a result of his studies in Cracow and Padua, Copernicus may be said to have mastered all the knowledge of the day in mathematics, astronomy, medicine, and theology. Copernicus appears to have planned a systematic program of astronomical work. Although he did not make extensive observations, he did enough to enable him to recalculate the major components of the supposed orbits of the Sun, Moon, and planets around the Earth. He published 27 such observations made during the years 1497–1529, and a few others have been found entered in books in his private library. He also published for his uncle in 1509 a Latin translation of Greek verses of Theophylactus, a Byzantine poet of the 7th century AD, and from 1519 to 1528 prepared an exposition of the principles of currency reform for certain Polish provinces; the latter was not published in Warsaw, however, until 1816.

Copernicus' fame as an industrious student of astronomy rapidly increased, and in 1514 he was invited to give his opinion on calendar reform, which was then being considered by the Lateran Council, a general meeting of the church authorities. He refused to express any firm views, for he felt that the positions of the Sun and Moon were not known with sufficient accuracy to permit a proper reassessment.

Dissatisfaction with the Ptolemaic system

Yet, as his studies progressed, Copernicus became increasingly dissatisfied with the Ptolemaic system of astronomy. He was not alone in this dissatisfaction; indeed, he himself said that the many divergent views prevalent in his day gave him cause for profound thought. Ptolemy's system, which contained not only original work but also a synthesis of the views of previous Greek philosophers, was basically geocentric and circular in conception. By the 16th century this geocentric interpretation of the heavens had become firmly entrenched in astronomical thought, virtually as an article of faith. Although certain Greek philosophers had suggested, as far back as the 3rd century BC, that the Sun—and not the Earth—was the centre of the universe, their ideas had not been widely accepted. Difficulties had arisen when ancient astronomers sought to account for the accumulated observations of the Sun, Moon, and planets. Accordingly, Ptolemy (*q.v.*) in the 2nd century AD had devised an elaborate geocentric model of the heavens composed of large circles, called deferents, and small circles, called epicycles. Each planet rode on the circumference of an epicycle, the centre of which revolved on the deferent. Ptolemy used this system to account for observed ir-

regularities of the planets, such as changes in brightness, and particularly for their puzzling retrogressive motions, when they seemed to stop and move backward and forward in a loop. Moreover, to account for observed variations in velocity, Ptolemy introduced the equant, which was an imaginary point in space where uniform, circular speed would indeed be observed. This system enabled astronomers to account for the phenomena and to make predictions. As observations in succeeding centuries became more accurate, however, it became increasingly difficult to compute the future positions of the heavenly bodies, and much of the flexibility and elegance of the Ptolemaic system was thereby lost.

Copernicus concluded that, in view of the many circles and their displacements from the center of the Earth that the Ptolemaic system required to account for the observed motions of heavenly bodies, a simpler, alternative explanation might be possible. In consequence, he read the works of many original Greek authors and found that, indeed, heliocentric ideas had been suggested. The idea of a moving Earth seemed absurd at first, but, when Copernicus applied this assumption, the result was an aesthetically superior, although not much simpler, system, even though, as might be expected, he still believed that the planets moved with uniform circular motion. After many years of mathematical calculations, he became convinced that his new idea was true, yet he made no attempt to publish.

From about 1510 to 1514, Copernicus prepared a short manuscript to summarize his new idea, *De hypothesibus motuum coelestium a se constitutis commentariolus*, which he privately circulated among friends in 1514. Its main points were: that the apparent daily motion of the stars, the annual motion of the Sun, and the retrogressive behaviour of the planets result from the Earth's daily rotation on its axis and yearly revolution around the Sun, which is stationary at the centre of the planetary system. The Earth, therefore, is not the centre of the universe but only of the Moon's orbit. As the years passed, he developed his argument with diagrams and mathematical calculations. Lectures on the principles expounded in the *Commentariolus* were given in Rome in 1533 before Pope Clement VII, who approved, and a formal request to publish was made to Copernicus in 1536. But he continued to hesitate. It was only through the efforts of his friends—in particular, his pupil and disciple Georg Joachim Rhäticus, who studied with him for two years—that he finally published his work. In 1540 Rhäticus was permitted to take the completed manuscript to Nürnberg, Germany, for printing. Because of opposition from Martin Luther, Philipp Melanchthon, and other reformers, Rhäticus left Nürnberg and went to Leipzig, where he passed on the task of publication to Andreas Osiander. Apparently fearing criticism of a treatise that proposed an annual motion of the Earth around a stationary Sun, Osiander, on his own responsibility, inserted a preface emphasizing that the hypothesis of a stationary Sun was only a convenient means for simplifying planetary computations.

The Copernican system

A careful examination of the text makes it clear, however, that Copernicus had really come to believe in the heliocentric system—rather, heliostatic, since he placed the Sun at some distance from the centre—as a true picture of the universe. He wrote *On the Revolutions of the Celestial Spheres*, in six sections, as a mathematical reinterpretation of Ptolemy. He wished to provide an alternative computational scheme that would make possible more accurate predictions that would be used in calendar reform and eclipses, and that would, at the same time, explain the troublesome variations of brightness, retrogressions, and velocity with a simpler geometric system of points and circles.

In the first section, Copernicus gave some basic mathematical rules, countered the old arguments about the fixity of the Earth, and discussed the order of the planets from the Sun. He could no longer accept the old arrangement—Earth, Moon, Mercury, Venus, Sun, Mars, Jupiter, and Saturn, since this had been a consequence of an Earth-centred, or geocentric, system. He found it necessary to adapt it to his Sun-centred, or heliocentric system and adopted the following order from the stationary Sun: Mercury, Venus, Earth with the Moon orbiting around it, Mars, Jupiter, and Saturn. In the second section, Copernicus applied the basic mathematical rules of the previous section to the apparent motions of the stars and planets, and attributed the motion of the Sun to the motion of the Earth. The third section contains a mathematical description of the Earth's motion, including the precession of the equinoxes, which is caused by the gyration of the Earth's axis. Sections four, five, and six deal with the motions of the Moon and of the five remaining planets.

In his heliocentric theory, Copernicus found himself able to describe the movements of the Moon and planets in a more elegant way than Ptolemy in his geocentric system. To fit the observations Ptolemy had been forced to offset the centres of regular motion a slight way from the centre of the Earth, and this Copernicus believed to conflict with the basic rule of true circular motion. In *De revolutionibus* the centres all lay at the centre of the Sun, although, because Copernicus still adopted circular motions at an unvarying speed, his system proved to be virtually as complex as Ptolemy's. Nevertheless, Copernicus believed that his system was aesthetically more satisfying and that it was a true picture of the divinely ordained cosmos.

A copy of the great work is believed to have been brought to Copernicus at Frauenburg on the last day of his life, May 24, 1543.

Assessment

The Copernican system appealed to many independent-minded astronomers and mathematicians. Its attraction was due not only to its elegance but also, in part, to its break with traditional doctrines: in particular, it opposed Aristotle, who had argued cogently for the fixity of the Earth; furthermore, it provided an alternative to Ptolemy's geocentric universe. In Western Christendom both of these views had been elevated almost to the level of religious dogma; to many thoughtful observers, however, they stifled development and were overdue for rejection.

Scientifically, the Copernican theory demanded two important changes in outlook. The first change had to do with the apparent size of the universe. The stars always appeared in precisely the same fixed positions, but if the Earth were in orbit round the Sun, they should display a small periodic change. Copernicus explained that the starry sphere was too far distant for the change to be detected. His theory thus led to the belief in a much larger universe than previously conceived and, in England, where the theory was openly accepted with enthusiasm, to the idea of an infinite universe with the stars scattered throughout space. The second change concerned the reason why bodies fall to the ground. Aristotle had taught that they fell to their "natural place," which was the centre of the universe. But because, according to the heliocentric theory, the Earth no longer coincided with the centre of the universe, a new explanation was needed. This re-examination of the laws governing falling bodies led eventually to Newton's concept of universal gravitation.

The dethronement of the Earth from the centre of the universe caused profound shock. No longer could the Earth be considered the epitome of creation, for it was only a planet like the other planets. No longer was the Earth the centre of all change and decay with the changeless universe encompassing it. And the belief in a correspondence between man, the microcosm, as a mirror of the surrounding universe, the macrocosm, was no longer valid. The successful challenge to the entire system of ancient authority required a complete change in man's philosophical conception of the universe. This is what is rightly termed "the Copernican Revolution."

BIBLIOGRAPHY. For the general reader two short biographies are available: ANGUS ARMITAGE, *Sun, Stand Thou Still* (1947); and JOZEF RUDNICKI, *Nicholas Copernicus* (1943); the latter was written for the 400th celebration of the publication of *De revolutionibus orbium coelestium* and is particularly well illustrated. ANGUS ARMITAGE, *Copernicus: The Founder*

of Modern Astronomy (1938), is a more detailed biography. For an analysis of the impact of Copernicus' ideas, see THOMAS S. KUHN, *The Copernican Revolution* (1957). *De revolutionibus* has been translated by CHARLES G. WALLIS and published in "Great Books of the Western World," vol. 16 (1952). The specialized reader might refer to EDWARD ROSEN, *Three Copernican Treatises*, 3rd rev. ed. (1971). For a thorough discussion of Copernicus' work, see J.L.E. DREYER, *A History of Astronomy from Thales to Kepler*, 2nd ed. (1953); and ALEXANDRE KOYRE, *From the Closed World to the Infinite Universe* (1957). For the reception of the Copernican theory in England, see F.R. JOHNSON, *Astronomical Thought in Renaissance England* (1937, reprinted 1968).

(C.A.R.)

Copland, Aaron

Generally recognized as one of the significant musical artists of this century, Aaron Copland succeeded so well in assimilating the materials of American folksong into his own highly personal style that, in the eyes of the world, he came to be regarded as *the* American composer of his time.

The son of a Russian-Jewish immigrant to America, Copland was born on November 14, 1900, in Brooklyn and attended public schools there. "I was born," he wrote, "on a street that can only be described as drab. Music was the last thing anyone would have connected with it." Yet an older sister taught him the piano, and by the time he was 15 he had decided to become a composer. As a first step Copland tried to learn harmony through a correspondence course. Haltingly and in an environment not particularly conducive to art, he struggled toward his goal.

By courtesy of the Boston Symphony Orchestra



Copland.

Years in Paris

In the summer of 1921 Copland attended the newly founded school for Americans at Fontainebleau, where he came under the influence of Nadia Boulanger, a brilliant teacher who shaped the outlook of an entire generation of American musicians. He decided to stay on in Paris, where he became Boulanger's first American student in composition. During his three years in Paris Copland's music became more and more experimental. He returned to New York in 1924 with an important commission: Nadia Boulanger had asked him to write an organ concerto for her American appearances. Copland composed the piece while working as the pianist of a hotel trio at a summer resort in Pennsylvania. That season the *Symphony for Organ and Orchestra* had its premiere in Carnegie Hall with the New York Symphony under the direction of the composer and conductor Walter Damrosch.

In his growth as a composer Copland mirrored the important trends of his time. After his return from Paris, he worked with jazz rhythms in *Music for the Theater* (1925) and the *Piano Concerto* (1926). There followed a period during which he was strongly influenced by Stravinsky's Neoclassicism, turning toward an abstract style he described as "more spare in sonority, more lean in texture." This outlook prevailed in the *Piano Variations* (1930), *Short Symphony* (1933), and *Statements for Orchestra* (1933–35). After this last work, there occurred a change of direction that was to usher in the most productive phase of Copland's career. He well summed up the new orientation: "During these years I began to feel an increasing dissatisfaction with the relations of the music-loving public and the living composer. It seemed to me that we composers were in danger of working in a vacuum." Furthermore, he realized that a new public for modern music was being created by the new media of radio, phonograph, and film scores: "It made no sense to ignore them and to continue writing as if they did not exist. I felt that it was worth the effort to see if I couldn't say what I had to say in the simplest possible terms." Copland therefore was led to what became a most significant development after the 1930s: the attempt to simplify the new music in order that it would have meaning for a large public.

Mature career

The decade that followed saw the production of the scores that spread Copland's fame throughout the world. Most important of these were three ballets based on American folk material: *Billy the Kid* (1938), *Rodeo* (1942), and *Appalachian Spring* (1944). To this group belong also *El Salón México* (1936), an orchestral piece based on Mexican melodies and rhythms; two works for high-school students—the "play opera" *The Second Hurricane* (1937) and *An Outdoor Overture* (1938); and a series of film scores, of which the best known are *Of Mice and Men* (1939), *Our Town* (1940), *The Red Pony* (1948), and *The Heiress* (1949). Typical too of the Copland style are two major works that were written in time of war—*Lincoln Portrait* (1942), for speaker and chorus, on a text drawn from Lincoln's speeches, and *Letter from Home* (1944), as well as the melodious *Third Symphony* (1946).

In his later years Copland refined his treatment of Americana. "I no longer feel the need of seeking out conscious Americanism. Because we live here and work here, we can be certain that when our music is mature it will also be American in quality." His later works include an opera, *The Tender Land* (1954); *Twelve Poems of Emily Dickinson*, for voice and piano (1950); and the delightful *Nonet* of 1960. During these years Copland also produced a number of works in which he showed himself increasingly receptive to the serial techniques of the so-called twelve-tone school of composer Arnold Schoenberg. Notable among such works are the stark and dissonant *Fantasy for Piano* (1957) and *Connotations for Orchestra* (1962), which was commissioned for the opening of Lincoln Center for the Performing Arts in New York City.

For the better part of four decades, as composer, teacher, writer of books and articles on music, organizer of musical events, and more recently as a much sought after conductor, Copland has succeeded in expressing "the deepest reactions of the American consciousness to the American scene."

MAJOR WORKS

OPERAS: *The Second Hurricane* (1937); *The Tender Land* (1954).

BALLETS: *Hear Ye, Hear Ye* (1934); *Billy the Kid* (1938); *Rodeo* (1942); *Appalachian Spring* (1944); *Dance Panels* (1959).

ORCHESTRAL MUSIC: *Symphony for Organ and Orchestra* (1924); *Piano Concerto* (1926); three symphonies (1928, 1933, and 1946); *El Sálon México* (1936); *Billy the Kid* (1938), suite from the ballet; *An Outdoor Overture* (1938); *Rodeo* (1942), four dance episodes from the ballet; *Lincoln Portrait*, for speaker and orchestra (1942); *Letter from Home* (1944, rev. 1962); *Appalachian Spring* (1945), suite from the ballet; *Concerto for Clarinet and String Orchestra*, with harp and piano (1948); *The Tender Land: Orchestral Suite* (1957).

WORKS FOR BAND: *Billy the Kid* (1938); *Variations on a Shaker Melody* (1956); *Emblems*, for symphonic band (1964).

CHAMBER MUSIC: *Two Pieces for String Quartet* (1928); *Quartet for Piano and Strings* (1950); *Nonet for Strings* (1960).

VOCAL MUSIC (SONGS): *Twelve Poems of Emily Dickinson* (1950); *The Tender Land: "Laurie's Song"* (1954).
CHORAL MUSIC: *Old American Songs*, two sets (1950, 1952); *Canticle of Freedom* (1955, rev. 1965); *The Promise of Living* (1964).
FILM SCORES: *The City* (1939); *Of Mice and Men* (1939); *Our Town* (1940); *The Cummington Story* (1945); *The Red Pony* (1948); *The Heiress* (1949); *Something Wild* (1961).

BIBLIOGRAPHY. AARON COPLAND, *What to Listen for in Music* (1939, *The New Music, 1900–1960*, rev. ed. (1968), *Music and Imagination* (1952), and *Copland on Music* (1960), works that display Copland's writing style and his unusual ability to explain sophisticated musical concepts to the general public; JULIA SMITH, *Aaron Copland* (1955), a comprehensive account of the composer's life and works; ARTHUR BERGER, *Aaron Copland* (1953), a carefully detailed analysis of Copland's music.

(J.Ma.)

Copper Products and Production

Copper is one of the most useful and widely used of the metallic elements. This article describes copper mining, refining and recovery, the production and uses of the metal and its alloys, chemical compounds of copper, and its economic importance. For information on the element copper, its physical and chemical properties and occurrence, see TRANSITION ELEMENTS AND THEIR COMPOUNDS; NATIVE ELEMENTS; and ORE DEPOSITS.

HISTORY

Copper was discovered and first used by Neolithic man during the Late Stone Age. Though the exact time of this discovery will probably never be known, it is believed to have been about 8000 BC. Copper is found in the free metallic state in nature; this native copper is the material that Neolithic man employed as a substitute for stone. From it he fashioned crude hammers and knives and, later, other utensils. The malleability of the material made it relatively simple to shape implements by beating the metal. Pounding hardened the copper so that more durable edges resulted; the bright reddish colour of the metal and its durability made it highly prized.

Discovery of casting and reducing

The search for copper during this early period led to the discovery and the working of deposits of native copper. Sometime after 6000 BC the discovery was made that the metal could be melted in the campfire and cast into the desired shape. Then followed the discovery of the relation of metallic copper to copper-bearing rock and the possibility of reducing ores to the metal by the use of fire and charcoal. This was the dawn of the metallic age and the birth of metallurgy.

The early development of copper probably was most advanced in Egypt. As early as 5000 BC copper weapons and implements were left in graves for the use of the dead. Definite records have been found of the working of copper mines on the Sinai Peninsula about 3800 BC, and the discovery of crucibles at these mines indicates that the art of extracting the metal included some refining. Copper was hammered into thin sheets and the sheets formed into pipes and other objects. During this period bronze first appeared. The oldest known piece of this material is a bronze rod found in the pyramid at Maydūm (Medum), the date of origin being generally accepted as about 3700 BC.

Qualities of bronze

Bronze, an alloy of copper and tin, is both harder and tougher than either; it was widely employed to fashion weapons and objects of art. The period of its extensive and characteristic use has been designated the Bronze Age. From Egypt the use of bronze rapidly spread over the Mediterranean area: to Crete in 3000 BC, to Sicily in 2500 BC, to France and other parts of Europe in 2000 BC, and to Britain and the Scandinavian area in 1800 BC.

About 3000 BC copper was produced extensively on the island of Cyprus. The copper deposits there were highly prized by the successive masters of the island—Egyptians, Assyrians, Phoenicians, Greeks, Persians, and Romans. Cyprus was almost the sole source of copper to the Romans, who called it *aes cyprium* ("ore of Cyprus"), which was shortened to *cyprium* and later corrupted to cuprum. From this name comes the English name copper. The first two letters of the Latin name constitute the chemical symbol Cu.

When copper and bronze were first used in Asia is not known. The epics of *Shu Ching* mention the use of copper in China as early as 2500 BC, but nothing is known of the state of the art at that time or of the use of the metal prior to that time. Bronze vessels of great beauty made during the Shang dynasty, 1766–1122 BC, have been found, indicating an advanced art. The source of the metals, however, is unknown.

The Copper Age in the Americas probably dawned between AD 100 and 200. Native copper was mined and used extensively and, though some bronze appeared in South America, its use developed slowly until after the arrival of Columbus and other European explorers. Both North and South America passed more or less directly from the Copper Age into the Iron Age.

Later uses for copper

As man learned to fashion his weapons from iron and steel, copper began to assume another role. Being a durable metal and possessed of great beauty, it was used extensively for household utensils, water pipes, and for marine uses and other purposes that required resistance to corrosion. The unusual ability of this metal to conduct electric current accounts for its greatest use during the 20th century.

MINING TECHNIQUES

Open-pit and underground mining. Though it sometimes occurs in native form, copper is most often found mixed with other minerals in the form of various ores (see below). In these ores, the amount of copper may vary from less than 1 percent to more than 10 percent. When the ore contains small quantities of the desired metal, very large amounts of rock must be mined and processed to recover a profitable quantity of metallic copper. In deposits of this type, the open-pit method, involving the removal of ore from extensive deposits exposed by the large open pits, is used. In cases in which the percentage of copper in the ore is considerably higher and the deposit far below the surface, the underground method, which requires sinking a shaft to the deposit, is employed. For a more complete description of these mining methods, see MINING AND QUARRYING.

Important ores. Principal forms in which copper ores are found include native copper, porphyry copper, massive deposits, and mixed ores. Native copper is simply the metal found unadulterated in nature. Porphyry copper deposits, in which the copper materials are more or less uniformly scattered throughout the rock, account for the greatest tonnage of metal in the producing areas of the world. The copper minerals in the upper portions of such deposits are in general oxides (copper chemically combined with oxygen), those in the lower levels sulfides (copper with sulfur). The host rock is porphyry, schist, or other rock. Massive deposits are of higher metal content but of more limited extent and may be oxidized in the upper portion with sulfides lower down. In mixed ores, nickel, zinc, or lead usually accompany the copper; when such ore is mined, these other metals also are refined and sold as by-products. Table 1 shows the ore minerals of copper and their compositions.

ORE PROCESSING

Copper metal of very high purity is recovered from mined rock by a three-step process. Since much copper ore is found mixed with large amounts of material that does not contain copper, the first step (ore dressing) involves separating the ore from this worthless rock or gangue. The second step frees copper metal from the concentrate; *i.e.*, separates it from other elements, such as sulfur and oxygen, with which it is chemically combined. The third step, refining, involves removing impurities from the copper metal in order to produce a metal of very high purity.

Ore dressing. Copper ores usually are treated either by smelting (pyrometallurgical methods) or by leaching (hydrometallurgical methods). Generally speaking, sulfide ores are first concentrated by the selective flotation method described below, while oxide ores are leached.

Table 1: General Classification of the Ore Minerals of Copper

	formula	copper (percent)*
Native copper ore		
Native copper	Cu	99.9
Sulfide ores		
Chalcocite	Cu_2S	79.9
Covellite	CuS	66.5
Chalcopyrite	$CuFeS_2$	34.6
Bornite	Cu_5FeS_4	63.3
Enargite	Cu_3AsS_4	48.4
Tetrahedrite	Cu_3SbS_2	46.7
Oxide ores		
Cuprite	Cu_2O	88.8
Tenorite	CuO	79.9
Malachite	$CuCO_3 \cdot Cu(OH)_2$	57.5
Azurite	$2CuCO_3 \cdot Cu(OH)_2$	55.3
Chalcanthite	$CuSO_4 \cdot 5H_2O$	25.5
Brochantite	$CuSO_4 \cdot 3Cu(OH)_2$	56.2

*Approximate.

Selective flotation. In selective flotation the finely ground ore, mixed with water and selected reagents, is violently agitated with air to produce a heavy froth. The reagents are so selected that a great attraction exists between the surface of the valuable mineral particles and the air bubbles, with only a small attraction between the gangue particles, which are readily wetted by the water, and the air. As a result, the mineral particles cling to the bubbles of the froth and are carried to the surface while the gangue particles sink to the bottom of the container. Removing the froth then separates the minerals from the gangue material. Perfect separation is impossible, but with constant supervision, repeated treatment, and exacting control it is possible to concentrate 95 percent or more of the copper ore into 10 to 20 percent of the original weight and to eliminate most of the worthless gangue.

Roasting. In roasting, the copper concentrate is heated in air to expel volatile matter, such as arsenic and antimony, and to expel part of the sulfur from sulfide ores by oxidation to produce sulfur dioxide gas and copper oxide. The material that results from roasting, known as calcine, is a fine mixture of the sulfides and oxides of iron and copper, gangue material, and nonvolatile impurities. The total sulfur content has been reduced to that desired for the subsequent smelting operation.

Metal recovery. To separate metallic copper from the substances with which it is associated in copper ores, smelting of matte or electrolytic techniques may be employed. Sometimes copper is recovered from scrap.

Matte smelting and converting. The object of the first step in smelting is to produce a molten, artificial sulfide of copper and iron. This material, known as matte, should contain all the copper and the desired amount of iron and must be sufficiently heavy so that when melted, it will separate from the gangue material, other undesirable compounds, and the balance of the iron. When metallic oxides fuse in the presence of fire, they react as bases and will combine with acid anhydrides to form stable compounds of relatively low weight. These compounds, known as slags, are usually formed by the use of silica, SiO_2, as the acid anhydride. The formation temperature and fluidity of the slag is a function of its composition. The action of the slag on the furnace lining is a function of its acidity. It is necessary then to balance the composition of the slag to achieve the desired results. This is done by adding to the furnace selected materials to give a slag of the desired composition and properties. These materials are known as fluxes; the most common are limestone and silica.

Matte is generally smelted in large oblong furnaces from 20 to 30 feet (six to nine metres) in width and from 90 to 130 feet (27 to 40 metres) in length, constructed with heavy, silica brick side walls and a low roof. These are called reverberatory furnaces because heat is radiated from the roof onto the material treated. Matte may also be smelted in blast furnaces or electric furnaces.

Producing blister copper

The conversion of matte involves the transformation of molten matte into copper metal (blister copper). Matte converting is based on the fact that copper has a lower affinity for oxygen than has iron or sulfur and the fact that the oxidation of iron and sulfur liberates large quantities of heat. Molten matte can be oxidized by introducing into it a stream of air. At the point at which the air enters, iron, sulfur, and copper are oxidized. The copper oxide immediately reacts with any iron sulfide still present to reform the sulfide of copper and form the oxide of iron. As air continues to flow through the molten mass, a time arrives when all the iron is present as the oxide and all the copper as the sulfide, and the sulfur that was originally with the iron has left the charge as sulfur dioxide. Since iron oxide forms a fusible slag with silica, it is necessary only to add to the molten mass sufficient silica to form the desired slag, and all the iron may be removed as an iron-silicate slag, while the heavy copper sulfide remains behind. The oxidation of the iron and its sulfur liberates sufficient heat to keep the material molten and to melt the necessary silica and any other solid charge that it may be necessary to add. If air continues to enter the charge after the removal of the slag, the oxidation of the sulfur and the copper continues, and the copper oxide so formed at once reacts with the copper sulfide remaining in the charge to form sulfur dioxide and metallic copper. These reactions liberate just sufficient heat to keep the charge molten. Ultimately, all sulfur is oxidized, and metallic copper and a small amount of copper oxide are present in the charge. Small amounts of minor impurities are also present.

Hydrometallurgy. Hydrometallurgy, or leaching, involves the treatment of the ore with a suitable solvent that takes the copper compound into solution and leaves all or a major part of the undesirable material unaltered. The copper is then recovered in relatively pure form from the solution.

The oxide ores of copper lend themselves to leaching. Sulfuric acid readily dissolves these oxides and has little effect on the common gangue materials, except for some of the salts of iron. Many oxide ores are not amenable to concentration, so it is necessary to treat or leach the ore as mined. The material is agitated or soaked with the leach solution. Elaborate systems of washing and filtering or settling are employed to recover the optimum amount of copper. After use, the leaching solutions are usually subjected to some type of purification for removal of the soluble iron and other objectionable impurities. This often involves neutralization and oxidation followed by removal of the precipitated ferric salts. For a more extensive description of hydrometallurgy, see METALLURGY.

Electrolytic decomposition of copper sulfate

Electrolytic recovery. When copper ores are leached with sulfuric acid, the metal may be recovered from the leach solution, in which it exists in the form of copper sulfate, by passing an electric current through the solution (electrolysis); this breaks down the sulfate and yields copper. An insoluble anode (positive terminal of an electrolytic cell) and a copper cathode (negative terminal) are used in the process; they are immersed in the leaching solution (copper sulfate), and an electric current is passed through the cell, and copper from the solution is deposited on the cathode.

Recovery from scrap. Because of its durability, large quantities of scrap or reclaimed copper become available each year from obsolete or discarded machinery and equipment. This material is readily converted into usable metal by its addition to smelter feed, or by remelting and refining (see also MATERIALS SALVAGE).

In the early 1970s pollution problems and economic considerations were bringing some shifts in the use of various methods of ore dressing and metal recovery. Smelting was under criticism in many places because of the release of sulfur dioxide to the atmosphere. Continuous smelting processes that minimize the problem were developed in Australia and Canada, and a combined roasting and smelting process, called flash smelting, in Finland and Canada. Higher prices for copper at the same time were encouraging the use of leaching, effec-

tive in exploiting low-grade copper-bearing material. In this procedure, dilute sulfuric acid is allowed to trickle through the ore, extracting the copper in the form of a solution of copper sulfate. The copper is recovered from the solution by precipitation on scrap iron.

Refining. Refining involves the removal of impurities from copper metal to yield a product containing less than $\frac{1}{10}$ of 1 percent of impurities.

Fire refining. Blister copper contains copper oxide and small quantities of other impurities. Some of these impurities can best be removed by fire refining, which is done either in small reverberatory furnaces or in revolving furnaces, equipped with fuel burners that melt the charge, if necessary, and maintain it in a molten condition throughout the refining operation. Air is forced through the molten material to ensure complete oxidation of all impurities, and then the oxides are allowed to rise to the surface of the quiet pool from which they are skimmed. The oxidizing treatment is followed by a reducing treatment known as poling, in which the ends of green logs are forced into the pool of molten metal. The highly reducing gases resulting from the destructive distillation of the green logs reduce (remove the oxygen from) most of the copper oxide present in the metal. This reduction also can be effected by bubbling reformed natural gas through the bath. During this operation, frequent samples are examined to determine the degree of deoxidation. When the proper stage has been reached, the sample has a distinctive metallic rose colour. This metal is called tough-pitch copper, and may be readily cast into dense slabs.

Oxidizing impurities

Commercial tough-pitch copper is usable in this form. It still contains, however, any gold and silver recoverable from the original ore and traces of other impurities. The gold and silver are often worth recovering, and the removal of other impurities to produce the pure metal can be brought about by electrolytic refining.

Electrolytic refining. When an impure copper anode and a copper cathode are immersed in a solution of copper sulfate and sulfuric acid and an electric current is passed through the cell, copper ions from the solution migrate to the cathode and are deposited as metallic copper, and copper from the anode enters the solution. For each ion deposited on the cathode, an ion enters the solution from the anode. Impurities in the anode are liberated and settle to the bottom of the cell as a slime or accumulate in the solution. The net result is the consumption of the impure anode and the deposition of pure copper on the cathode along with the formation of a cell slime that contains the gold, silver, and other insoluble components. (A.W.S.)

THE METAL AND ITS ALLOYS: PRODUCTION AND USES

The major portion of the world's production of copper is utilized by the electrical industries; most of the remainder is combined with other metals to form alloys.

Essentially pure metal. Typical samples of electrolytic copper contain from 99.92 percent to 99.96 percent copper. About 0.03 percent oxygen is purposely left in the copper, since this amount slightly improves density and conductivity. Copper in this condition has a conductivity of 100 percent to 102 percent of the International Annealed Copper Standard. On this standard, 100 percent denotes a resistance of 0.15328 ohm for a length of one metre (39.37 inches) weighing one gram ($\frac{1}{28}$ ounce) at 20° C (68° F); this standard has been universally adopted for industrial purposes.

Electrical conductors. For making copper wire, electrolytic copper is first cast into wirebars, which are made in several standard sizes varying in weight from 135 to 500 pounds (60 to 225 kilograms). The wirebars are then reheated to 700° to 850° C (1,290° to 1,560° F) and are rolled without further reheating to rods approximately $\frac{3}{8}$ inch or about ten millimetres in diameter. The rod is drawn cold into wire, through dies of successively smaller diameters until the desired size is reached. The dies are usually of tungsten carbide; for finer wires, diamond dies are used.

Making copper wire

Much copper wire is marketed in the form of bare coils; a considerable tonnage is subsequently covered with paper, fabric, rubber, plastic, or other insulating material for use in the form of covered conductors. Much of the wire is also supplied stranded; all of these operations are carried out on special machines that are largely automatic in operation.

Copper cables are often covered to render them resistant to moisture. Lead covering, extruded onto the outside of the cable by means of a special apparatus, often is used; hemp or metal armouring sometimes may be added for additional protection.

The electrical industries also use large quantities of bare copper strip for incorporation in electrical machinery. This ribbonlike form of the metal is produced mainly from wirebars that are rolled in a mill similar to that used for the production of wire. Copper strip of greater width and much thinner gauge also is produced in long lengths and supplied in the form of coils.

Sheet and strip. The term copper strip as distinct from copper sheet is usually applied to material less than 24 inches (61 centimetres) wide that is supplied in long lengths. The majority of the strip used is less than 12 inches (30 centimetres) wide. In the preliminary stages of manufacture, the copper castings are rolled hot, but in the later stages of manufacture all the rolling is carried out cold, the material being coiled on coiling drums on each side of the rolling mills. Material produced by this method is of very even gauge and possesses an exceptionally good surface finish. The coils can be handled easily and are in general use for the manufacture of stampings in the electrical and other industries. Copper strip is supplied in various degrees of hardness according to the rolling it has received subsequent to the last annealing. These various tempers are selected according to the amount of subsequent mechanical deformation to which they will be subjected.

Hardness of copper strip

Copper sheets are produced by somewhat similar methods of manufacture. In the United States the majority of copper sheets are made from electrolytic copper, but in Europe fire-refined arsenical copper is used frequently in the manufacture of sheets and plates.

Alloys. In variety of uses, the alloys of copper surpass all other nonferrous alloys and comprise mixtures of copper with zinc, tin, nickel, aluminum, lead, manganese, and other elements.

Brass. Brass is an alloy consisting mainly if not exclusively of copper and zinc. The brasses may be conveniently divided into two groups in terms of their malleability, the dividing line being approximately the composition of 55 percent copper and 45 percent zinc. All the higher copper alloys are workable either hot or cold and in some cases both hot and cold, while the remainder are not malleable at all. The unworkable brasses, known as the white brasses, are not industrially important.

The general mechanical properties of brass vary widely; indeed, it is this wide range in tensile strength, elongation, and hardness that makes brass such an important alloy. Brass is readily drawn into fine wire, rolled into very thin strips, or drawn into tubes and extruded as rods or sections (see also ZINC PRODUCTS AND PRODUCTION).

Bronze. Bronze, an alloy formed by adding tin to copper, fuses at a lower temperature than copper and is thus better suited for casting; it also is harder and less malleable. A soft bronze or gunmetal is formed from 16 parts of copper to one of tin; a harder gunmetal, such as that used in the past for bronze cannon, contained about eight parts copper for each part of tin.

Phosphor bronze

Bronze is made harder and stronger when it is alloyed with phosphorus. Alloys prepared in this way, known as phosphor bronzes, may contain only about 1 percent of phosphorus in the ingot and a mere trace after casting, but their value is nevertheless enhanced for purposes in which a hard, strong metal is required, as for pumps, plungers, valves, and the bushings of bearings. Bronze is also improved by the presence of manganese in small quantities. Various grades of manganese bronze, in which there is little or no tin but a considerable percentage of zinc, are used in mechanical engineering (see also TIN PRODUCTS AND PRODUCTION).

Nickel. Because copper and nickel are completely miscible (mix thoroughly) in the solid state, forming a complete series of solid solutions, the useful range of alloys is not confined within any definite limits of composition, although certain compositions have come into general use. Additions of 2 to 45 percent nickel to copper provide a series of alloys that are considerably stronger and more resistant to oxidation at high temperatures than is copper. Of these cupronickels, the one containing 30 percent nickel is the most important; it is widely used for steam-condenser tubes.

The alloy formed of 20 percent nickel with the remainder copper is one of the most ductile of commercial alloys and may be subjected to the most severe cold-working without the need of any intermediate annealing. It is also readily forged and rolled at a temperature above 800° C (1,500° F). These properties make it a very suitable alloy for drop forgings and cold stamping and pressing. It has also found a variety of uses in automobile construction for exposed fittings, as it takes a high polish and is resistant to atmospheric tarnishing. Other uses include bullet sheathing, a widespread application. Another alloy in this series, containing either 45 percent or 40 percent nickel, became widely known under the name constantan. It has a high electrical resistance, which remains almost constant over a wide temperature range.

Monel metal

Monel metal is a so-called natural alloy prepared by the reduction of a copper-nickel ore; it contains 65 to 70 percent nickel, iron and manganese in small amounts, and certain impurities that influence its properties to some extent. It has been widely used for various engineering and ornamental purposes, and possesses exceptionally high strength at both normal and elevated temperatures. Alloys of similar nickel content are also manufactured by melting nickel and copper together.

Beryllium-copper. Unlike many kinds of steel, most copper alloys are not susceptible to improvements of hardness and strength by processes of heat treatment. One useful exception is the heat-treatable alloy beryllium-copper. This consists of copper with addition of about 2 percent of beryllium, with or without a smaller addition of nickel or cobalt.

When beryllium-copper is heated to about 800° C (1,500° F), quenched in cold water and then reheated to 275° C (525° F), it develops a tensile strength comparable to some of the stronger varieties of steel.

Other copper alloys. Copper also forms an important series of alloys with aluminum, classed under the general term aluminum bronzes. They may be classified into two main groups: those containing up to 7.5 percent aluminum are extremely ductile, whereas those containing 8 to 11 percent possess high tensile strength in the cast state. The ductile alloys containing less than 7.5 percent are especially useful for deep stamping, spinning, and severe cold-working of all kinds. They are useful as a substitute for brass and possess greater strength and resistance to atmospheric corrosion.

Silicon bronze usually contains about 96 percent copper. The remainder may be silicon alone, but more often a little manganese, tin, iron, or zinc is also added. These alloys were developed originally for the chemical industry because of their exceptional resistance to corrosion in many liquids. Their application later extended far beyond this field, chiefly because of their good casting qualities, strength, hardness, and ease of welding.

Manganese bronze

Manganese bronze is made in several varieties, exhibiting a range of compositions and properties. One type is in reality a brass to which a very little manganese has been added as a deoxidizer, less than 0.5 percent manganese remaining in the alloy. Another kind contains 2 to 5 percent manganese together with 2 to 4 percent iron and 3 to 7.5 percent aluminum. It has exceptionally high strength and is called high-tensile manganese bronze, or manganese-aluminum bronze. (A.Bu.)

CHEMICAL COMPOUNDS: PREPARATION AND USES

Copper forms two series of compounds, one in which it exhibits a valence (degree of combining power) of 1 (cuprous) and the other a valence of 2 (cupric); several unstable compounds in which a valence of 3 is exhibited are also known. Since the cuprous ion is unstable in aqueous solution, its salts readily decompose to form the metal and cupric salts.

Oxides. Copper forms two oxides in accordance with its two valences: cuprous oxide, Cu_2O, and cupric oxide, CuO. Cuprous oxide, a red crystalline material, can be produced by electrolytic or furnace methods. It is reduced readily by hydrogen, carbon monoxide, charcoal, or iron to metallic copper. It imparts a red colour to glass and is used for antifouling paints. It is soluble in mineral acids to form colourless cuprous salts, most of which rapidly oxidize to the cupric state. Cupric oxide, a black powder, can be prepared by the ignition of suitable salts such as the carbonate, the hydroxide, or the nitrate of copper, or by heating of cuprous oxide. This compound oxidizes carbon compounds and finds a wide laboratory and commercial use for this purpose. Since it imparts a green colour to glass, it is used extensively for that purpose. It is soluble in mineral acids and forms with them blue or green solutions.

Halides. Cuprous chloride, CuCl, can be prepared by treating metallic copper and cuprous oxide with hydrochloric acid or by treating metallic copper and cupric chloride with hydrochloric acid. The hydrochloric acid solution of cuprous chloride readily absorbs carbon monoxide and acetylene and is used for this purpose in gas analysis. Cupric chloride, $CuCl_2$, can be prepared by dissolving cupric oxide in hydrochloric acid. This material finds some use as the base salt for the manufacture of pigments. Cuprous iodide, CuI, is prepared by the direct combination of copper and iodine. Cupric iodide, CuI_2, exists only in combination with ammonium salts or in complex organic compounds.

Blue vitriol

Sulfates. Cupric sulfate, $CuSO_4$, commonly known as blue vitriol, is the most important salt of copper. It usually crystallizes as $CuSO_4 \cdot 5H_2O$ and has a bright blue colour. It is prepared by the treatment of copper oxides with sulfuric acid. While readily soluble in water it is insoluble in alcohol. The anhydrous salt is hygroscopic and is useful as a desiccating agent. Copper is readily displaced from aqueous solutions of the salt by metallic iron. Copper sulfate is the basic salt in the electrolytic refining of copper, and it also finds a wide use in the preparation of pigments.

Carbonates. Basic copper carbonates are formed when an alkaline carbonate is added to the solution of a copper salt. These compounds, which have a bright blue or green colour and are used in the preparation of pigments, occur in nature as the minerals azurite and malachite.

Table 2: World Mine Production of Copper by Major Producing Countries
(1970 preliminary)

	amount (short tons)	total production (percent)
United States	1,719,657	26.2
Chile	755,700	11.5
Zambia	753,968	11.5
Canada	676,006	10.3
U.S.S.R.	630,000	9.6
Zaire (Congo [Kinshasa])	425,138	6.5
Peru	233,751	3.6
South Africa	164,469	2.5
Australia	160,551	2.4
Philippines	160,303	2.4
Japan	136,851	2.1
China, People's Republic of	110,000	1.7
Yugoslavia	100,099	1.5
Poland	79,000	1.2
Mexico	68,361	1.0
Bulgaria	44,000	0.7
Finland	34,333	0.5
Turkey	30,010	0.5
South West Africa	29,406	0.4
Sweden	25,500	0.4
Others	229,540	3.5
World total	6,566,643	100.0

Source: U.S. Department of the Interior, Bureau of Mines, *Minerals Yearbook, 1970.*

Other compounds. Copper forms a series of salts with arsenic; all are bright green in colour and poisonous. Wide use has been made of these compounds in insecticides; some copper salts are used as superficial antiseptics. Cupric nitrate, $Cu(NO_3)_2$, can be prepared by dissolving metallic copper in nitric acid. The hydrated crystals are deep blue in colour. Copper silicates occur in nature and impart blue colour to many minerals. (A.W.S.)

ECONOMIC IMPORTANCE

Although commercial deposits of copper ores occur in almost every continent, five general areas contain about 75 percent of the world's known reserve. These areas are: the Rocky Mountain and Great Basin area of the United States; the west slope of the Andes mountains in Chile and Peru; the central plateau of Africa in Zaire (formerly Congo [Kinshasa]) and Zambia; the Precambrian shield area of central Canada and its extension into northern Michigan; and the Soviet Union. These areas have been listed in order of importance based on past production tonnages. The greatest known reserve of copper ore in one body is the deposit at Chuquicamata, Chile. Many lesser deposits are being exploited, such as those in Alaska, China, Australia, and Europe.

Table 2 shows the most important producing nations and their percentage of the world total.

BIBLIOGRAPHY. PAUL E. QUENEAU (ed.), *Extractive Metallurgy of Copper, Nickel, and Cobalt* (1961), a collection of technical papers with a valuable bibliography of over 1,000 references, the majority referring to copper; J.N. ANDERSON and PAUL E. QUENEAU (eds.), *Pyrometallurgical Processes in Nonferrous Metallurgy* (1967), technical articles confined to pyrometallurgy with about one-third concerned with copper; GEORGE R. ST. PIERRE (ed.), *Physical Chemistry of Process Metallurgy*, 2 vol. (1961), highly technical articles (only a few exclusively on copper); ALLISON BUTTS (ed.), *Copper: The Science and Technology of the Metal, Its Alloys and Compounds* (1954), authoritative chapters on all aspects of copper, each written by an expert, though now somewhat outdated; JOSEPH R. BOLDT, JR., *The Winning of Nickel: Its Geology, Mining, and Extractive Metallurgy* (1967), primarily concerned with nickel, but most nickel ores contain copper—hence, much of interest; ALEXANDER SUTULOV, *Copper Production in Russia* (1967); RALPH HULTGREN and P.D. DESAI, *Selected Thermodynamic Values and Phase Diagrams for Copper and Some of Its Binary Alloys* (1971), the most recent technical data of this type.

(A.Bu./A.W.S.)

Copyright Law

Copyright is the term used in English-speaking countries to characterize the legal recognition of rights to control or benefit from the communication of works of authorship. The word has a double meaning, stemming from its etymology. It denotes not only the right "to copy" but also the right to own and control "the copy," that is, the prototypic work of authorship itself. The word copyright is sometimes used interchangeably with vague terms such as literary property or intellectual property. None of these English-language labels, however, identify the primary beneficiary of the rights they cover. In contrast, equivalent terminology in other languages—such as *droit d'auteur* (French), *Urheberrecht* (German), and *derecho del autor* (Spanish)—does identify the beneficiary and can usually be translated simply as "the right of the author."

Copyrights, patents, and trademarks

Though similar in some ways, copyrights and patents are fundamentally different forms of legal protection. Copyrights protect works that owe their origin to the expressive efforts of a writer, composer, artist, or other creative individual. As long as a work is original in the sense that it was created independently, it can be copyrighted even if a closely similar work is already in existence. A copyright owner has rights only against those who use his work for their own purposes without his permission.

Conversely, both the criteria for obtaining a patent and the protection it offers are much greater than those applicable under copyright law. Patents protect the discoveries of inventors. A valid patent generally requires the invention of something that is both "novel," in the sense of never having existed before, and "unobvious," in the sense of being beyond the ordinary skill of an artificer or expert in the field. For as long as it lasts, the owner of a valid patent has a monopoly right to prevent anyone from using his invention, including other inventors who have made the same discovery independently.

Both copyrights and patents are frequently confused with trademarks, which are words or other symbols that have come to identify the source or sponsorship of merchandise or services. Legal protection of trademarks is based not upon creative authorship, as in the case of copyrights, or upon inventive discovery, as in the case of patents, but upon the investment of time, money, and skill in selecting a mark and inducing the public to identify it with a particular source of goods or services.

THEORIES OF COPYRIGHT

The literature of copyright law is crowded with efforts to fit copyright under a single heading, including property, monopoly, statutory reward, natural right, and personal right. To be consistent, writers espousing a particular theory were often forced into unrealistic positions. Those who argued, for example, that copyright is a form of property no different from goods or land were led to demand unlimited, perpetual rights in copyrighted works. Their opponents, relying on the theory that copyright is a monopoly and that all monopolies are evil, contended that authors should be given as little protection as possible. The value of this extended dialectic is that, taken as a whole, the philosophical arguments illuminate the special characteristics of copyright and demonstrate its unique legal nature.

Copyright as property. In its broadest sense, property is something of value created or acquired lawfully by a particular owner, who is entitled to transfer or deal with it in any way that does not conflict with the public interest. It is the duty of the law, to the extent consistent with the social norms on which it is based, to protect property against unauthorized appropriation or use.

Although copyright cannot be considered an absolute property right, it can be classed as a form of property. As property, it is artificial and limited in the sense that the law of a country creates it and imposes conditions and limitations upon it. The differences between copyright and ordinary forms of property stem from certain unique characteristics. In the first place, the purpose of creating a copyrighted work is to communicate its contents to the minds of others. Yet, once this disclosure takes place, the intangible work cannot be physically possessed or controlled. Indeed, it has become capable of instantaneous and unlimited reproduction and dissemination. The task of the law in protecting this kind of property thus becomes much more difficult.

There are, of course, other rights or forms of intangible property, apart from copyright, that the law protects as fully as it does land or goods. But in cases such as those involving securities or breaches of contract, the owner's rights are applicable only as against a particular person or limited group. In contrast, the copyright owner is owed a duty by the public at large and can enforce his rights against anyone and everyone who violates this duty.

Copyright as protection of originality and creativity

Finally, material subject to copyright is something more than a mere product of labour and investment. It emerges from the exercise of the creativeness and originality that go to make up a nation's culture. Despite the practical problems of enforcement and the burdens it places on the using public, copyright is thus regarded throughout the world as a form of property worthy of special protection in the ultimate public interest.

Copyright as monopoly. Both patents and copyrights are monopolies, giving their owners the power to control the market for a particular item. A copyright is a more limited monopoly than a patent, since it does not prevent competition from similar works that have been created independently. But, granting that a copyright is to some extent a form of monopoly, the question arises as to whether it shares the harmful economic and social effects usually attributed to monopolies.

The owner of copyright in a book or play has a monopoly in that particular work and, if it is a work of exceptional importance or popularity, his control over the extent of its dissemination and the prices charged for its use, if arbitrary or unreasonable, may carry the monopoly taint. But anyone else may write a book or play on the same subject, and there are usually many other books or plays competing for the public's attention.

Copyright in a single work is a small monopoly that can promote competition by forcing others to create their own works. This conclusion does not necessarily extend to situations in which many copyrights are pooled and controlled on a collective basis and in which the opportunities for monopoly abuse can become very real.

Copyright as a personal right. Copyrights, considered as unique forms of property and monopoly, are essentially pecuniary rights, but underlying them is a strong personal element. Authorship is, or should be, an individual creative process, something quite different from ordinary labour or investment. Copyright by its nature is linked inseparably with the individual author since it is he who should be stimulated to create and rewarded for his work.

Most countries recognize this personal factor directly in their copyright laws, notably through provisions giving an author a nontransferable "moral right" to prevent distortion of his work and to assure that he is identified as its author. Similarly, the duration of the copyright term throughout most of the world is based on the life of the individual author.

The purpose of copyright. Paralleling the dispute as to the legal nature of copyright has been a sharp conflict over the purpose a copyright law is intended to serve: should it protect the author in something that is rightly his, or should it benefit the public by stimulating creation, or should it do both?

Two differing approaches to copyright

The view that the author should be the fundamental beneficiary of copyright protection is the foundation of copyright legislation in a number of countries, particularly in western Europe. In general, this was the philosophy underlying the Berne Convention for the Protection of Literary and Artistic Works (1886, as subsequently revised: see below). Where this view prevails, the law starts from the premise that protection should be as long and as broad as possible and should provide only those exceptions and limitations essential in the public interest. The opposing view, which has its roots in English copyright law and is embodied in the copyright systems of the United States, of several east European nations, and of some developing countries, lies behind the Universal Copyright Convention adopted in Geneva in 1952. Under this philosophy, the law should give only as much protection as is necessary to induce authors to create and disseminate their works.

The middle ground between these opposing philosophies has been gaining acceptance since World War II and appears to have influenced recent legislation in a number of countries. Its basic premise is that the purpose of copyright is both to stimulate the creation and public dissemination of works and to give their authors a generous reward for their contributions to society. These aims are usually not in conflict, but where they are, they must be balanced to reach a fair result. This process of balancing is an infinitely delicate one, and among the many factors to be weighed are the educational needs and economic situation within a country, the importance of promoting national authorship, and the demands of the various new technological methods for disseminating the works of authors.

HISTORICAL DEVELOPMENT OF COPYRIGHT BEFORE 1886

The word plagiarism is reputed to have been coined by Martial who, in one of his epigrams, likened his poems to freed slaves and termed a rival poet, who had represented the poems as his own, an abductor (Latin *plagiarius*) of them. There are a number of other instances of literary theft throughout the classical and Middle ages, but, although the practice was sometimes the object of moral censure, it carried no legal consequences.

Copyright in the modern sense was born in the late 15th century, the offspring of Gutenberg's invention of printing and of the expansion throughout Europe of the learning and religious ferment accompanying the Renaissance and Reformation. At about the time William Caxton established a printing press in Westminster in 1476, the city of Venice inaugurated a system of granting "privileges," or monopoly rights, to print certain books.

The practice of sovereign grants of exclusive publishing rights spread quickly to other countries and became a common trade practice during the 16th and 17th centuries. The printer or publisher seeking the monopoly was willing to pay for the privilege and to submit the work for official approval. For the ruler making the grant, the system was thus a source of revenue and, more important, an opportunity for exercising political or religious censorship.

For more than 200 years this inchoate form of copyright was a matter involving tradesman and sovereign, and the individual author was rarely even an indirect beneficiary of the transaction. At the same time, the unauthorized reproduction of books, which had once been considered merely reprehensible, was gradually coming to be recognized as an illegal act.

In England the system of royal licenses to individual printers was organized into a definite procedure with the restoration of Roman Catholicism under Mary I. In 1555, reaction to the widespread persecution of Protestants under the reinstated heresy laws led the crown to seek methods for enforcing tighter censorship. In 1556, Queen Mary chartered the Stationers' Company, giving the members of this guild of London printers monopoly rights in the books they published. All books were required to be submitted for official approval and to be entered on the company's register; both unauthorized printing and failure to register were punished by decrees of the Court of Star Chamber.

This unvarnished restraint on freedom of the press persisted for more than a century, but the unpopularity of the Star Chamber led to a new legislative system recognizing perpetual rights in registered books and providing penalties for their violation. Parliament abolished the Star Chamber in 1641 but passed a new licensing act in 1643, which, with various lapses and renewals, survived until 1694. Indignation at the arbitrary power conferred by these licensing acts led to their final lapse in that year. The result was flourishing book piracy.

The Statute of Anne of 1710

Pressure by owners of literary property to restore their protection prompted Parliament to act in 1710 but in an unexpected way. The Statute of Anne of April 10, 1710, was hardly an exemplary piece of legislative craftsmanship, but in changing the conceptual nature of copyright, it became the most important single event in copyright history. Two of the principles on which it rests were revolutionary: recognition of the individual author as the fountainhead of protection and adoption of the principle of a limited term of protection for published works. This statute influenced the legal development of copyright protection the world over.

The statute of 1710 limited copyright protection in published works to a maximum of 28 years; but it left unanswered an important question. Assuming that authors had always enjoyed rights in their unpublished works, based on unwritten principles of the English common law, did these rights still continue after publication of a work? In other words, the question was whether the Statute of Anne added to common-law protection already available or cut off all common-law rights, thus protecting published works only for the term of years and under the conditions specified in the statute.

This controversy produced two landmark decisions in the history of copyright law. In 1769 the case of *Millar* v. *Taylor* held that authors have common-law rights in their creations, that these rights are not destroyed by publication of the works, and that they were not superseded by the Statute of Anne. Five years later, in 1774, the House of Lords in *Donaldson* v. *Becket* upheld the Millar decision on the first two points, acknowledging that before the Statute of Anne the common law had

provided perpetual copyright protection for all works, unpublished or published. It overturned the Millar case on the third point, however, holding that the statute destroyed and replaced all common-law rights in works after they had been published and that once the term of protection provided by the statute expires, the work enters the public domain and is free for anyone to use.

Twelve of 13 states of the newly independent United States of America passed copyright legislation between 1783 and 1789, but it soon became apparent that these statutes, each requiring separate enforcement, were ineffective. The result, in 1789, was the adoption of a provision in the U.S. Constitution empowering Congress "to promote the progress of science and useful arts by securing for limited times, to authors and inventors, the exclusive right to their respective writings and discoveries." Congress immediately exercised this power in 1790 by enacting a federal copyright statute closely patterned on the Statute of Anne.

The great English cases of *Millar* v. *Taylor* and *Donaldson* v. *Becket* found their American counterpart in the 1834 decision of the United States Supreme Court in *Wheaton* v. *Peters*. This important case dealt with the same basic question: What common-law rights continue to exist after a work has been published? Unlike *Donaldson* v. *Becket*, which held that common-law rights survive publication but are cut off by statute, the doctrine of the Wheaton case is that, regardless of the existence of a statute, common-law rights in the United States are cut off by publication of the work. Under this rule, a work enters the public domain unless there is a statute to protect it; and, if there is, protection is wholly dependent upon the provisions of that statute.

Wheaton v. Peters

Denmark had adopted a short copyright ordinance in 1741, and in 1793 France adopted legislation that was to serve as a model for copyright statutes in many civil law countries. Thereafter, throughout the 19th century, most independent nations enacted copyright laws; whatever their provenance and detailed differences, almost all of these statutes recognized the author as the fundamental beneficiary and offered protection for limited periods of time. The general trends in 19th-century copyright legislation, reflecting the spread of the Industrial Revolution and the striking improvements in transportation, communications, and printing techniques, were toward increasing the classes of works eligible for protection, broadening the exclusive rights in these works, and providing a longer term of copyright.

THE ORIGINS OF INTERNATIONAL COPYRIGHT PROTECTION

The Industrial Revolution and the expanding technology of communications brought with them not only a growth in the number and scope of domestic copyright statutes but also the need for reciprocal protection of works between countries. At first international copyright protection was quite unusual. As a rule, national copyright laws denied any protection to works of foreign authors, and what few exceptions there were derived from bilateral treaties negotiated on a strict give-and-take basis.

A turning point occurred in 1852, when France extended protection to all works of authorship regardless of their national origin. Although this generous act set no pattern, it accelerated the movement toward a more broadly based multilateral system of international copyright. The Association Littéraire et Artistique Internationale (ALAI) was formed in Paris and took the lead in seeking ways to establish an international union of countries pledged to the protection of authors' rights. It sponsored a series of meetings and, in 1883, prepared a draft treaty which was considered at intergovernmental conferences in 1884, 1885, and 1886. This draft provided the basis for the Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886.

Development of the Berne Union. The original Berne Convention of 1886, although a modest beginning, was of fundamental importance: it was not only the first multilateral copyright convention in history but it also established some enduring international copyright principles. Rather than adopting a merely reciprocal principle by which a country protects foreign works only to the extent that its own works are protected in return, the Berne Convention was based on the principle of national treatment or assimilation under which a country agrees to give foreign authors the same protection it accords its own authors. The convention set up an international copyright union of all member states, still known as the Berne Union. It required that, among union members, the right of translation (*i.e.*, the right to control the act of translating a copyrighted work into another language) be protected for a minimum of 10 years. The 1886 text established no other specific minimum requirements, but in the successive revisions of the Berne Convention the "minima" have expanded to the point that in the early 1970s the convention required members to offer a fairly high level of copyright protection to works of other member countries.

The original Berne Union consisted of 14 countries, all but one of which remained members. The convention has gone through a series of partial or complete revisions, in 1896, 1908, 1914, 1928, 1948, 1967, and 1971. In the early 1970s the union consisted of 59 countries, of which all but 16 were bound by the Brussels revision of 1948. Some 28 member countries were European, 13 were located in Asia or Asia Minor, 10 in Africa, five in Latin America, two in Australasia, and one (Canada) in North America. Neither the United States nor the Soviet Union has ever been a member.

Revisions of the Berne Convention

A revision of the Berne Convention in 1967 proved extremely controversial because of the broad concessions it offered to developing countries, and it failed to achieve the number of ratifications necessary to bring it into effect. This failure brought on a renewed effort to revise the Berne Convention in conjunction with revision of the other worldwide copyright treaty, the Universal Convention of 1952. Revisions of both conventions were signed at Paris on July 24, 1971, and will be discussed together below. The 1971 Paris text of the Berne Convention superseded the Brussels text of 1948 for some purposes, but during the mid-1970s the Brussels text remained the dominant text of the Berne Convention.

The Brussels text of the Berne Convention, 1948. Like the Rome text of 1928, which preceded it, the Brussels Act of the Berne Convention broadly defines the "literary and artistic" works a member country is required to protect. These include "every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression." In addition to more traditional examples of copyrightable subject matter, such as books, dramas, musical compositions, and works of art, the Brussels text specifies as copyrightable a wide variety of subject matter including choreography, pantomimes, cinematographic works, architecture, and photographs.

From the ten-year exclusive translation right of 1886, the Berne Convention has expanded to embrace a wide range of exclusive rights that all members are required to grant to works first published in other member countries. Under the Brussels Act of 1948 these minimum rights include, in general, rights of public performance, broadcasting, wire diffusion, adaptation, arrangement, translation, recording, and motion picture adaptation. Most, though not all, of these rights are subject to limitations.

Probably the most important innovation of the Brussels revision of 1948 was the introduction of the author's "moral right" as a mandatory requirement. Under Article 6, member countries are obliged to grant to an author of another member country, regardless of the ownership of copyright in his works, the "right, during his lifetime, to claim authorship of the work and to object to any distortion, mutilation or other alteration thereof, or any other action in relation to the said work, which would be prejudicial to his honour or reputation."

The Brussels revision of the Berne Convention took another major step toward stronger copyright protection when it established as mandatory a minimum term of copyright protection for "the life of the author and fifty years after his death." For "anonymous and pseudonymous works" the minimum term is set at 50 years from

publication, and member countries are left free to establish the duration of protection for motion pictures and photographs.

The Berne Convention specifically prohibits a country from making protection conditional upon the fulfillment of any formal requirements such as registration or the use of a copyright notice. An unpublished work is automatically protected if its author is a national of a Berne Union country, and a published work receives protection if it is first published in a Union country, either alone or simultaneously with publication elsewhere.

INTER-AMERICAN COPYRIGHT CONVENTIONS

During the formative stage of the Berne Convention in the 1880s, efforts were being made to develop multilateral copyright arrangements in the Americas. These efforts produced a series of Pan-American copyright conventions, the most recent of which were revisions adopted at Havana in 1928 and at Washington in 1946. The most important of the Pan-American conventions, and the only one of which the United States is a member, is the Buenos Aires Convention of 1910. In effect, it provides that a work is to be protected in a member country if it has been copyrighted in another member country and bears a form of copyright notice.

The trend since World War II has been away from regional copyright conventions and toward global arrangements that can accommodate a wide diversity of legal systems, economic conditions, and stages of cultural development. The need for a bridge between the Pan-American conventions and the Berne Convention was a motivating force behind development of the Universal Copyright Convention, and the latter has since superseded the Pan-American conventions in many cases.

THE UNIVERSAL COPYRIGHT CONVENTION

Complexity of U.S. copyright procedures

More than anything else, the copyright situation in the United States, with its notice, registration, deposit, affidavit, and manufacturing requirements, was responsible for the development of the second worldwide multilateral copyright convention, the Universal Copyright Convention. The UCC, as it is usually called, was adopted at Geneva on September 6, 1952, and took effect on September 16, 1955. In the early 1970s, 60 countries were parties to the UCC, including 40 that were also members of the Berne Union. Among UCC members, 26 were located in Europe, 14 in Latin America, eight in Africa, eight in Asia and Asia Minor, two in Australasia, and two in North America.

The United States offered no protection to foreign authors until 1891, did not participate actively in the adoption or revisions of the Berne Convention, and established a copyright system sharply diverging from that underlying the Berne Union. A series of legislative programs during the 1920s and 1930s, aimed at U.S. adherence to the Berne Convention and general revision of the domestic law, met with complete failure. After World War II there was general recognition in the United States and most other countries that U.S. participation in a broad multilateral copyright arrangement was important for the future of international copyright. It was also realized that it would be futile to continue thinking of the Berne Convention as that arrangement, since U.S. copyright practice differed so much from that of the Berne Convention countries. The approach adopted was a new "common denominator" convention, which was intended to establish a minimum level of international copyright relations throughout the world without weakening or supplanting the Berne Convention.

At least in comparison with the 1948 Brussels revision of the Berne Convention, the Universal Copyright Convention represents a rather low-level compromise arrangement, resembling in many ways the original 1886 Berne act. Like the Berne Convention, the UCC is based on the principle of national treatment: works of other UCC countries must be protected to the same extent as domestic works. The field of application of the UCC is somewhat broader than that of the Berne Convention, since, in addition to unpublished works by citizens of UCC countries and works first published in member countries, the treaty obligations apply to published works by authors who are citizens of member countries, regardless of where first publication takes place. But the Universal Copyright Convention's minimum requirements as to exclusive rights are extremely modest. Aside from a vague obligation to "provide for the adequate and effective protection of the rights of authors and other copyright proprietors," the UCC merely establishes a requirement that contracting states give exclusive translation rights for at least seven years and thereafter establish a compulsory licensing system for translations.

The great compromise of the UCC, embodied in Article III, provides that the formal requirements of a contracting state's copyright law, such as notice, registration, and domestic manufacture, are satisfied for foreign UCC works

> if from the time of the first publication all the copies of the work . . . bear the symbol © accompanied by the name of the copyright proprietor and the year of first publication placed in such manner and location as to give reasonable notice of claim of copyright.

This provided a sturdy and much-travelled bridge between the U.S. system, with its complicated formal requirements, and the Berne system, which specifically forbids any such formalities.

The other main compromises in the Universal Copyright Convention were really concessions to the U.S. system. The requirements as to the minimum duration of protection are very permissive: subject to a range of detailed qualifications and exceptions, the minimum term is to be either 25 years from the death of the author or from the date of first publication. Another main stumbling block impeding U.S. adherence to the Berne Convention had been its retroactive effect, requiring protection for works that had previously been considered in the public domain. In contrast, the UCC specifically frees a contracting state from any obligation to protect works that are permanently in its public domain on the date the convention becomes effective in that state.

Later developments. A crisis in international copyright relations was precipitated by the adoption at Stockholm, on July 14, 1967, of a revision of the Berne Convention. The crisis did not result from the rather minor and technical revisions in the body of the convention itself but from the "Stockholm Protocol Regarding Developing Countries." This instrument, though self-contained, is an integral part of the Stockholm text and consists of extremely sweeping concessions to developing countries, allowing them to make unauthorized use of works of other union countries under broad and loosely defined compulsory licensing systems. As it turned out, the concessions were so broad that no developed country accepted the new text, and the result was an impasse that for a time threatened the future of the Berne Union. It seems probable that the Stockholm revision, as such, will never come into effect.

The Stockholm protocol of 1967

In an effort to break this impasse, and following a number of preparatory meetings, international conferences for the revision of both the Berne Convention and the Universal Copyright Convention were held at Paris in July 1971. The coordinated program of these conferences resulted in a new text of the Berne Convention that retains the revisions in the body of the convention already adopted at the Stockholm Conference but drops the protocol in favour of an appendix making more realistic concessions to developing countries with respect to reproduction and translation of materials having educational value. The new text successfully meets the needs of developing countries without damaging the legitimate copyright interests of developed countries. Equivalent changes were made in the Universal Copyright Convention, which was also amended to require member countries to accord at least some minimum level of protection to the rights of reproduction, public performance, and broadcasting.

Aside from the concessions to developing countries, the changes in the Berne Convention made at Stockholm appear destined to come into effect in most Berne Union

countries. Of these changes, the most basic involve the criteria of eligibility for protection of published works under the convention: the Stockholm text retains the traditional criterion that the work must be protected if it was first published in a Union country but makes it secondary to a new criterion that the work must be protected if the author's nationality is that of a Union country. The definition of "published works" has also been revised to cover "works published with the consent of their authors, whatever may be the means of manufacture of the copies, provided that the availability of such copies has been such as to satisfy the reasonable requirements of the public, having regard to the nature of the work."

In the area of exclusive rights, the Stockholm Conference finally added the right of reproduction (*i.e.*, the right to control the duplication of the work in tangible form) to the rights a Union country must protect. Certain "fair use" qualifications (that is, for example, exceptions to copyright restrictions as when quoting from a book in a review of it) accompanied the granting of reproduction rights, and exemptions dealing with news reporting and instructional uses were broadened. The Stockholm text also included a package of compromise provisions aimed at facilitating the international exchange of motion pictures and television programs. The obligations of a Union country with respect to the author's "moral rights" were strengthened, and minimum terms were adopted for the duration of copyright protection in cinematographic works (50 years) and in works of photography and applied art (25 years).

Basic congruence of most national copyright systems

Although they differ greatly in any number of detailed points and often approach copyright from conflicting theoretical premises, the copyright statutes of most major countries of the world follow a consistent pattern. This general pattern is based directly or indirectly upon the framework and provisions of the various revisions of the Berne Convention, the fundamental paradigm for copyright statutes as different as those of France, the United Kingdom, Japan, the Soviet Union, and Brazil.

As a rule, in most countries except the United States, copyright in both published and unpublished works is secured automatically by the author. As long as his work is original, in the sense that he created it independently, the author need do nothing to secure and maintain his copyright, which lasts throughout his lifetime and for a period of years after his death. The postmortem period varies from 15 years in the Soviet Union to 70 years in the Federal Republic of Germany, with the vast majority following the Berne formula of 50 years.

THE SPECIAL SITUATION IN THE UNITED STATES

Copyright in the United States is essentially and radically different from copyright in nearly every other country. Under the law of the United States, there are actually two systems of copyright protection: a "common law" copyright based on the laws of each of the 50 states, automatically protecting the authors of unpublished works, and a statutory copyright protecting works published under the federal copyright act of 1909. A "common law" copyright comes into existence as soon as a work is created and carries with it much the same protection as a statutory copyright but lasts only as long as the work remains unpublished. If a work is published without observing the strict formal requirements of the act of 1909, both common law and statutory copyright protection are lost permanently and the work is thrown into the public domain.

United States copyright in published works is secured by the act of publication with the notice of copyright prescribed in the statute. For most types of works this notice must consist of the word Copyright, the abbreviation Copr., or the symbol ©, accompanied by the year of first publication and the name of the copyright owner. The date may be omitted for certain kinds of graphic, pictorial, or three-dimensional works. For a book or similar printed publication the copyright notice must appear on the "title page or page immediately following." The location specified for musical compositions is the "title page or first page of music," and that for periodicals is the "title page, the first page of text, or under the title heading." Under a recent amendment of the statute, certain sound recordings can be protected if they are published with a special copyright notice.

Although publication of a work with the copyright notice is the crucial act upon which copyright protection in the United States depends, the concept of "publication" has become increasingly obscure as the new media for disseminating works have increased in number and popularity. The sale of a handful of copies can constitute publication for copyright purposes, but the same is not true of the reading or performance of a work before an audience of millions. The copyright effects of disseminating, under various conditions, film prints, phonograph records, audio and video recordings in different forms, computer tapes, and similar artifacts could hardly be more unclear under U.S. law.

As soon as copyright in a work has been secured by publication with notice, a U.S. copyright owner is under an obligation to send to the copyright office in Washington, D.C., an application for registration, two copies of the best edition of the work, and a registration fee ($6). Another requirement is the notorious "manufacturing clause" of the U.S. copyright statute, which requires that the typesetting, printing, and binding of books or periodicals in the English language must be done in the United States. Since the widespread adoption of the Universal Copyright Convention, which exempts UCC works of foreign origin from the manufacturing requirements, the clause now has its main impact upon books and periodical material by American authors. If the copies of these works are manufactured and first published abroad, protection in the United States can be secured only by registering a claim to an "ad interim" copyright within six months of first publication, and by publishing an edition manufactured in the United States within five years.

The U.S. "manufacturing clause"

United States copyright protection for published works begins on the date of first publication, while works registered in the U.S. Copyright Office in unpublished form are protected from the date of registration. In both cases copyright lasts for 28 years and may be renewed for a second term of 28 years but only if an acceptable application and renewal fee are received in the Copyright Office within the last year of the original copyright term. If the author is no longer living, the author's widow, widower, or children, his executors, or his next of kin become entitled to renew in that order. In certain exceptional cases, including periodicals and works made for hire, the owner of the copyright is entitled to claim the renewal.

The maximum term of copyright protection in the United States is set in the Act of 1909 at 56 years. Special legislation has, however, extended some expiring copyrights by as much as ten years. Upon the expiration of a copyright all exclusive rights end, and the work enters the U.S. public domain.

BIBLIOGRAPHY

United States: R.R. BOWKER, *Copyright: Its History and Its Law* (1912); ARTHUR W. WEIL, *American Copyright Law* (1917); HORACE G. BALL, *The Law of Copyright and Literary Property* (1944); STANLEY ROTHENBERG, *Legal Protection of Literature, Art and Music* (1960); HERBERT A. HOWELL, *Copyright Law*, rev. and ed. by ALAN LATMAN, 4th ed. (1962); MELVILLE B. NIMMER, *Nimmer on Copyright* (1963); BRUCE W. BUGBEE, *The Genesis of American Patent and Copyright Law* (1967).

Canada: HAROLD G. FOX, *The Canadian Law of Copyright and Industrial Designs*, 2nd ed. (1967).

Great Britain: WALTER ARTHUR COPINGER and F.E. SKONE JAMES, *Copyright*, 10th ed. (1965).

International copyright: WILLIAM BRIGGS, *The Law of International Copyright* (1906); STEPHEN P. LADAS, *The International Protection of Literary and Artistic Property*, 2 vol. (1938); ARPAD L. BOGSCH, *The Law of Copyright Under the Universal Convention*, 3rd rev. ed. (1968).

History: HARRY RANSOM, *The First Copyright Statute* (1956); LYMAN RAY PATTERSON, *Copyright in Historical Perspective* (1968).

Specific aspects: AUGUSTINE BIRRELL, *Copyright in Books* (1899, reprinted 1970); HARRY P. WARNER, *Radio and Tele-*

vision Rights (1953); SAMUEL SPRING, *Risks and Rights in Publishing, Television, Radio, Motion Pictures, Advertising, and the Theatre*, 2nd ed. rev. (1956); SIDNEY SHEMEL and M. WILLIAM KRASILOVSKY, *This Business of Music* (1964) and *More About This Business of Music* (1967); RICHARD WINCOR, *Literary Property* (1967); PHILIP WITTENBERG, *The Protection of Literary Property* (1968).

(B.A.R.)

Coraciiformes

The birds of the order Coraciiformes include the kingfishers, todies, motmots, bee-eaters, rollers, hoopoes, and hornbills and are collectively referred to as roller-like birds. Among the members of the order that have attracted man's special attention are certain kingfishers that plunge headfirst into water for fish and are associated with classical mythology. According to the ancient Greeks, Ceyx and his wife Alcyone were shipwrecked at Delphi and changed into kingfishers. The Chinese used the shining blue feathers of some types of kingfishers to decorate picture screens. Bee-eaters (Meropidae) have been accused of preying on commercially valuable honeybees, and the North American belted kingfisher (*Megaceryle alcyon*) is sometimes considered a pest at fish hatcheries, because it preys on young game fish. The kookaburra (*Dacelo novaeguineae*) has a loud, laughing or braying voice that is commonly associated with the Australian outback, or back-country. To the biologist, the sealing of the female of certain species of hornbills in her nest during incubation and brooding is one of the most intriguing behavioral modifications among birds.

General features. The roller-like birds are active by day (diurnal) and range in length from the size of a small sparrow (about ten centimetres, or four inches) to over 100 centimetres (40 inches). They have compact bodies, short to moderately long necks, large heads, rather long bills, small feet, and ample wings. The tail varies from short to very long and may be forked, square, or graduated; the outer or central tail feathers are, in some species, pointed or spatulate at the tip. All of these birds regularly perch in trees, where some feed; others fly in search of food, and a few walk or hop on the ground. The group's food ranges from invertebrates (including insects) and small vertebrates to berries and fruit.

Drawing by R. Keane

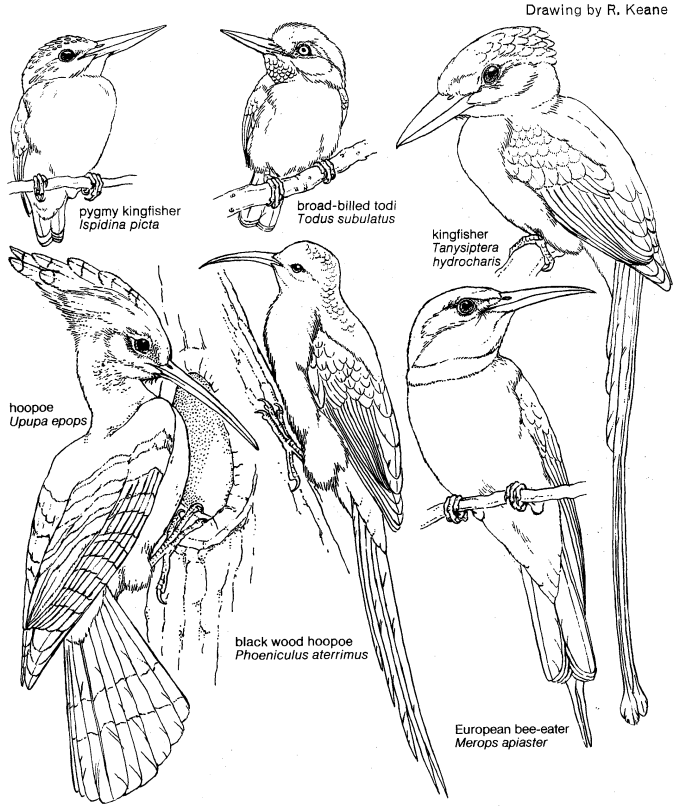


Figure 1: Body plans of smaller Coraciiformes.

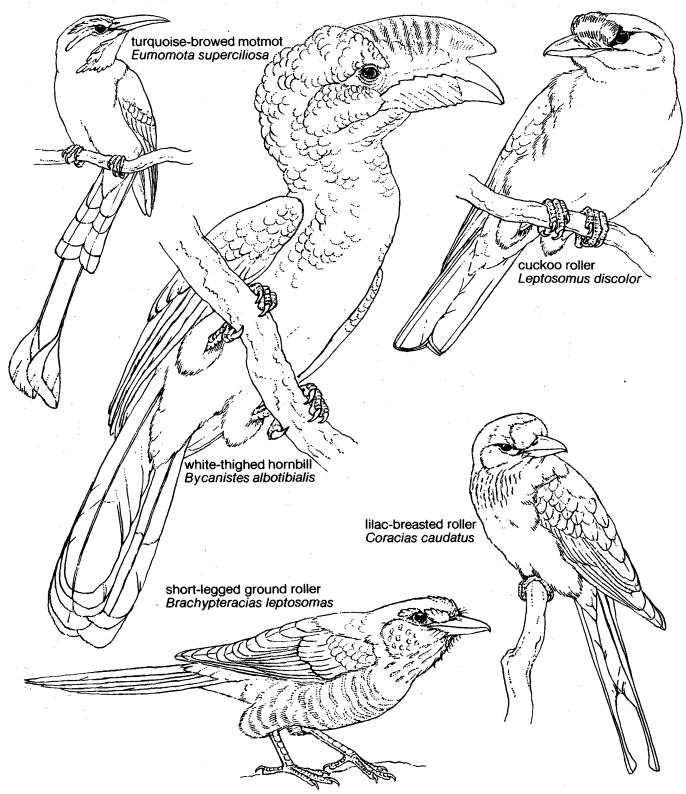


Figure 2: Body plans of larger Coraciiformes.
Drawing by R. Keane

Distribution. Collectively, the ten families of the order are almost worldwide in temperate and tropical areas, with the greatest number and diversity in the warmer parts of the African, southern Asian, and Papuan areas. Many species are common and conspicuous, and a few tolerate human settlement, although their direct importance to man is minimal.

Families of coraciiform birds

Only the kingfishers (family Alcedinidae) are found in both Eastern and Western Hemispheres. The motmots (Momotidae) and todies (Todidae) are restricted to the New World tropics, the bee-eaters (Meropidae), rollers (Coraciidae), and the hoopoes (Upupidae) to temperate and tropical regions of the Old World, and the hornbills (Bucerotidae) to the tropics of Africa and Asia. The wood hoopoes (Phoeniculidae) are found only in tropical and subtropical Africa; the cuckoo roller (Leptosomatidae) and the ground rollers (Brachypteriacidae) are found only on Madagascar.

NATURAL HISTORY

Life history. Most species live permanently in one region, but temperate-zone species move nearer the tropics for the winter. In the Old World tropics that have a dry–wet seasonal change, local movements of bee-eaters sometimes take advantage of the related crops of insects. In East Africa, where there are two dry and two wet seasons, two periods of breeding may occur a year; however, the picture is not clear.

As a group, these birds are well endowed with voices, but their vocalizations are usually referred to as calls rather than songs. Some are harsh; others are soft or are whistled or hissed. Some are given as single notes, others as series in a trill, a rattle, or a hooting, and still others as a cacophony or medley of notes. Some calls are given from perches; others are given on the wing. Certain utterances may be related to courtship and mating or to territory; others seem to be simply a part of the bird's general daily activity. Pairs are formed to the accompaniment of simple posturing displays and series of calls, ranging from harsh to soft. Certain rollers use tumbling display flights. Both sexes usually share in the nest duties.

Many roller-like birds are solitary in feeding and nest-

ing, but some bee-eaters are gregarious and also nest in colonies. Hornbills and wood hoopoes move about in small parties most of the year, but they nest solitarily. The nest site is always in a cavity, which may be a hole or crevice in a tree, a bank, or a wall. The cavity may be among rocks or be a tunnel dug by the birds in the ground, or it can be an abandoned termite nest.

Walling in of the female hornbill

Normally, little or no nest material is added, except by the hoopoe (*Upupa epops*). In some (perhaps most) hornbills, however, the female enters the nest cavity before laying starts; the male brings mud and debris, which the female takes and plasters around the entrance until only a slit is left. The male passes food to the female through the slit until after the young are more or less grown. At that time she breaks out of the nest, and then both parents feed the young.

The members of the order lay two to nine eggs, with the tropical species laying the smaller clutches. Among the hornbills the larger species lay fewer eggs than do some smaller ones. The eggs are usually white but, in the hoopoe, may be olive, brownish, bluish, or greenish, and sometimes spotted. Incubation is performed by both male and female in kingfishers, todies, motmots, and bee-eaters and by the female alone in the hoopoe and hornbills when she is fed by the male. Incubation periods of 18 to 22 days have been recorded for some of the smaller members of the order.

The young are nidicolous (dependent upon the parents) and are naked, except for the hoopoe and some kingfishers that have varying amounts of down at birth. The recorded nestling period (25 to 28 days) of certain of the smaller species is only slightly longer than the incubation period, but large species have longer nestling periods, and the great hornbill (*Buceros bicornis*) has a total period (incubation to fledging) of three to four months. Except for the hoopoe and the walled-in hornbills, the young are fed by both parents. The male of these species brings food to the female, who passes it on to the young. The rate of feeding the young is apparently variable. In large species of hornbills, the male is recorded as making one trip per hour to the nest, during which he brings a gullet full of fruits and regurgitates them one at a time and passes them to the female. A much smaller species brings one item of animal food and holds it in his bill each trip and makes about six trips an hour.

Coraciiforms of several groups practice no nest sanitation, and regurgitated insects and other remains accumulate and are used to make a platform on which the young rest in later nest life. Hornbill nests are kept more or less clean, and, in one small species in which the female breaks out of the nest when the young are half grown, the young perform nest sanitation and also help to seal up the nest again after the female emerges.

When the young leave the nest, they are able to fly and usually are similar to, although sometimes duller or paler than, the adult in plumage. There is believed to be one molt a year, which occurs after the nesting season. In an unusual modification, the female of the African hornbills is walled in during incubation and throughout the nest life of the young. She molts during this period, losing and renewing all of her flight feathers.

Locomotion and feeding. Coraciiform birds tend to perch in trees and shrubs when at rest. Some favour exposed perches on which they are conspicuous, others seek the protection of foliage or the shade of the forest. Lacking cryptic coloration—many are nearly uniform in colour or boldly patterned—they do not rely on concealment for protection. Their flight varies from weak and laboured to strong, well sustained, and direct. The flight of some, such as the rollers, is swift and graceful. Some species, such as the kingfishers, use little bipedal locomotion; others (*e.g.*, some hornbills) hop, walk, or scramble in the treetops, creep along branches (wood hoopoes), or walk or hop on the ground (hoopoe, other hornbills).

The food of the roller-like birds includes a wide variety of organisms. Among the animals taken are worms, snails, crustaceans, insects, fish, amphibians, reptiles, small birds, and mammals. Vegetable food consists chiefly of the fruits of trees, usually gathered in the trees but sometimes picked up on the ground.

Types of foraging behaviour

Some birds of this order seem to choose animal food more for its size than its type, and a single method of feeding tends to predominate in each family. Some families (Upupidae and Leptosomatidae) contain only one species; others are large and predictably diverse—*e.g.*, the hornbills (Bucerotidae), with 45 species, and the kingfishers (Alcedinidae), with about 90 species. Each family or group of families tends to have a characteristic pattern of feeding behaviour, and the foraging patterns fall into four categories, or feeding niches: (1) watchful waiting on a perch, (2) aerial—*i.e.*, spending much time on the wing, (3) searching on foot among branches of trees, and (4) walking on the ground.

The watchful waiters, the kingfishers, motmots, and todies, tend to sit quietly for long periods. When they see their prey on a leaf, a branch, the ground, or even in water, they fly out with swift, direct flight, seize the prey with the bill, and return to the perch. The todies catch more flying insects than do members of the other two families. A few kingfishers plunge headfirst into water from perches or from hovering flight, but these number only a few of the species-rich family Alcedinidae. The shovel-billed kingfisher (*Clytoceyx rex*) of New Guinea is partly terrestrial and is known to feed on beetles and earthworms; the latter are apparently dug from the soil of the forest floor with the bird's short, heavy bill. The ruddy kingfisher (*Halcyon coromanda*), widespread in Southeast Asia, eats many large land snails. It seizes a snail with its bill and beats it against a rock until the shell is broken and the meat can be extracted.

The term temperament, although tinged with human associations, seems applicable to certain traits that are common to many species within a family, as contrasted with members of another group. Kingfishers, motmots, and todies are stolid, phlegmatic birds that sit quietly for varying periods of time between sallies for food. Kingfishers often bob their heads and the forepart of their bodies when nervous or mildly alarmed; when startled into flight, some give sharp calls. Motmots have the habit of moving their tails from side to side.

Bee-eaters and rollers are aerial feeders but spend much time perched quietly, leaving their perches to make sallies for passing insects. Bee-eaters, especially, often spend long periods on the wing, gliding in circles while looking for insects, especially bees and wasps. Some forest bee-eaters perch in foliage and near flowers, securing their prey without flying. Rollers spend more time perched, but they too are graceful in flight and capture much of their food by hawking or by darting down to the ground. Members of both groups are often seen aloft, apparently not feeding but flying for diversion. The cuckoo roller (*Leptosomus discolor*) also flies above the forest canopy, but it is looking for large insects and small lizards in the outermost foliage. It may either seize them while on the wing or alight to capture them.

Feeding while clambering among the branches of trees is carried on by many of the larger hornbills and by the small wood hoopoes, but in quite different ways. The hornbills fly over or through the forest with strong, often noisy flight and, on alighting, scramble or hop among the branches reaching out for fruit, small animals, or both. The wood hoopoes have weak flight, and they do not fly much; they fly chiefly from one tree or clump of trees to the next, climbing about the trunks and branches of trees and lianas in acrobatic poses as they seek insects in crevices and on the bark surface.

Walking on the ground is the usual mode of feeding for the common hoopoe (*Upupa*), the ground rollers, and for a few hornbills. The hoopoe walks with quick steps, bobbing its head in time with the steps and pausing to probe with its long bill in the ground and in crevices, in search of large arthropods and small vertebrates. Its flight is strong and direct. When perched, it may quietly flash its long crest open and shut. The ground rollers, most of which are birds of the deep forest, also feed on the ground on food similar to that of the hoopoe. When disturbed, they fly or jump to low perches.

Social feeding by ground hornbills

The ground hornbills (*Burcorvus* species) exhibit a definite social organization when foraging. Three or four members of a group searching for insects and other small animals on the ground may keep near each other, with the result that prey frightened into activity by one bird may be caught by one of the others. Several other species of hornbills occasionally forage solitarily on the ground.

Habitat selection and ecological diversity. Coraciiform birds, diverse in their structure and behaviour, occupy a variety of habitats. Each species is restricted in distribution by requirements of feeding and nesting areas. A species may be considered to occupy a feeding and, during the breeding season, a nesting habitat, and these may or may not be contiguous. Some species, such as certain African kingfishers, which nest in cavities excavated in termite mounds and feed on termites and other insects near the nest site, may be said to nest within their feeding habitat. Aerial feeders, such as bee-eaters, which nest in burrows, may be considered to occupy two habitats, one for feeding and one for nesting.

The factors that influence habitat suitability are evident in a few cases. One bee-eater requires a cut bank in a grassland area for its burrow, but must be near a forest because it feeds over the forest and the forest edge. The importance of suitable perches is illustrated by observations made on an insect-rich region of East African grassland, from which kingfishers were absent until a road was built. With the roads came telegraph wires, which provided the perches.

Relative diversity of Old and New World coraciiform families

The widest range of habitats occupied by members of the order Coraciiformes is found in Africa (including Madagascar), where eight of the ten families are represented. These exhibit all four basic feeding modes, but on Madagascar some niches are occupied by families different from those on the African mainland. Ground feeding in the forest, for example, is limited to the ground rollers in Madagascar; in Africa this niche is occupied by the hoopoe (except in dense forest) and by birds of other orders. The four coraciiform families found in temperate Eurasia occupy only three niches: ground feeding (hoopoe), darting from a perch (kingfishers, some bee-eaters), and aerial hawking (other bee-eaters, rollers). In Australia only the last two niches are occupied by coraciiforms, and these are the same families that hold these niches in Eurasia. In the New World, where three families are found, only one feeding mode is used, that of darting out from a perch; but this mode is geographically and ecologically subdivided between the diminutive todies, which are limited to the Greater Antilles where neither the kingfishers nor the motmots are represented, and the kingfishers (with the exception of the pygmy kingfisher, *Chloroceryle aenea*), which are specialized fish eaters.

Social behaviour. A number of coraciiform birds are markedly social, feeding in small parties and nesting in colonies. Some wood hoopoes forage in conspicuous, noisy bands of five to ten individuals. The acrobatic, climbing activity of a band is sometimes interrupted when the birds of a whole party bow and sway their bodies, pump their tails up and down, and join in a chorus of chattering calls. Moving from tree to tree, one bird follows another in weak undulating flight. Forest species of wood hoopoes are less social, and lone individuals sometimes call while perched high in a tree. Many hornbills may be seen flying through or over the forest, the beats of their broad wings giving a characteristic loud whooshing noise. They are usually found in small parties and actively move about in the branches, sometimes giving conversational notes. A study of a young captive hand-reared hornbill of a small species of *Lophocerus* (*Tockus*) has provided surprising data. This bird seemed to have a remarkably active, alert, intelligent personality recalling that of captive crows. It greeted its foster parents by raising its wings, pointing its beak upward, raising its head feathers, and chattering. It was jealous of attention given other animals, kept close to its foster parents when out of doors, and alighted on their shoulders. When accompanying them on walks, it flew from tree to tree. It was busy and mischievous, attracted by anything bright, and fond of picking at knots or holes. This young hornbill had a passion for pulling up seedlings and sometimes amused itself by darting in and tweaking the tail of a larger, more lethargic, young *Bycanistes* hornbill.

Relationships with other species. Certain types of social behaviour of the Coraciiformes involve other birds or unrelated animals. Although some of these interactions are occasional and opportunistic, others are regular parts of everyday life and may be called symbiotic—*i.e.*, one that brings mutual benefit to the different species involved.

Utilization of grass fires

The regular swarming of many bird species about grass fires to capture animals driven out of hiding by the flames is a phenomenon often related to human activity, for such events are often caused accidentally or deliberately by man. Among the birds that gather are both rollers and bee-eaters; they swoop down near the flames and into the smoke to seize fleeing insects. After the fire has passed, certain hornbills find good foraging on foot over the newly exposed ground.

More notable are a number of interspecific nesting relationships. Some bee-eaters make their colonial burrows in the same banks in which certain smaller swallows dig their burrows; there seems to be no conflict between the larger bee-eaters and the smaller swallows, despite the similarity in nesting and feeding habits. In southern Africa, the little bee-eater (*Melittophagus pusillus*) sometimes makes its nest burrow in the wall of the very much larger burrow of the aardvark (*Orycteropus afer*), and there is no further relationship between the bird and the mammal.

Sporadic incidents occur between species when one or both are foraging; a kingfisher may pilfer a food item from a dipper (*Cinclus*), and a savanna kingfisher will occasionally fly down to seize a grasshopper flushed by a man. Many associations are more frequent. Some bee-eaters in Africa often accompany large bustards, other large walking birds, and zebras and other game animals to feed on the insects roused from the grass by the animals. The bee-eater even uses the bustard's back as a perch. The bee-eater may also accompany an automobile driven through these grasslands to secure insects. There is also a regular association between hornbills and bands of monkeys in the treetops of African forests, with the birds seeking the insects stirred into activity by the fruit-eating monkeys.

In many parts of the Old World tropics, where large arboreal termite mounds are common and conspicuous, certain kingfishers usually excavate their burrows in them; in fact, some species are believed to nest only in them. The presence or absence of the termites might be expected to have an important effect on the populations of such kingfishers. The hoopoe commonly nests near buildings, especially in South Africa, and it is possible that the availability of such sites may affect the local abundance of the species. The presence of woodpecker holes used by hoopoes and wood hoopoes may also affect the size of their breeding populations.

Some African species of kingfishers, bee-eaters, hoopoes, and wood hoopoes are victimized by obligate social parasites, the honey guides (Indicatoridae, related to the woodpeckers). The honey guide lays its eggs in the host's nest and, with its bill or claws, often punctures the shell of the foster parents' eggs so they do not hatch. If the foster parents' eggs do hatch, the nestling honey guide usually disposes of the host's young by throwing them from the nest or by biting, crowding, or starving them to death. The honey guide's young are thus raised at the expense of the young of the host species. Apparently kingfishers, bee-eaters, hoopoes, and wood hoopoes are of great importance in the ecology of honey guides, and the frequency with which the roller-like birds are victimized by honey guides may be a serious factor in their population status.

A remarkable insect fauna has been found in the nest of an African hornbill. Though some nest sanitation is practiced by the birds, it is not complete. In one nest.

more than 400 individual insects, mostly larvae, were found (about half were moth larvae); they represented eight species and were feeding on the droppings and debris in the nest cavity, which was remarkably clean and had little odour. The hornbill provides microhabitats for the insects (albeit scattered and seasonal), and the scavenging of the insects may be of advantage to the hornbill.

FORM AND FUNCTION

Size and plumage. The coraciiform birds are a rather heterogeneous order, united mainly by features of their internal anatomy. Some characteristics of the beak and feet serve to separate them from other orders, such as perching birds (Passeriformes) and the woodpeckers and their allies (Piciformes), which appear to be their closest relatives.

No single coraciiform family encompasses the entire size range of the order. The todies are the smallest, with lengths of nine to about 11.5 centimetres (3.5 to 4.5 inches), and the hornbills, which range in length from about 40 to 160 centimetres (16 to 63 inches), are the largest. The kingfishers are from ten to nearly 46 centimetres long (four to 18 inches), the longest being those with extended tail feathers. Motmots and bee-eaters are in the same general size range as the kingfishers, but the smallest of them are larger than the tiniest kingfishers (*Ceyx*, *Ispidina*), and the largest motmots, although about 50 centimetres (20 inches) long, have not nearly the body bulk of the chunkier, but slightly smaller, kookaburras (about 45 centimetres, or 18 inches). The smaller families have, predictably, less size variation.

Ornamental coloration

The plumage of the roller-like birds is firm and often highly colourful. The bee-eaters are collectively and individually among the most brilliantly coloured of all birds; one individual may be marked with green, yellow, red, blue, and black. Many kingfishers are also brightly coloured, with a tendency toward metallic blues and blue greens. The beak is often bright red or orange. Most hornbills, with ornamentation frequently found on the beak, are strikingly patterned in black, white, and shades of gray and are sometimes accented with rufous or yellow; many have areas of bare skin, blue, red, yellow, or black in colour, around the face.

Morphological specializations. A few of the external features, such as modifications of the wings and feet for locomotion and of the beak for food handling, are obviously related to behaviour and habitat. Even in these important aspects of the body plan, the common heritage is evident in such features as the small feet and fusion of the front toes.

The most obvious adaptation to behaviour is the shape of the wing. The size and shape of the wing correlates well with the type of flight. Aerial feeders have the longest, most pointed wings; the most extreme forms are found in the bee-eaters, but they are also well developed in the rollers and the cuckoo roller. The birds that watch for their prey and fly out after it (*e.g.*, kingfishers) have moderate, rounded wings, while the ground feeders (ground rollers, hoopoes) have broad wings, as do those that feed on foot in the trees (hornbills, wood hoopoes).

The tail is highly diversified in length and shape. Forked tails occur only in the best fliers (bee-eaters and rollers), though some of these birds have square tails or elongated central tail feathers instead. Elongated central tail feathers also occur in the hornbills that practice direct flight, as well as in ground rollers that fly little. The small kingfishers have the shortest tails. The exact shape of the tail does not correlate well with locomotion, nor does the presence of spatulate tips on elongated tail feathers, which occur in motmots, rollers, and kingfishers and are perhaps of social importance.

The characteristic short tarsus (the lower leg) of the order and the fusion of the three front toes (except in the cuckoo roller) seem a heritage that has been modified but little with modification of behaviour. The foot is used only for perching in most family groups. In cases in which it is used extensively for terrestrial locomotion (ground rollers, hoopoe), the tarsus is lengthened somewhat in the former but not in the latter, and is also somewhat lengthened in the aberrant hornbill, known as the ground hornbill. The hornbills that feed in the tree branches have a broad pad on the toes, evidently an adaptation for perching. The wood hoopoes have long slender toes with long, sharp, curved claws, an obvious adaptation for the bark-climbing habits of these birds.

Drawing by R. Keane

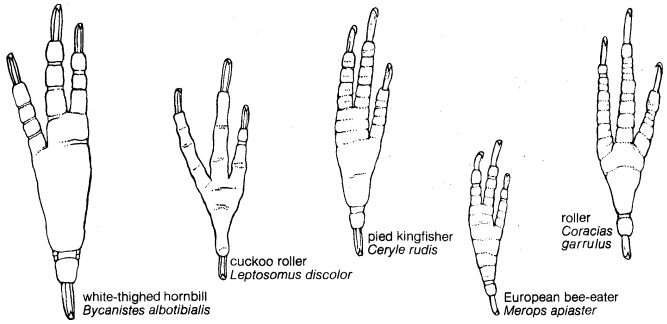


Figure 3: Feet of some Coraciiformes, plantar (bottom) view.

The most unusual foot in the order belongs to the cuckoo roller, whose outer toe is capable of being reversed. This makes a better perching foot for the bird that flies over the forest trees, scanning the branches for its prey, and alighting suddenly to seize a caterpillar, chameleon, or grasshopper.

Diversity in the bill

The bills show remarkable diversity in bulk and shape. Basically, the long stout bill so common in this order seems to be an adaptation for seizing and subduing active animal prey that is large in proportion to the size of the bird. This is true for rollers, ground rollers, cuckoo rollers, and motmots, all of which have only moderately long and stout bills. The larger, rather stouter, straight bill of the kingfishers is an exaggerated version in birds that often take small invertebrates. The hornbills, with their very large, laterally compressed bills, often somewhat ornamented with a prominent horny casque in the male (smaller in the female), seem to have carried bill size beyond the point of a strictly functional feeding organ. Species that feed largely on fruit plucked from branches as well as species that take lizards and snakes and dig in the ground for insects all have exaggerated bills.

An advantage of a long bill can be seen in hornbills that feed on fruits among the outer branches of forest trees; the long bill enables the bird to reach fruit on slender, outer twigs. There is also the probability, evidenced by the sexual difference in bill ornamentation, that the bill serves in courtship and perhaps in other social contacts within the small parties characteristic of hornbills.

Another type of long bill in the order is that of the hoopoes and the wood hoopoes, which is a slender and slightly, to strongly, downcurved bill. The former use the bill to probe in the ground as it walks along, the latter to poke into crevices and crannies of the barks and branches of trees.

Similarities to birds of other orders. The size of the birds of this order and their propensity to take rather large prey bring them into competition with many other species of other groups. It is perhaps instructive to compare certain Old World groups with representatives of different orders in the New World tropics (see above *Natural history*). The most striking parallel is seen between the toucans (Rhamphastidae, order Piciformes) and the hornbills, which, with their enormous bills, small feet, general diet, behaviour, and appearance, are remarkably alike. A similar degree of convergence is seen between the bee-eaters and the jacamars (Galbulidae, order Piciformes) of tropical America and also between some wood hoopoes (*Rhinopomastes*) and American wood hewers (*Campylorhamphus*). Similar ecological conditions apparently have brought similar adaptations in external structure and behaviour between birds of quite unrelated orders, in widely separated areas.

EVOLUTION AND CLASSIFICATION

Evolution. The history and some of the family relationships within the Coraciiformes are obscure, so that any treatment of the evolution of the order must be considered speculative. The present distribution and abundance of the ten families suggest an Old World origin, probably in the Ethiopian–Indian region, but the southern Palaearctic (Eurasia) may also have been involved; the limited number of fossils, consisting of a few hornbills from the Eocene (about 50,000,000 years ago) and Miocene (about 15,000,000 years ago), a roller from the upper Eocene or lower Oligocene (about 38,000,000 years ago), a possible wood hoopoe from the Miocene, and a kingfisher from the lower Oligocene, all have migrated from Europe. At the periphery of the Ethiopian region, the island of Madagascar has the endemic ground-roller and cuckoo-roller families, probably derived from separate colonizations of the early roller stock; the island was later colonized by the modern rollers, bee-eaters, kingfishers, and the hoopoe. Extending eastward through southern Eurasia, the modern hoopoe (*Upupa epops*) reaches Malaysia (making it the most widespread single coraciiform species); the hornbills have reached the Papuan area; and the roller, bee-eater, and kingfisher families have reached the Australian continent.

In the New World, the early arrival of a protokingfisher stock, via the Bering Strait, probably gave rise to the motmots of Central and South America and the todies of the West Indies. Later, a specialized, fish-eating branch of the kingfishers colonized the New World, evolving one species in the Nearctic region (North America) and several in the Neotropical region.

Considering the relative paucity of roller-like birds in tropical America and their comparative abundance and diversity in the Old World tropics, it seems likely that, by the time the coraciiform stock had reached the neotropics, many niches occupied by members of this order in the Old World were already filled by members of various piciform and passeriform families. The passeriform suborder Tyranni, with more than 600 New World species, is particularly diverse. The presence of highly adapted potential competitors, such as the toucans, jacamars, and puffbirds, endemic members of the Neotropical avifauna, may have retarded the colonization and evolution of the Coraciiformes in the New World.

Classification. *Distinguishing taxonomic features.* The external characteristics on which families are based are the size and shape of the beak and wing and the arrangement and amount of fusion (syndactyly) of the three front toes. The ten families are united, additionally, by features of the palate bones, the tendons of the leg, the configuration of the leg muscles, and the body pterylosis (pattern of feathers).

Annotated classification. The following classification is based on an arrangement used in 1960 by Alexander Wetmore, a U.S. ornithologist.

ORDER CORACIIFORMES (roller-like birds)
Small to medium-large land birds, with body lengths of about 10 to 160 cm. Most species conspicuously coloured. Beak prominent; straight, or slightly or strongly downcurved. Some syndactyly of toes (I, II, and III) in most families. Cavity nesters; young hatched blind and naked (except in Upupidae). Worldwide in temperate and tropical regions. About 191 species.

Family Alcedinidae (kingfishers)
Oligocene to Recent. Chiefly arboreal; short tarsus, small feet; syndactyl. Beak medium to long, straight, stout, usually spearlike. Wings short, rounded. Food: invertebrates and small vertebrates, including fish. About 90 species; distribution that of order, but greatest diversity in Indo-Australian Region; length 10–45 cm.

Family Todidae (todies)
Recent. Chiefly arboreal. Long, straight, flattened, blunt bill. Toes syndactyl. Wings short, rounded. Food: invertebrates, insects. Five species; West Indies; length 9–12 cm.

Family Momotidae (motmots)
Eocene to Recent. Chiefly arboreal. Moderately long, stout, pointed, slightly decurved bill usually with serrate cutting margins. Tarsus very short, toes syndactyl. Wing short and rounded. Food: invertebrates, lizards, and some fruit. About 8 species; South and Central America; length 17–50 cm.

Family Meropidae (bee-eaters)
Pleistocene to Recent. Arboreal and aerial. Bill long, compressed, tapering to a fine point, and slightly decurved. Tarsus short, anterior 3 toes slender, weak, and syndactyl. Wing long and pointed. Food: insects. About 24 species; Africa, southern Eurasia to Australia; 15–35 cm long, including elongated tail feathers.

Family Coraciidae (rollers)
Eocene to Recent. Chiefly arboreal and aerial. Bill stout, crowlike, slightly downcurved, terminally hooked. Tarsus short, foot strong; inner and central toes united at base. Wing long, moderately pointed. Food: chiefly insects. About 12 species; temperate and tropical parts of the Old World, greatest number of species in Africa; length 25–32 cm.

Family Brachypteraciidae (ground rollers)
Recent. Chiefly terrestrial in forest and desert brush. Roller-like birds with longer tarsus and short, rounded wings. Food: small animals of forest floor or desert brush. Five species; Madagascar; length 30–40 cm (including long, graduated tail of some species).

Family Leptosomatidae (cuckoo roller)
Recent. Arboreal and aerial. Bill moderately long, stout, slightly decurved, and terminally hooked. Tarsus very short and, unique in this order, toes semi-zygodactyl (the outer, anterior toe reversible). Wings long, moderately broad and somewhat pointed. Also unique in having a pair of powder-down patches, one each side of rump. Food: large insects, lizards. One species; Madagascar; length about 43 cm.

Family Upupidae (hoopoe)
Pleistocene and Recent; terrestrial and arboreal. Bill long, slender, slightly decurved. Tarsus short, slender; toes, long, with central and outer ones fused at base, claws short. Wing moderate, broad. Food: arthropods, and other invertebrates. One species; Africa, southern Eurasia, and Malaysia; length about 29 cm.

Family Phoeniculidae (wood hoopoes)
Miocene (Europe), Recent (Africa); arboreal. Bill long, slender, slightly curved to sickle-shaped. Tarsus very short; toes long, central and outer ones fused at base. Claws long, curved and sharp. Food: invertebrates chiefly. Six species. Length 22–38 cm.

Family Bucerotidae (hornbills)
Eocene (Europe) to present. Chiefly arboreal (1 species chiefly terrestrial). Large, slightly curved bill, often with casque or sculpturing (larger in males). Tarsus short to very short, toes syndactyl. Wings moderate to long and broad. Unique in order in having eyelashes. Food: insects, small vertebrates, and fruit. About 45 species; Africa, southern Asia to Papuan area; length 40–160 cm.

Critical appraisal. Paleontologists have called the order Coraciiformes a miscellaneous assemblage of birds with large bills and have pointed out that the anatomical differences between the families of birds within this and other orders are minor compared with differences found within some mammalian orders, such as the carnivores which include both cats and seals. The classification of birds into higher categories is justified by expediency and tradition. The roller-like birds include ten groups of birds here called families, the largest, the kingfishers and the hornbills, containing 90 and 45 species respectively, the smallest, the hoopoe and the cuckoo roller, a single species each. Each family is fairly uniform within itself, but between families some anatomical characters have a disconcerting habit of lacking a common linkage. The order does seem to have a common heritage, however, judging from the sum of the characters that place them high in degree of specialization, just below the perching birds (Passeriformes) and the woodpecker-like birds (Piciformes). They have been variously said to be distantly related to the cuckoos, parrots, owls, and even the nightjars, depending on the relative weight given to characters such as the condition of the bony palate, the notching of the sternum, the arrangement of the toes, the thigh muscles and foot tendons, the structure and location of the syrinx (sound-producing organ), the development of the ceca (blind pockets in the digestive tract), and the arrangement of the intestines, the arrangement and type of feathers and down, and the type of nesting.

Some authorities would include rollers, ground rollers, and cuckoo rollers as subfamilies of the Coraciidae. There is also a question as to whether the hoopoes and wood hoopoes are more nearly related to the hornbills or the rollers.

The families fall into six or seven well-defined groups: (1) kingfishers, todies, and motmots, (2) bee-eaters, (3) rollers and ground rollers, (4) cuckoo rollers, (5) hoopoes, (6) wood hoopoes, which are sometimes united with the Upupidae, and (7) hornbills.

BIBLIOGRAPHY. A. NEWTON, *A Dictionary of Birds*, 4 pt. (1893–96), a most important historical review of the classification and natural history of birds, with data on roller-like birds arranged alphabetically; E. STRESEMANN, "The Status of Avian Systematics and Its Unsolved Problems," *Auk*, 76: 269–280 (1959), an important essay with bibliography; R.E. MOREAU, "The Comparative Breeding Biology of the African Hornbills (*Bucerotidae*)," *Proc. Zool. Soc. Lond.*, 107A:331–346 (1937); "The Nesting of African Birds in Association with Other Living Things," *Ibis*, 6:240–263 (1942), includes insect scavengers found in nests occupied by hornbills; H. FRIEDMANN, "The Honey-Guides," *Bull. U.S. Natn. Mus. 208* (1955), includes a section on host relationships of honeyguides, birds which victimize kingfishers, bee-eaters, hoopoes, and wood hoopoes.

(Au.L.R.)

Coral Islands, Coral Reefs, and Atolls

The words coral reef have for centuries evoked a conflicting response in the minds of sailors, "coral" conjuring visions of colour and beauty and "reef" arousing feelings of apprehension. These reefs are masses of carbonate of lime built up from the sea floor by the accumulation of the skeletons of a profusion of animals and algae; eventually they rise to the surface of the water and become a danger to mariners. The several reef types include fringing reefs, which grow along the shores of a continent or island; patch or platform reefs, which occur as isolated patches on a continental shelf; barrier reefs, which commonly are found growing on or near the edge of a continental shelf, roughly parallel with the coastline; and atolls, ring-shaped reefs that in their grandest oceanic form mark the rims of sunken, truncated volcanic cones or guyots (tablemounts).

In ancient times "coral" meant the precious coral of commerce, *Corallium nobile* of the Mediterranean. Subsequently, in the phrase coral reef, it came to imply all the sedentary creatures growing on the reefs; namely, soft coral, stony coral, black coral, blue coral, organ-pipe coral, coralline algae, and so on. Today, when zoologists speak of corals alone, however, they generally mean only the stony corals or Scleractinia. The Scleractinia of today are not confined to coral reefs; those that form reefs are dependent on warmth and sunlight (*i.e.*, are hermatypic). Reef corals grow best in shallow, sunlit water, between the low-water mark and six fathoms (36 feet, or 11 metres), but they can still construct reefs in water as deep as 22 fathoms (132 feet), and they may have a sparse existence between 22 and 30 fathoms (132 and 180 feet). They prefer water of normal salinity and with an annual maximum temperature above 72° F (22° C) but below 82° F (28° C); but their reef-building activities may be carried on in waters whose minimum temperature in winter is not less than 59° F (15° C).

Hermatypic corals and cold-water forms

A second group of corals in present-day seas grow in thickets and coppices that develop banks rather than reefs on the outer, deeper, colder, and darker parts of continental shelves and platforms. These organisms flourish in water with a winter minimum temperature ranging between about 39° and 59° F (4° and 15° C) at depths of about 33 fathoms to 110 fathoms (200 to 660 feet). In any one thicket there are commonly only two genera of delicately branching corals involved. Such coral banks are known along the eastern Atlantic shelf edge (or continental slope) from Norway to the Cape Verde Islands and again off the Niger River Delta and in the west Atlantic around the Gulf of Mexico, the Bahamas, and the Orinoco River Delta. Off New Zealand such banks have been recognized on the Campbell Plateau and the Chatham Rise; they also occur in the northwest Pacific near Japan.

The third coral assemblage of the modern seas is associated with even colder or deeper seas; it consists of small, solitary corals of relatively few genera, known from the abyssal floors of the oceans and from the shelves around Antarctica, Patagonia, and the Falkland Islands in waters 35° to 43° F (2° to 6° C) in temperature.

Buried fossil reefs on ancient continental shelves are targets for petroleum-exploration companies. The porosity of reefs and the characteristic curvature of nonporous enclosing sediments cause them to be prospective reservoirs for oil and gas. The rich oil fields of Alberta, for example, are associated with Devonian reefs (about 345,-

By courtesy of the Geological Society of America

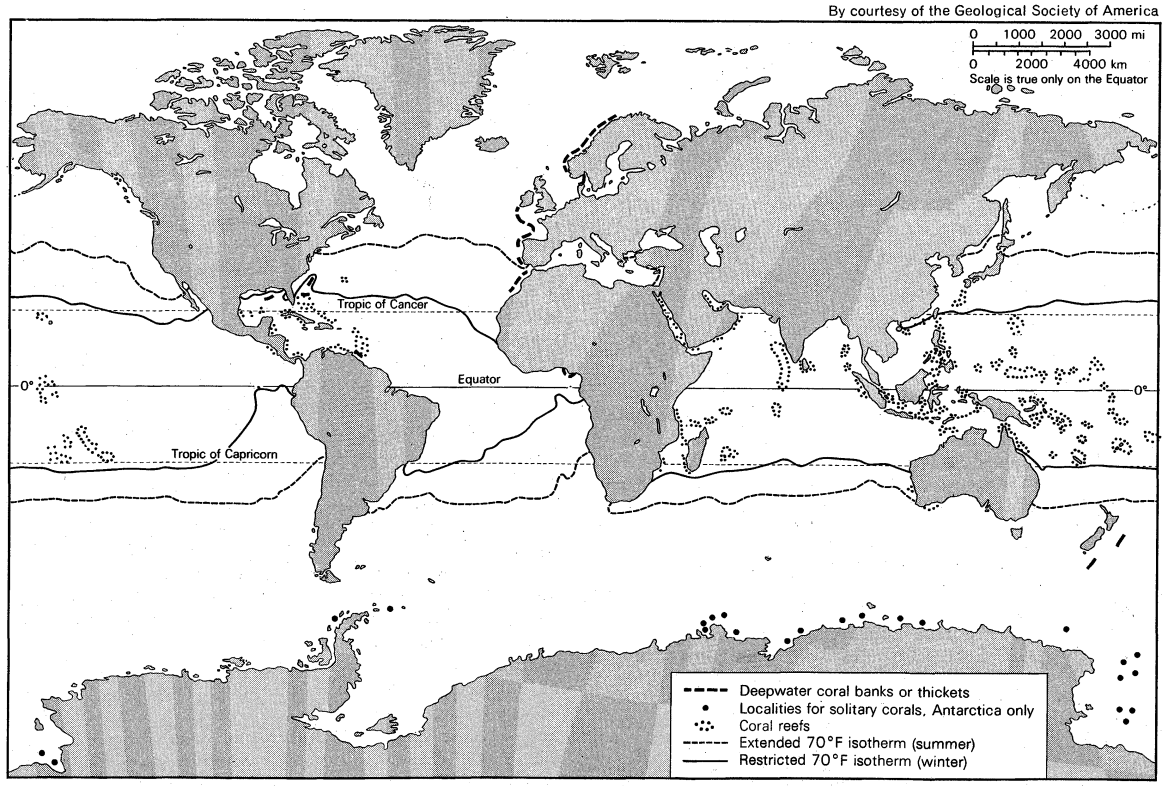


Figure 1: Distribution of coral reefs and their relation to controlling sea-surface temperature.

000,000 to 395,000,000 years old). Fossil reefs recently have become targets for metal prospecting because some corals contain small percentages of metals such as zinc and copper, biogenetically extracted from seawater. A living coral reef may also have economic potential, in that it constitutes a major tourist attraction.

This article describes the reef-building organisms themselves, modes and geographic distribution of reef accumulation today, reefs of the geological past and their environmental significance, factors involved in reef growth, and the several theories of reef development. For relevant information on water and terrain characteristics that influence the location of coral reefs, see OCEANS AND SEAS; LAGOONS; CONTINENTAL SHELF AND SLOPE; MARINE SEDIMENTS; OCEAN BASINS. See also FOSSIL RECORD, for discussion of ancient reefs and their component organisms, and the several geographic articles on particular reefs and atolls, particularly GREAT BARRIER REEF. The concentration of elements by invertebrate organisms in the sea is treated in ELEMENTS, PHYSIOLOGICAL CONCENTRATION OF.

REEF-BUILDING ORGANISMS

The role of organisms in the construction of a reef is threefold: whole skeletons may form a loose or tight framework; broken-up and overturned skeletons form the vast quantities of lime sand and mud that fills the interstices of the framework; and organisms such as encrusting algae, particularly red algae and blue-green algae, cement the framework and the interstitial material together.

Precambrian and Paleozoic forms. In Precambrian sediments (those older than 570,000,000 years) there are no records of animals with mineral skeletons, but some limestones, not unlike the laterally extensive reefs of today, were formed. These limestones were formed by stromatolites, successive mats or felts of tiny filaments of blue-green algae; each mat traps silt particles and becomes encased in a mold of fine calcareous dust that the algae precipitate by physiological processes. Stromatolites are thus sedimentological (layered) rather than skeletal; commonly they developed colonies of a consistent shape and show growth laminae (annual layers) but little or no organic microstructure.

Stromatolites and Archaeocyathids as reef builders

The passage from Precambrian to Paleozoic time (570,000,000 to 225,000,000 years ago) was strikingly marked by the ability of most animal phyla to develop mineral skeletons, commonly calcareous. The first animals to take a framework role in the formation of reefs were Archaeocyatha (primitive invertebrates resembling sponges); the corals did not appear until later. The Archaeocyatha are a phylum with a very brief history. They appeared early in the Lower Cambrian and, except for a few stragglers, became extinct in the early Middle Cambrian. The basic form of their skeleton is an inverted cone ranging from 10 to 20 millimetres (0.4 to 0.8 inch) in average diameter and from 80 to 200 millimetres (3 to 8 inches) in height. The mineral of the skeleton is finely crystalline calcite. The cone may or may not be secured by holdfasts; in some there is only one wall, but in the great majority there are two walls with a central cavity. The walls are held together by radial-longitudinal and transverse plates, and all contain minute pores. Most Archaeocyatha are solitary, and colonial skeletons are rare. The Early Cambrian reefs were small, and the largest of those—from Siberia, where they are best known—attained a diameter of only 33 metres (108 feet) and a height of 25 metres (82 feet). The blue-green algae probably were the main organisms of these reefs, which are thought to have formed on a sea floor about 20 to 50 metres (70 to 160 feet) deep, but both solitary and some colonial forms of Archaeocyatha were also associated. In smaller growths, which are thought to have developed in deeper waters (50 to 100 metres [160 to 330 feet]) red-algal bushes are the main component, with Archaeocyatha found between the bushes. Associates were gastropods, brachiopods, trilobites, and other invertebrates. Also present are skeletons attributed to stromatoporoids, an extinct order of hydrozoans that formed calcareous skeletons of successive lamellae (layers), usually irregularly perforated and connected one to another by pillars of varying types.

In the Ordovician (430,000,000 to 500,000,000 years ago) the first coral skeletons are found, but they are of different orders from Recent corals (those of the last 10,000 years). The exclusively colonial Tabulata, with relatively simple skeletons, were the first to appear, followed by the Rugosa, which included many solitary as well as colonial forms. Stromatoporoids became common, as did the Bryozoa (sea mats). All these orders are found in the small lenticular, or horizontally extended, thin reefs of the Ordovician carbonate provinces. The small reefs of the eastern United States and of Scandinavia are probably the best known.

Colonial forms and stromatoporoids

In the Silurian Period (395,000,000 to 430,000,000 years ago) and the Devonian Period (345,000,000 to 395,000,000 years ago), stromatoporoids, corals, and Bryozoa again were the most abundant organisms of the fossil reefs, which were larger and more like those of today than the Ordovician reefs. Calcareous algae were few, but blue-green algae such as *Stromatactis* are thought to have played an important part. Echinoderms were important associates, particularly the crinoids, or sea lilies, the plates of whose calcareous skeletons were commonly massed by the local currents into banks (crinoidal limestones) abutting the reefs. Brachiopods, bivalves, gastropods, trilobites, and many others abounded in the interstices of the reef framework. The Silurian reefs of the Great Lakes region of North America and the beautifully exposed reefs of the Swedish island of Gotland resemble the patch reefs of the modern Great Barrier Reef. Wenlock Edge in Great Britain was probably a barrier reef. Stromatoporoids and tabulate and rugose corals all reached their acme in the Middle Devonian. In the Belgian Upper Devonian reefs it appears that profuse growths of corals formed on a subsiding sea floor beneath the zone of wave turbulence and grew upward into the zone of turbulence, where stromatoporoids were favoured. The Late Devonian reefs of the Kimberley district of Western Australia show all the sedimentation patterns of modern reefs, as do the Early Carboniferous (280,000,000 to 345,000,000 years ago) reefs of Derbyshire in England.

Carboniferous and Permian (225,000,000 to 280,000,000 years ago) reefs are strikingly different from those of the Devonian because of the great rarity of stromatoporoids and because Bryozoa generally are more abundant than corals. *Stromatactis* cavities abound in bryozoan reef knolls, which may not have grown to wave base (the lower limit of effective wave action). Banks of crinoidal debris are not uncommon. In the Urals, in the Soviet Union, Upper Paleozoic reefs were built up during continued complete submergence through successive ages. The shallowest reef builders were a group of marine hydroids called hydractinia; somewhat lower were the corals; and deeper still were porcelain-like algae or tubiphytes. The Permian barrier reef of Texas, the famous Capitan Reef, is built almost entirely of interlocking remains of encrusting algae and calcareous sponges. Behind the reef wall is a fossil-rich zone less than half a mile wide with filamentous and massive algae and pisolites (calcareous spherules larger than two millimetres in diameter); bryozoa; fusulines (single-celled marine organisms); large, heavy gastropods; heavy-shelled brachiopods; bivalves; straight cephalopods; and holothurians. In the forereef, on the upper slopes, is a belt of filamentous algae with sponges. Corals are rare.

Crinoids, Bryozoa, and Late Paleozoic reefs

Mesozoic and Cenozoic forms. In the Early Mesozoic reefs (from 65,000,000 to 225,000,000 years ago), the order of corals that flourishes today, the Scleractinia, first appears. It may or may not have evolved from the rugose corals that became extinct at the end of the Paleozoic. The Triassic reefs of the Alpine region were dominantly constructed of algae, but more normal coralline reefs are present in very thick (up to one kilometre [0.6 mile]) limestone. Jurassic reefs are quite like those of today. The famous reef structures of the southern Jura contain many corals, algae, and stromatoporoids, as well as rudistid bivalves, oysters, brachiopods, gastro-

pods, echinoderms, and other invertebrates; these occur in nests in the white oolite among the irregular masses of unbedded coral limestone. Rudistid bivalves are very large; one valve is horn-shaped and the other fits into it like a lid; in the Cretaceous Period (65,000,000 to 136,000,000 years ago), particularly in the Tethyan region, an area occupied by ancient continental seas, they grew profusely and formed rudistid reefs; they were extinct by the end of the Mesozoic.

Cenozoic reefs (originating during the last 65,000,000 years), particularly the Miocene reefs below the atolls and barrier reefs of the Pacific, are very similar to modern types. Corals and calcareous algae are dominant and nearly all invertebrate orders are represented in their interstices or framework.

PRESENT REEF ACCUMULATION

Tropical water conditions. Water conditions favourable to the growth of reefs exist in tropical or near-tropical surface waters. Regional differences may result from the presence or absence of upwelling currents of colder waters or from the varying relation of precipitation to evaporation.

Effects of light, turbidity, and nutrients

Tropical seas are well lit, the hours of daylight varying with the latitude. Light intensity and radiant energy also vary with the depth. Thus, at latitude 32°44′ N (Madeira Island) the "day" in March has a length of 11 hours at a depth of 20 metres (70 feet), five hours at 30 metres (100 feet), and only about a quarter of an hour at 40 metres (130 feet). Nearer the pole these figures decrease further. Light intensity has a profound effect on the growth of the individual reef-coral skeleton, because of the symbiont zooxanthellae of reef corals (see below *Biological factors*). The number of species present on a reef may also be related to light intensity and radiant energy.

Turbidity may be high in lagoons, where shallow water lies over a silt-covered sea floor and where storms and windy periods cause considerable disturbance of the bottom silt. The average transparency may be low (about 12 metres [39 feet]), and light penetration is reduced.

Inside the Great Barrier Reef, on the shallow continental shelf of Queensland, the oxygen content of the water is high, exceeding 90 percent saturation most of the time; in deeper water, during the calm periods of the rainy season, the saturation may fall to about 80 percent. Plant nutrients such as phosphate and nitrate show no seasonal change in quantity; both are present in very small quantities throughout the year. Constant mixing of the shallow sea prevents any stratification of the nutrients. As a result, growth of phytoplankton is possible and almost uniform throughout the year, providing a constant supply of food for the zooplankton which in turn form the chief food supply of the corals. Some nutrients enter the lagoonal waters with the oceanic water that flows through the reef openings, but the dissolved phosphates in the lagoons are probably derived chiefly from bacterial decomposition of the organic matter on the sea-bottom, as well as from detritus swept in from the reef surfaces. This environmental pattern is typical of many atoll lagoons.

Geochemistry of reefs. Minute quantities of metallic elements are present in solution in seawater and also occur in marine invertebrate skeletons, though not in the same proportions as in the surrounding water. Magnesium and strontium are the most frequently occurring trace elements in reef skeletons and are measured in parts per thousand, but barium, manganese, and iron are also present and can be measured in parts per million. In Pacific corals, 2.17 parts per million of uranium have been found; and in Florida coral, 2.36–2.95 parts per million. Strontium is concentrated in aragonitic skeletons, and magnesium in calcitic skeletons; coral aragonite has a higher strontium content than (some) molluscan aragonite; the magnesium content in the calcite of coralline algae is high; that of barnacle shells is low (11.5 parts per thousand). By identifying these trace elements and their degree of assimilation in different organisms, sediments formed predominantly of coral-skeletal detritus can be distinguished from sediment derived chiefly from mollusks or coralline algae.

Trace elements in reef skeletons

Quite recently it has been shown by atomic-absorption spectrophotometry that ultratraces of metals are present in the aragonite skeleton of the hydrozoan coral *Millepora* from a reef flat on the Coral Sea Plateau off Queensland. These are, in parts per 1,000,000,000: lead (100), copper (71), cadmium (23), cobalt (17), nickel (1,480), iron (507), and zinc (507).

The Pine Point lead–zinc ore body on the southern shore of Great Slave Lake, Canada, occurs mainly within or close to the Middle Devonian reef sequence known as the Presqu'ile Formation. Geologists speculate that these metals may have been extracted from seawater by Devonian corals or other animals and subsequently concentrated within the reef rock by postdepositional changes.

Another aspect of reef geochemistry is the carbon and oxygen isotopic composition of coral skeletons and shells. Determination of the number of carbon isotopes present provides a method of assessing the age of a sample, and determinations of oxygen isotopes present are useful in indicating water-temperature changes that occurred during the period of growth of the reef.

Winds, currents, temperature, and salinity. Winds and currents are important in shaping individual reefs and in determining the orientation, shape, and position of the coral sand cays, or "low islands," that develop on reefs. Currents are primarily those generated by the prevailing winds, but, in areas where the tidal range is great, tidal effects may become paramount.

Cays may be round; oval, or boat-shaped; or irregular in outline. They originate when sediment is lifted from the reef surface and carried leeward by waves or tidal currents and then deposited where the water velocity is reduced abruptly. Thus, they commonly form on the more protected leeward end of the reef. Wind action at low tide on these deposits may build dunes above the high-water mark. Beach rock may form by carbonate cementation of grains in deposits lying between tide levels; it then acts as a stabilizing factor. Storm waves may drive forward coral fragments derived from "stag-horn" corals growing on the windward slopes of the reef, forming shingle banks; successive, superposed banks may thus be formed. The shingle on the banks may become cemented and thus add considerable stability to the cay, as does the growth of vegetation. Hurricanes, however, may carve back the shorelines of even stabilized cays. Huge, isolated boulders of coral or coral limestone are fairly common along reef margins. Some may be remnants of a once-emergent reef platform; others are hurricane or storm jetsam.

Coral reefs are best developed where the mean annual surface-water temperatures are approximately 73° to 77° F (23° to 25° C); no significant reefs occur where such temperatures fall below about 64° F (18° C), although a few reef-coral species can exist in temperatures considerably below this. Seasonal temperature differences on any one reef are usually slight, as are differences due to depths of water or situation on the reef.

Restrictions on reef formation

Seawater of normal oceanic salinity (between 30 and 40 parts per thousand), to which corals are restricted, is normally supersaturated in calcium carbonate ($CaCO_3$), so that adequate ionized calcium (Ca^{2+}) is available for the skeleton-forming process. Floods of freshwater may destroy life on inshore fringing reefs. A luxuriant reef on Stone Island, near Bowen, Queensland, was killed to a depth of 3 metres (10 feet) below mean tide level by a week of cyclonic rains, in which 90.7 centimetres (35.7 inches) of rain coincided with full-moon spring tides.

Biological factors. The most significant biological determinant of reef accumulation is the presence of zooxanthellae in the living tissues of all reef corals and of many massive-shelled mollusks (Tridacnidae) and other shelled invertebrates, as well as in the soft-bodied hydrozoans, scyphozoans, and anthozoans. Zooxanthellae are now known to represent the vegetative stages of dinoflagellate algae, and their association with reef corals is symbiotic—that is, mutually helpful. In temperate seas they occur only occasionally. Their profusion in reef animals is no doubt connected with the greater light intensity and radiant energy of reef waters, for, like other

plants, zooxanthellae require sunlight for photosynthesis. They remove at the source part of the carbon dioxide (CO_2), together with nitrogen, phosphorus, and (doubtless) sulfur, produced by metabolic breakdown within the coral and which would otherwise be excreted by the corals. They greatly aid in the formation of the coral skeleton by increasing the speed with which the CO_2 produced in coral metabolism is removed and the speed with which the skeletal $CaCO_3$ is formed. Corals also may gain some nutrient from their zooxanthellae, but they probably do not need the oxygen produced during photosynthesis.

The productivity of reefs is a current focus of interest. A constant supply of food in the form of zooplankton is essential to reef corals, which are carnivorous. The zooplankton supply is dependent upon an adequate phytoplankton supply, and the phytoplankton, in turn, require an adequate supply of plant nutrients dissolved in the water. An atoll in the open ocean may be compared with an oasis in a desert, as a localized centre of high productivity. In the warm, well-lit, and well-mixed lagoon waters there is a rapid turnover of the endemic planktonic (floating or swimming) and benthic (bottom-dwelling) population, perhaps 12.5 times per year. Carbonate skeletal matter is accumulated perhaps 1,000 times faster on the summit of an atoll than in the surrounding deeps.

The menace of the crown-of-thorns starfish

Certain biological factors may contribute to the destruction of coral reefs: the fish and invertebrates that feed on the soft tissues of reef builders and the organisms that bore into coral rock. Of the former, the most destructive such enemy yet known is *Acanthaster planci*, the crown-of-thorns starfish, which, during the last decade, has multiplied spectacularly and has removed the soft tissues from large areas of many reefs in the southwest Pacific. *A. planci* feeds by everting its stomach and liquifying and absorbing the tissues of the corals. Coral-rock borers include boring algae, boring sponges (of great significance), various polychaete and sipunculid worms, and many bivalves and a few gastropods. These organisms usually penetrate the rock mechanically but in some cases do so chemically. Extensive damage is caused both by their own activities and by the assistance they give to the erosive action of the sea.

Indo-Pacific and Atlantic communities. Two major coral-reef communities are clearly distinguishable in shallow tropical seas—the Indo-Pacific and the Atlantic. The corals of the Indo-Pacific community comprise about 80 genera and about 700 species, but the number of these that may be found on any one reef or set of reefs varies from region to region. Members of the community may be found from the Red Sea and east coast of Africa eastward across the Indian Ocean, through the East Indies, on to the Tuamotu Archipelago; and, indeed, they are still recognizable as far east as Panama. Within this great province the richest faunas may be found in optimum temperature conditions between the East Indies and the central Pacific; the number of species and of genera decreases both latitudinally and longitudinally from this central region. Latitudinal limits to distribution are in part imposed by the inhibiting effect of the lower temperatures on spawning.

The Atlantic coral-reef community is found in the Gulf of Mexico, the Caribbean Sea, and the West Indies; on the northern and eastern coasts of South America, south to Rio de Janeiro; and in the Gulf of Guinea off West Africa. It includes only 26 genera and 41 species. *Acropora* and *Porites*, as in the Indo-Pacific, are the two most important coral genera, but in the Atlantic each is represented by only three species, as against 150 and 30, respectively, in the Indo-Pacific. Nevertheless, there is prolific growth on many Atlantic reefs. On Pacific reefs, many of the reef-coral species can live exposed between tides, but there are few Atlantic reefs where living coral can be studied exposed at low tide.

A notable feature of Atlantic coral-reef communities is the fact that particular environmental niches may be occupied by a genus different from that occupying the same niche in an Indo-Pacific community, yet these separate genera may exhibit the same growth form and superficial appearance. Thus, the Atlantic faviid *Montastrea* occupies the niches proper to the similar looking faviid *Plesiastrea* of the Indo-Pacific. There are at least 10 other pairs of similar genera. Many important Indo-Pacific genera are not represented in the Atlantic by such ecological equivalents. These include *Pocillopora*, *Stylophora*, *Seriatophora*, and *Montipora*, although all occur in the Tertiary (about 2,500,000 to 65,000,000 years ago) faunas of both the Atlantic and the Indo-Pacific. The Atlantic reef coral fauna as a whole seems to be a weakening relic of that of the mid-Tertiary Tethyan seas.

Many characteristic reef-coral associates are either absent or greatly reduced in variety in the Atlantic. Thus, calcareous algae play a minor role; there are no giant clams like *Tridacna*, no coral gall crabs, and no giant anemones with commensal fish and crustaceans; many mollusks present in great variety in the Indo-Pacific are missing or uncommon. One of the most striking differences is the absence of the alcyonarian corals *Heliopora* and *Tubipora* in the Atlantic; in addition, soft corals are a minor element. Gorgonian corals are very abundant in the Atlantic, but both alcyonarians and gorgonians are relatively insignificant in the Indo-Pacific.

ORIGINS AND DEVELOPMENT OF REEFS

Darwin's theory of atolls

The English evolutionist Charles Darwin concluded in 1842 that barrier reefs began as reefs fringing the land around which they now form a barrier, and that oceanic atoll reefs began as reefs fringing a volcanic island. Subsidence of the land fringed was thought to allow the reef to grow upward (and outward over its own forereef debris). Maximum growth would occur at the seaward edge, and lagoons would develop between the ascending barrier, or atoll, reef and the land or volcanic cone. When the volcanic cone became completely submerged, the

From *Geomorphology* by A. Lobeck. Copyright 1939. Used with permission of McGraw-Hill Book Company.

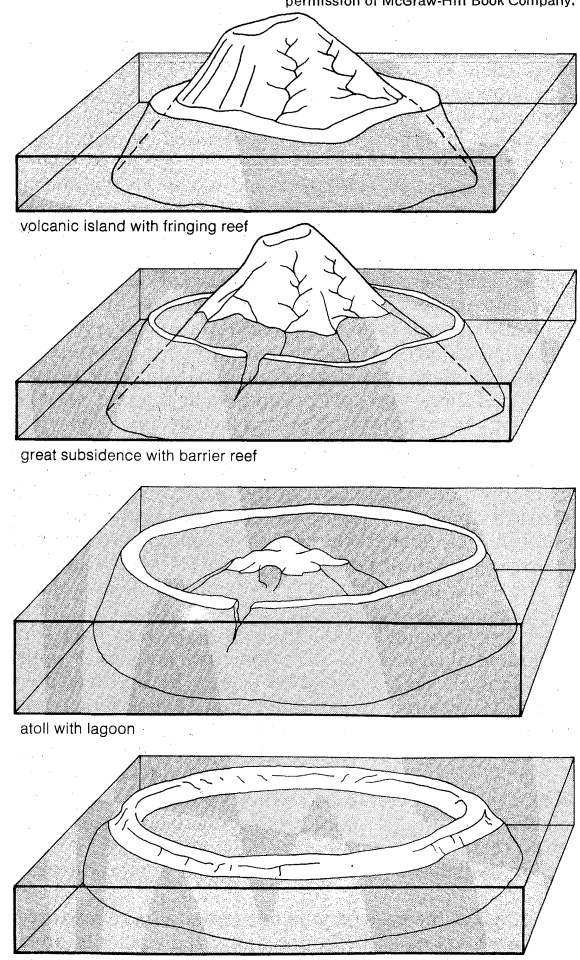


Figure 2: Stages in reef development on a subsiding island.

atoll lagoon would contain only coral islands. Fundamentally, Darwin's concept is still valid, although many consider submergence by the rise of sea level, following melting of Pleistocene ice sheets, to be a better explanation of the latest upward growth of many reefs, particularly on continental shelves. A reef whose surface lies above high-tide mark, either by uplift or by eustatic regression of the sea (that determined by ice sheet–sea level relations), is subject to planing by marine erosion. If planing off is complete, a flat-topped submerged platform results; if subsidence or eustatic submergence intervenes, a wave-cut terrace is left around the reef. Terraces that may have formed in this way are known around many reefs. Some annular reefs may develop without relation to subsiding volcanic cones. When reef platforms have been uplifted above sea level, they are subjected to subaerial erosion. Surface slope, or gradient, determines the amount of runoff and is a prime factor in this erosion. Two secondary processes also are involved: (1) case hardening of steep, bare limestone surfaces by recrystallization caused by alternate wetting and drying, so that walls or knife edges result from weathering; and (2) continuous subsoil solution, if surfaces are nearly horizontal and runoff is diminished. These processes combine to produce a prominent rim and a saucer-shaped interior in emerged limestone islands. With submergence, algal and coral growth resumes, the fastest growth being on the rim and on any pinnacles that may be left. Thus an atoll or annular reef may develop along the rim around the low-lying central region, which becomes a lagoon, and coral knolls grow on former pinnacles in the lagoonal area.

Types of reefs. Fringing reefs form a veneer in the shallow water near or at the shore of the mainland or of islands. On shorelines where bays receive large quantities of terrestrial mud, sand, and freshwater, fringing reefs are intermittent and are restricted to promontories. Along limestone coasts, however, coastal erosion is by solution, little mud or sand is supplied, and coral growth may be almost continuous along the shore. Fringing reefs may extend as far as 1,500 metres (4,900 feet) from shore; they show ecological zonation parallel to the shore. Along mainland shores with easily eroded rocks, mudbanks may be washed into the reef flat and colonized by mangroves. Inner parts of reef flats may show low mesas of middle Holocene (perhaps 5,000 years old) and even Pleistocene emergent reefs that have not been quite planed off by marine abrasion. Geomorphological features peculiar to raised fringing reefs have been described for the Solomon Islands, and there the complex problem of dead reefs and dead patches on reefs arises. The surface of the reef flat is mostly of dead coral, but pools occur in which live coral colonies flourish. An algal rim formed by growing red calcareous algae may develop on a fringing-reef margin if the reef faces strong waves and swells the year around. This rim is commonly less spectacular than the algal ridge of the windward edge of Pacific oceanic atolls and may be developed merely as an algal platform, or pavement, on which algal encrustation over corals is thin. The seaward slope of a fringing reef, like the seaward slope of an atoll reef, is characterized by a zone of grooves and spurs to a depth of perhaps 15 metres (50 feet), and it is in this zone that vigorous growth occurs. Under stable shelf conditions, the fringing reef will extend outward as the spurs elongate and the grooves fill or roof over with coral and algal growth. In tectonic belts (zones of uplift and deformation), fringing reefs may be uplifted intermittently, resulting in parallel, stepped subaerial terraces such as those of the Finsch Coast of New Guinea.

Growth of fringing reefs

Platform and patch reefs are characteristic of continental shelves; they may or may not lie behind a barrier reef. Reefs grow actively outward as well as upward, especially in the stable conditions of a continental shelf. Any given reef, having depth and temperature fixed by its location, will have its shape determined by the direction and force of the water currents that bring food and by the shape of the base on which it grows. Where the forces of growth are equal in all directions, radial expansion results in platform-like reefs. With further radial growth, lagoonal platform reefs develop. If the reef grows on a sand bank, elongation may result. The shape of an elongated platform reef may be determined by the orientation of rising and falling tidal currents; these may be directly opposed to each other. The boat-shaped reefs of Torres Strait, between Australia and New Guinea, apparently developed in such a pattern. Where wave-generated currents are asymmetric, horseshoe reefs develop, with convexity facing the current and the leeward ends curving round to partly surround a lagoon. Low Isles, made famous by the Great Barrier Reef Expedition of 1928–29, is the best known example of this type. A sand cay (or cays) commonly develops on one or both of the leeward wings. Those parts of Pacific platform reefs that face strong and persistent currents characteristically have a low algal rim from which radiate grooves and spurs.

Growth of platform and patch reefs

Barrier reefs commonly present to the ocean and the trade winds a steep wall, in some dropping abruptly 1,000 to 5,000 metres (3,300 to 16,400 feet). On the lagoon side they grade off gently with a wedge of sediment dotted by small patch reefs, coral knolls, and coral heads; depths in the lagoon may reach 50 to 80 metres (165 to 260 feet). According to the strength of surf and swell, an algal ridge, rim, platform, or pavement develops and is commonly emergent at low water. The seaward slope has radial grooves and spurs, the grooves forming surf channels and imparting great wave resistance. Ecological zonation on the seaward slope is difficult to study because of the dangers of surf and swell. Zonation on the reef flat, parallel to the wall, is dependent mainly upon depth.

Character of barrier and ribbon reefs

From *Encyclopedia of Geomorphology* by R. Fairbridge, copyright © 1968 by Litton Educational Publishing, Inc.; and C. Yonge, *Advances in Marine Biology*, Vol. 1 (1963), Academic Press

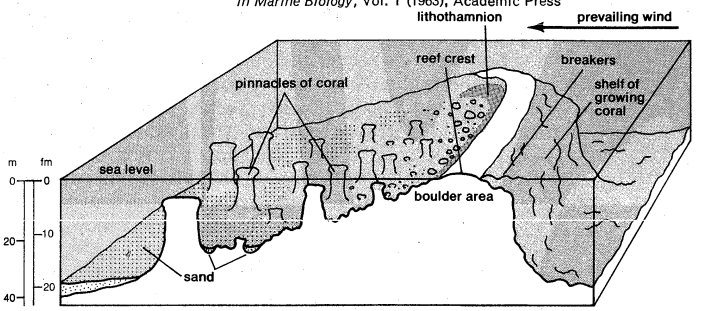


Figure 3: Section of an outer barrier reef, showing the relation of form to prevailing winds.

A characteristic form in a barrier reef system is the wall, or ribbon, reef, emergent at low tide, such as Yonge Reef, Queensland. A ribbon reef flat is commonly only 300 to 450 metres (1,000 to 1,500 feet) wide from the seaward wall to its lagoonward edge; its ends may curve leeward and border the passages between it and the next reefs in line. Rarely, there may be an unvegetated sand cay. A wall reef may develop irregular leeward prongs normal (perpendicular) to its axis by vigorous coral growth favoured by augmented turbulence associated with a high tidal range. The open leeward zone may be very wide (three kilometres or two miles) and support large, scattered reef clumps that are always submerged.

Atolls. The oceanic atoll reefs of the Pacific rise from volcanic cones that have subsided, probably intermittently, in areas of oceanic deeps. According to the Darwinian subsidence theory, the annular atoll reefs extend and grade downward into barrier reefs, which originated as reefs that fringed the volcanic cone. On the other hand, the compound atoll reefs of the Indian Ocean, such as the Maldive Islands and Laccadives, and of the Coral Sea Plateau of the Pacific Ocean, are believed to have grown above foundered continental (rather than oceanic) crustal segments. The Nicaraguan atolls rise from a sea floor of 1,000 metres (3,300 feet) or more, in a volcanic province.

Atoll environments and characteristics

The basic environments of the atoll are the seaward slope, the reef flat, and the lagoon (including patch reefs). Each of these has windward sections intergrading with leeward sections. The ecological environments and depth zones of the oceanic atolls of the Pacific have been well studied, and the results from Bikini and other Marshall islands are summarized here.

The seaward slope of an atoll includes its intergrading windward and leeward sections. On the seaward slope to windward, observations are sparse because of the dangers of the heavy surf, but radial spurs and grooves are developed, down to perhaps 15 metres (50 feet), and surf-resistant coral heads grow in the grooves. The seaward slope to leeward of an atoll is much steeper, probably the result of smaller accumulation of talus; it has a rich upper zone dominated by great brackets of anastomosing (braided) and spreading branches of *Acropora*, beginning at about 3 metres (10 feet) and extending down to ten or 20 fathoms (60 to 120 feet). From about ten to 50 fathoms (60 to 300 feet), around the atolls of the Marshall Islands is a zone of such platelike corals as *Echinophyllia* and *Oxypora*. From 50 to about 80 fathoms (300 to 480 feet) in the Marshalls, below the zone of optimum growth of reef corals, there are delicate *Leptoseris* and other small colonies that are rather widely spaced. Below 80 fathoms (480 feet) only nonreef corals are known.

On the seaward slope of the windward side of Caribbean atolls, elkhorn *Acroporas* oriented with the surf grow profusely down to nearly three fathoms (18 feet), and below this domed colonies of *Montastrea* and *Porites* are common.

The reef flat of an atoll may be divided into outer (seaward) and inner (lagoonward) components, and these may each be described as windward or leeward, in relation to the direction of wind and waves. Thus, on the windward side of an atoll, the seaward part of the reef flat is windward, whereas the lagoonward part of the reef flat is leeward; and on the leeward side of an atoll, the seaward part of the reef flat is leeward, whereas the lagoonward part is windward.

The inner plus the outer flats of the windward side of a Pacific atoll are closely comparable with the ribbon reefs of the Pacific Barrier Reefs; the inner plus the outer flats of the leeward side are comparable to the windward side of an Atlantic atoll, for an algal ridge is absent from both.

The most notable feature of the seaward part of the windward reef flat of the Indo-Pacific atoll is the algal ridge, which dries at low water and is dominated by red, encrusting calcareous algae; the ridge is grooved radially seaward by surge channels. The seaward part of the reef flat of a windward reef extends inward from the algal ridge for a few tens or a few thousands of feet. Ecological depth zones may be distinguished that more or less parallel the algal ridge. First is a coral–algal zone, which may exhibit the greatest coral growth on the reef flat; then, in a low-tide depth of two to three feet or more, micro-atoll zones may be found or, alternatively, a green-algal zone.

The seaward part of the leeward sector of the reef flat of a Pacific atoll characteristically carries a richer growth of corals on the margins. The algal ridge is much less developed or even absent, and the slope from the shore (if there is a cay) toward the outer margin is relatively uniform. The marginal environments are like those of the windward flats, except that, where the algal ridge is very weak, the zone is dominated by low-growing species of *Pocillopora*, and, where the algal ridge is absent, by corymbose *Acropora*. Inward from these, the flat may be nearly barren of corals.

Lagoon reefs on atolls

The reef flat inside the cays and sloping gently away from them toward the lagoon is regarded as part of the lagoonal environment. The lagoonward part of the windward sector of an atoll reef flat is leeward to the prevailing wind and in character is broadly similar to the seaward part of the leeward sector that lacks an algal ridge. It is, however, often relatively narrow and more sloping. Depths no greater than a fathom may extend 30 metres (100 feet) out from the cay shore. The coral patches tend to be elongated normal (perpendicular), rather than parallel, to the shore, with irregular sand-floored patches between them. Massive and branching corals grow luxuriantly and freely in the reef patches, and single colonies may attain great size—up to 6 metres (20 feet) across and three or more metres (10 feet) high. Corals are dominant, but there is a rich variety of crustaceans, echinoids, mollusks, soft alcyonarians, and the green alga *Halimeda*. On the opposite side of the lagoon, where the lagoonward part of the leeward sector of the atoll reef flat is exposed to the prevailing winds and lagoon waves, red calcareous algae are able to grow to some extent, paving the reef flats and developing low algal ridges comparable to those of seaward reefs. The flat is rather wide; cemented platforms slope from the shore at first to depths of two to three feet at low tide, then rise gently upward to a margin that may be awash at low tide, with a fairly steep talus slope down to the lagoon floor. Reef-coral zonation is roughly parallel to the margin, but the zones are less well defined than in the seaward flats, and an uneven, thin veneer of sand and silt makes coral growth patchy.

From the edge of the flat there may be a fairly steep slope into the lagoon, which may reach a depth of 27 to 30 fathoms (162 to 180 feet). Shrublike *Acropora* is abundant. Coral knolls rising from the floor of the lagoon bear a profuse coral assemblage in and around their summits; both branching and massive colonies are found. Calcareous algae are a minor element, but soft corals, usually absent from seaward reefs and scarce on lagoonward reef flats, are common. The general assemblage is a modified leeward–lagoonward reef assemblage living under optimum conditions in somewhat deeper waters.

Although there is minor variation in development from atoll to atoll, ecological environments on atoll reefs are basically similar. Profusion and variety of growth, together with specialized forms of coral colonies, are to be found in similar situations, particularly in regard to depth and the movement of water rich in food. The ecological environments of the annular reefs of the Atlantic differ more in degree than in kind; the algal ridge or rim is absent, and there is much less coral emergent at low tide.

BIBLIOGRAPHY

Classic works: R.A. DALY, "Glacial Control Theory of Coral Reefs," *Proc. Am. Acad. Arts Sci.*, 51:157–251 (1915); CHARLES DARWIN, *The Structure and Distribution of Coral Reefs* (1842, reprinted 1962).

Classic review articles on reef structure, ecology, and growth: R.W. FAIRBRIDGE, "Recent and Pleistocene Coral Reefs of Australia," *J. Geol.*, 58:330–401 (1950); F.S. MACNEIL, "The Shape of Atolls: An Inheritance from Subaerial Erosion Forms," *Am. J. Sci.*, 252:402–427 (1954); A.P. ORR and F.W. MOORHOUSE, "Variations in Some Physical and Chemical Conditions On and Near Low Isles Reef," *Sci. Rep. Great Barrier Reef Expedition*, 2:87–98 (1933); T.A. STEPHENSON *et al.*, "The Structure and Ecology of Low Isles and Other Reefs," *ibid.*, 3:17–112 (1931); J.W. WELLS, "Recent Corals of the Marshall Islands," *Prof. Pap. U.S. Geol. Surv. 260–I*, pp. 385–486 (1954); "Coral Reefs," *Mem. Geol. Soc. Am.*, 67:609–631 (1957); C.M. YONGE, "The Biology of Coral Reefs," in F.S. RUSSELL (ed.), *Advances in Marine Geology*, vol. 1, pp. 209–260 (1963); "Living Corals," *Proc. Roy. Soc.*, Ser. B, 169:329–344 (1968).

Current works: W.G.H. MAXWELL, *Atlas of the Great Barrier Reef* (1968), an excellent description, with illustrations, of this well-known reef.

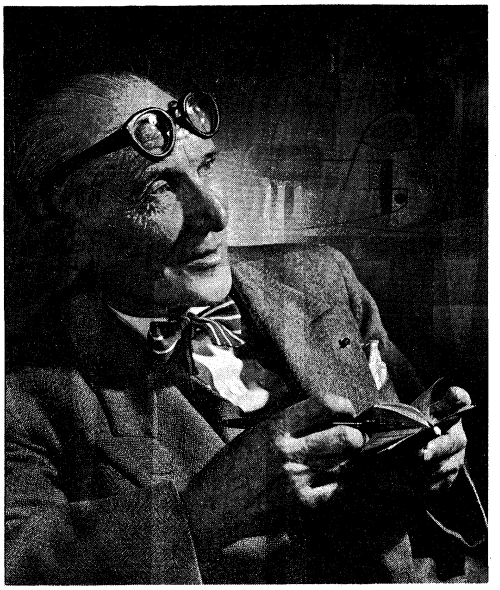
(D.Hi.)

Corbusier, Le

Architect, city planner, and painter, Le Corbusier belonged to the first generation of the so-called International school of architecture and was their most able propagandist in his numerous writings. In his architecture he joined the functionalist aspirations of his generation with a strong sense of expressionism. He was the first architect to make a studied use of rough-cast concrete, a technique that satisfied both his taste for asceticism and for sculptural forms.

Le Corbusier was born Charles-Édouard Jeanneret on October 6, 1887 in La Chaux-de-Fonds, a small town in the mountainous Swiss Jura region, since the 18th century the world's centre of precision watchmaking. All his life he was marked by the harshness of these surroundings and the puritanism of a Protestant environment. His father was an engraver and enameller of watch faces; his mother, a piano teacher.

Education. At 13 years of age, Le Corbusier left primary school to learn his father's trade at the École des Arts Décoratifs at La Chaux-de-Fonds. There, Charles



Le Corbusier, photograph by Yousuf Karsh, 1954.
© Karsh—Rapho Guillumette

L'Eplattenier, whom Le Corbusier later called his only teacher, taught him art history, drawing, and the naturalist aesthetics of Art Nouveau, a contemporary style of decoration that utilized an abundance of convoluted and sinuous plantlike motifs.

It was L'Eplattenier who decided that Le Corbusier, having completed three years of studies, should become an architect and gave him his first practice on local projects. From 1907 to 1911, on his advice, Le Corbusier undertook a series of trips that played a decisive role in the education of this self-taught architect who, unlike well-known contemporaries such as Walter Gropius and J.J.P. Oud, never continued his formal artistic training.

During these years of travel through central Europe and the Mediterranean, he made three major architectural discoveries. The Charterhouse of Ema at Galluzzo, in Tuscany, provided a contrast between vast collective spaces and "individual living cells" that formed the basis for his conception of residential buildings. Through the 16th-century Late Renaissance architecture of Andrea Palladio in the Veneto region of Italy and the ancient sites of Greece, he discovered classical proportion. Finally, popular architecture in the Mediterranean and in the Balkan peninsula gave him a repertory of geometric forms and also taught him the handling of light and the use of landscape as an architectural background.

Early architectural influences

During the same period Le Corbusier also met many of the European pioneers of the so-called Modern Movement in architecture. He claimed to have worked in Vienna for six months (1907) with Josef Hoffmann, head of the Viennese Secession (an association based on the principles of the English Arts and Crafts Movement). He was also associated in Paris for 15 months (1908–09) with Auguste Perret, an architect who was concerned with the importance of structure and the integrity of building materials and who first showed Le Corbusier the utilitarian and aesthetic aspects of reinforced concrete. Finally, Le Corbusier worked in Berlin for five months (1911) with Peter Behrens, who also formed the architects Walter Gropius and Ludwig Mies van der Rohe.

From 1912 to 1917, Le Corbusier was a teacher at his old school in La Chaux-de-Fonds. At the same time, he constructed several houses marked by the influence of Perret and Frank Lloyd Wright that he later considered less than worthy of inclusion among his "complete works."

Founding of "L'Esprit Nouveau." At the age of 30, in 1917, he returned to live in Paris, where his formation was completed a year later when he met the painter and designer Amédée Ozenfant, who introduced him to sophisticated contemporary art and had him paint his first pictures. Ozenfant initiated Le Corbusier into Purism, his new pictorial aesthetic that rejected the complicated abstractions of Cubism and returned to the pure, simple geometric forms of everyday objects. Relations between Ozenfant and Le Corbusier, for the few years of their friendship, were constant and fervent. In 1918 they wrote and published together the Purist manifesto, *Après le cubisme*.

In 1920, with the poet Paul Dermée, they founded a polemic avant-garde review, *L'Esprit Nouveau*. Open to the arts and humanities, with brilliant collaborators, it presented ideas in architecture and city planning already expressed by the turn of the century architects Adolf Loos and Henri van de Velde, fought against the "styles" of the past and against the elaborate nonstructural decoration then prevalent in architecture, and defended functionalism and the newfound values of industrial society.

The association with Ozenfant was the beginning of Le Corbusier's career as a writer. Ozenfant and Le Corbusier (then still known as Jeanneret) together wrote a series of articles for *L'Esprit Nouveau* that were to be signed with pseudonyms. Ozenfant chose Saugnier, the name of his grandmother, and suggested for Jeanneret the name Le Corbusier, the name of a paternal forebear. A group of articles written between 1920 and 1921 by Le Corbusier were collected and published under the title *Vers une architecture*, illustrated with photographs selected by Ozenfant. Later translated as *Toward a New Architecture*, the book is written in a telling style that was to be characteristic of Le Corbusier in his long career as a polemicist. "Architecture is the conscious, correct and magnificent interplay of volumes assembled under light," is a definition that has become classic. "A house is a machine for living in" and "a curved street is a donkey track, a straight street, a road for men" are among the famous declarations of Le Corbusier. His books, whose essential lines of thought were born of travels and lectures, hardly changed at all in 45 years, constituted a bible for succeeding generations of architects. Among the most famous are *Urbanisme*, *Croisade*, *Quand les cathédrales étaient blanches*, published after his first trip to the United States, *La Charte d'Athènes*, and *Entretien avec les étudiants*, *Propos d'urbanisme*, and *Les Trois Établissements humains*.

Le Corbusier's definitions of architecture

Association with Pierre Jeanneret. *L'Esprit Nouveau* was the springboard for Le Corbusier's entrance into practice. In 1922 he became associated with his cousin Pierre Jeanneret, a former associate of Perret, and together they opened a studio at 35 rue de Sèvres, which Le Corbusier kept open until his death. The association of the two cousins, however, lasted only until 1940. It corresponds to the first of the two main periods, separated by World War II, that can be distinguished in Le Corbusier's work; the second period covers the years from 1944 to the architect's death in 1965. It has become clear that, during the years 1922 to 1940, Pierre Jeanneret's contribution to what had generally been considered the work of Le Corbusier alone was, in fact, decisive. Though Le Corbusier did the drawings and designs, it was Jeanneret who oversaw the actual work and who tackled technological problems that Le Corbusier never pretended to master.

Early residential and city plans. The period from 1922 to 1940 was as remarkably rich in architecture as in city planning projects. As was always to be the case with Le Corbusier, unbuilt projects, as soon as they were published and well circulated, created as much of a stir as did the finished buildings. In the Salon d'Automne of 1922, Le Corbusier exhibited two projects that expressed his idea of social environment and contained the germ of all the works of his first period. The Citrohan House, model for an industrially produced residential unit, displays the five characteristics by which the architect five years later defined his conception of what was modern in architecture: pillars supporting the structure, thus freeing the ground beneath the building; a roof terrace, transformable into a garden and an essential part of the house; an open floor plan; a facade free of ornamentation; and windows in strips that affirm the independence

The Citrohan House

of the structural frame. The interior provides the typical spatial contrast between open, split-level living space and the cell-like bedrooms. Accompanying the model was a diorama of a city of 3,000,000 inhabitants that illustrated ahead of its time the concept of green parks and gardens at the foot of a cluster of skyscrapers that house business offices. Apartment houses were placed in a rectangle around and outside the centrally located skyscrapers.

The ideas for city planning set forth at the Salon d'Automne, an annual semi-official exhibition, were taken up again and developed in 1925 at the Exposition des Arts Décoratifs in Paris, in a pavilion that was to be a "manifesto of the *esprit nouveau*." In this little duplex-flat, the interior walls violently coloured under the influence of the painter Fernand Léger, Le Corbusier exhibited his first collection of industrially produced furniture. But, foremost was the Voisin Plan of Paris, a project patronized by the automobile manufacturer Voisin. The plan proposed the complete demolition of deteriorated neighbourhoods, the replacement of them by parks, and the reorganization of businesses and administrative offices into two groups of cross-shaped skyscrapers, 250 metres (820 feet) tall. These were to be connected to each other and to residential buildings by means of express highways, consecrating, as it were, the role of the automobile in an industrial society.

This plan aroused the indignation of Parisians; it was, however, the basis of several more plans for vertical cities that eliminated conventional streets and located various urban functions in skyscrapers. Among the applications of the "radical city" principle are a series of seven plans for Algiers drawn up between 1931 and 1942. This is basically a project for a "viaduct" city in which the business centre is linked to the residential area by an expressway passing over the roofs of the apartment buildings.

During these years, in fact, Le Corbusier's social ideals were realized only on two occasions. One of these was in 1925 and 1926 when, thanks to the financial support of an industrialist, Henri Frugès, he built at Pessac, near Bordeaux, a workers' city of 40 houses in the style of the Citrohan House; the scorn for local tradition and the unconventional use of colour provoked hostility on the part of municipal authorities, who refused to provide a public water supply. Pessac was thus deprived of inhabitants for six years, and Le Corbusier did not forget this affront. In 1927 the architect, together with other architects, Gropius, Mies van der Rohe, Oud, Mart Stam, and others, participated in the international exposition of the Deutscher Werkbund, an association of various groups from art and industry concerned with producing functional objects of high aesthetic value. For this exposition Le Corbusier constructed two houses in the experimental residential quarter of Weissenhof at Stuttgart.

City at Pessac

Although Le Corbusier was from the beginning most interested in building for large numbers of people, during the prewar period he built primarily for privileged individuals, friends or patrons who commissioned individual houses. These houses are conceived as "machines for living in." They were functional in design and ascetic in appearance, incorporating rigorous geometric forms and bare facades. The first was for Ozenfant in 1922, followed by, among others: the house of the Swiss collector Raoul La Roche (1923), which later became the quarters of the Le Corbusier Foundation in Paris (1968); the villa (1927) of Michael Stein, a brother of the expatriate American writer and patron of Fauvism and Cubism Gertrude Stein; the Savoye House (1929–30), at Poissy, set in a lush, rural landscape on slender concrete pillars and disposed around an interior ramp structure connecting its three floors like a staircase.

The Geneva competition and first widespread public notice. Le Corbusier's ambition, however, was not limited to domestic architecture. In 1927 he participated in the competition set by the League of Nations for the design of its new centre in Geneva. His project, with its wall of insulating and heating glass, is one of the finest examples of the architect's gift for functional analysis. For the first time anywhere, he proposed an office building for a political organization that was not a Neoclassical temple but corresponded in its structure and design to a strict analysis of function. This plan was to become the prototype of all future United Nations buildings. It probably would have shared a first prize but was eliminated on the grounds of not having been drawn up in India ink as the rules of the competition specified. After the disappointment of Pessac, this disqualification, which was almost certainly the result of a conspiracy on the part of conservative members of the jury, further embittered Le Corbusier in his attitude toward official architectural circles. The scandal accompanying the elimination of his design, however, gave him needed publicity by identifying him with modern avant-garde architecture. An immediate consequence of the Geneva affair was the creation, in La Sarraz, Switzerland, in 1928, of the International Congresses of Modern Architecture (CIAM), intended at first to defend the avant-garde architectural values defeated in Geneva. By 1930 the organization had become oriented toward city planning theory. Le Corbusier, as secretary of the French section, played a very influential role in the five prewar congresses and especially in the fourth which issued in 1933 a declaration that elaborated some of the basic principles of modern architecture. It was later published simultaneously by both Le Corbusier (under the title *La Charte d'Athens* in 1943) and the industrial architect José Luis Sert, each architect providing his own commentary.

Design for the League of Nations centre

The publicity from the Geneva competition also made possible for Le Corbusier a lecture tour in South America in 1929 that was the source for his *Précisions sur un état présent de l'architecture et de l'urbanisme* (1930; "Reflections on the Present State of Architecture and Urbanism") and a trip in 1928 to Moscow, where he was able to make contact with avant-garde constructivist architects and won the competition for the Centrosoyus building (1929–35), originally meant to house the employees of the Central Office of the Cooperative (food distribution) Organizations of the U.S.S.R. This building (today housing The Gosplan administration) followed the design of the Geneva project in its segregation of functions and in its direct linking of the various buildings.

Le Corbusier constructed two other important buildings during this period, the Salvation Army Hostel in Paris, with its attempt at a "breathing" glass wall conceived as an unopenable glass surface equipped with an air conditioning system (a technological and financial failure), and the Swiss Dormitory at the Cité Universitaire in Paris (1931–32). In the latter structure Le Corbusier once again set apart the dormitory area, consisting of three floors of cells, from the common services areas located in a separate building. The two segments were connected by a stairway tower. Surfaces were left largely unfinished, and, for the first time, the massive pillars took on a sculptural value that set them in counterpoint to the curve of the stairway. At this point Le Corbusier's rational functionalism began to be balanced by a desire for expression.

Activities of the late 1930s. The end of the 1930s saw such especially famous projects as the masterplans for Algiers (1938–42) and Buenos Aires (1938); the building for the Ministry of Education and Health in Rio de Janeiro (1936); and an infinitely expandable museum for Philippeville (1938), in French North Africa. There was also a trip to the United States (1935), where Le Corbusier was already famous. For Le Corbusier, the Americans were a "timid" nation unwilling to push to its logical consequences the "conquest of height," the only "solution to the essential problems posed by city planning."

Le Corbusier's diverse activities corresponded to a chosen life-style. He was not a teacher, like his colleague Gropius, but the boss, who shut himself up alone in his office while his collaborators, who had come from all over the world and some of whom (José Luis Sert, Kunio Maekawa, Gyorgy Kepes, Affonso Reidy) would later become famous, worked outside in the long hall that served as a studio. Le Corbusier came to his office only in the

afternoons. His break with Ozenfant, in 1925, had not interrupted his painting career, and he usually spent his mornings painting at home. He was, by the mid-1930s, marked by the influence of Fernand Léger, who remained one of his few good friends. From 1934 on, Le Corbusier lived on the top floor of a solid but unpretentious apartment block he had built in Boulogne, where he sometimes organized expositions of primitive and naive art and of his own paintings on the bare stone walls of his studio. He led a simple and secluded life there with Yvonne Gallis, a former model whom he had married in 1930, the year he acquired French citizenship. Mme Le Corbusier, a southerner of unpretentious origins, had an exuberant and pugnacious nature that contrasted with the rather awkward reserve of Le Corbusier; she was present at his side until her death, bringing him a sense of balance and humour. It is through her that a woman first appears in Le Corbusier's paintings, in 1930, and she is the inspiration for the female figure often found in Le Corbusier's sketches.

World War II and the German occupation of France marked an interruption for Le Corbusier of his activity as a builder and a traveller, and of his 20-year association with Pierre Jeanneret, who, unlike Le Corbusier, had joined the French Resistance. Although he was prepared to work with the Vichy government, there was little building being done at the time in France, and his only activities were painting, writing, and reflection. His thoughts during this time led to the elaboration of the first bases of the "Modulor" concept, a scale of harmonic measures that set architectural elements in proportion to human stature. This theory was finally perfected in 1950, and Le Corbusier used it in designing all his subsequent buildings, wishing them to incorporate "on a human scale." By the time the war ended, Le Corbusier had welded the attacks launched against him by representatives of traditional architecture into a myth. He had become, for the public, the Picasso of architecture, and, for architecture students, the symbol of modernity.

The Modulor theory

Postwar city and residential plans. When the war ended, Le Corbusier thought that he would finally be able to apply his theories of planning in the reconstruction of France. He prepared in 1945 two plans for the cities of Saint Dié and La Pallice-Rochelle. At Saint Dié, in the Vosges Mountains, he proposed regrouping the 30,000 inhabitants of the destroyed town into five functional skyscrapers. Despite the support of two cabinet ministers, these plans were rejected by local authorities.

The plans for these towns, subsequently circulated throughout the world, became doctrine. But Le Corbusier was bitter at not being able to participate in the reconstruction of his adopted country. His bitterness increased when he was named a member of the jury of architects for the construction of the United Nations building in New York instead of being asked to design it himself. This feeling is revealed in the edition of Le Corbusier's complete works, published under his direction in 1946. He creates the image of a solitary and persecuted genius, going so far as to accuse the architect of the United Nations building of plagiarizing his own designs.

At last, in 1945, thanks to the unlimited support of the French government, Le Corbusier had the opportunity to construct a large (private) housing complex; he was commissioned to build, in Marseille, a residential complex that embodied his vision of a social environment.

The Marseille project

The Marseille project (*unité d'habitation*) is a vertical community of 18 floors. The 1,800 inhabitants are housed in 23 types of duplex (*i.e.*, split-level) apartments. Common services include two "streets" inside the building, with shops, a school, a hotel, and, on the roof, a nursery, a kindergarten, a gymnasium, and an open-air theatre. The apartments are conceived as individual "villas" stacked in the concrete frame like bottles in a rack. The facade is composed of balconies interspersed with prefabricated concrete sunshades (*brises-soleil*), an element invented by Le Corbusier in 1933.

The Marseille *unité d'habitation* was completed in 1952 at considerable cost. From the time construction began, it was the object of controversy; it was called "the Marseille folly" by critics, and the city's inhabitants had baptized it "la Maison du Fada," ("the house of the mad architect"). Nevertheless, two more *unités* were built at other locations in France, at Nantes and Briey, as well as others in West Berlin. As art, the Marseille complex resembles a sculpture; its massive forms, unfinished materials, and painted decoration in violent colour relate it to the expressive qualities of archaic art. The Marseille *unité d'habitation* characterizes the bold expressionist direction of the second period of Le Corbusier's work.

Diversified projects. Two religious buildings in France carry this tendency further. Both were commissioned as a result of the influence of the Dominican father Reverend Couturier, creator of the review *L'Art Sacré*. The more lyrical of the two, the chapel Notre-Dame-du-Haut at Ronchamp (1950–55), sacrifices Le Corbusier's famous principles of apparent functionalism; the wall has been built to a double thickness for visual effect and the roof, which only appears to be suspended, actually rests on a forest of supports. More brutal and austere is the convent of Sainte-Marie-de-la-Tourette at Eveux-sur-Arbresle, near Lyons. The square building imposes a fortress of concrete in a natural setting. In the three-tiered facade of glass at la Tourette, Le Corbusier first employed panes of glass set at "musical" intervals (calculated by the Greek engineer and composer Yannis Xenakis) to obtain a lyrical effect.

Chapel at Ronchamp

Ronchamp and la Tourette established once and for all Le Corbusier's reputation in France. Two large expositions of his work were organized in Paris in 1953 and in 1962. He was invited in 1965 to design the civic centre of the town of Firminy-Vert. The minister of cultural affairs, André Malraux, ordered plans for the French embassy in Brasília and commissioned a Museum of the Twentieth Century neither of which was built.

Meanwhile, in 1952, the architect constructed for his own use, in Cap Martin, near Nice, a tiny cabin consisting of a single room containing two beds, a work table and the necessary plumbing installations, which he made the symbol of his life-style and where he liked to go to get away from the city. Le Corbusier continued to paint, and, from 1946 on, sculpted. He participated in the interior decoration of all his buildings and some tapestries have also been made according to designs.

At his studio there were frequent changes in personnel, often the result of Le Corbusier's shifting moods. Between 1948 and 1960 Xenakis especially left the mark of his personality on the productions of the workshop. In 1958, for the World's Fair at Brussels, Xenakis completed a design for an audio-visual spectacle involving music at the Philips Pavilion, which Le Corbusier designed.

International activity. Only from 1950 on did Le Corbusier become active on a large scale outside of France. He built the National Museum of Western Art in Tokyo (1960), the Carpenter Visual Art Center at Harvard University (1964), and designed an Exposition Pavilion in Zürich that was constructed posthumously (1964).

New capital of the Punjab

In 1951, the government of the Punjab named him architectural advisor for the construction of its new capital, Chandīgarh. For the first time in his life, Le Corbusier was able to apply his principles of city planning on a metropolitan scale. He assigned the completion of the residential sectors to Pierre Jeanneret, whom he had asked to join him, and to the English architects Jane Drew and Maxwell Fry. He reserved for himself the Capitol, or administrative complex, the axial plan of which was inspired by examples of 17th-century urban design in the French classical manner such as Versailles. Totally without reference to local tradition, he designed the Palace of Justice, the Secretariat, and the Palace of the Assembly. Unfinished concrete, with windows sheltered by enormous *brises-soleil*, the sculptural facades, swooping rooflines, and monumental ramps are principal elements of his architecture, which immediately influenced architects all over the world.

Le Corbusier was not greatly impressed by his late recognition. He never forgave French authorities for having kept him out of the main city-planning projects of the 1960s. His last years were also saddened by the death of

his wife, in 1957. Nevertheless, he continued to conceive new projects until the end of his life: an art centre for Frankfurt (1963), the Olivetti computer centre in Milan (1963), the Palais des Congrès (for the European parliament) in Strasbourg (1964), and the French embassy in Brasília (1964).

Le Corbusier died suddenly on August 27, 1965, while swimming at Cap Martin. The man who had thought himself so misunderstood in his own time was given a national funeral, and in 1968 the Le Corbusier Foundation was created.

MAJOR WORKS

House at Vaucresson near Paris (1922); Ozenfant House, Paris (1922); La Roche and Jeanneret houses, Paris (1923); Pavillon de l'Esprit Nouveau, Paris (1925); Les Terrasses, Garches, near Paris (1927); Savoye House, Poissy (1929–30); Swiss Dormitory, Cité Universitaire, Paris (1931–32); Unité d'habitation, Marseilles (1946–52); Notre-Dame-du-Haut, Ronchamp, France (1950–55); Villa Shodan, Ahmādabād, India (1955–56); Convent of Sainte-Marie de la Tourette, Eveux-sur-L'Arbresle, France (1957–61); Chandīgarh, India state capital of the Punjab (1950s; Palace of Justice, 1952–56, Secretariat, 1952–56); Carpenter Visual Art Center, Harvard University (1964).

PROJECTS: Domino housing project (not executed, 1914–15); Citrohan House projects (1919–22); A contemporary city for 3,000,000 inhabitants (1922); Voisin Plan of Paris (1925); Project of the Palais des Nations (1927–28); Plan for Algiers (1930–42); Plan for Saint-Dié (1945–46).

WRITINGS: *Vers une architecture* (1923; *Toward a New Architecture*, 1946); *L'Art décoratif d'aujourd'hui* (1924); *Urbanisme* (1925; *The City of Tomorrow*, 1929); *La Ville radieure* (1935); *Quand les cathédrales étaient blanches* (1937; *When the Cathedrals Were White*, 1947); *La Charte d'Athènes* (1943); *Les Trois Établissements humains* (1945); *Protos d'urbanisme* (1946); *Le Modulor I* (1948; *The Modulor*, 1954).

BIBLIOGRAPHY. There is no complete bibliography of Le Corbusier; however, the Le Corbusier Foundation of Paris, which houses the Le Corbusier archives, is now preparing one. The principal sources on the man and his work are the *Oeuvre complète* published as follows with the assistance of the architect: W. BOESIGER, *Le Corbusier et Pierre Jeanneret 1910–1929* (1929); M. BILL, *Le Corbusier et Pierre Jeanneret 1929–1936* (1936), *Le Corbusier et Pierre Jeanneret 1936–1938* (1939); W. BOESIGER, *Le Corbusier Oeuvre complète 1938–1946* (1946), *Le Corbusier Oeuvre complète 1946–1952* (1953), *Le Corbusier Oeuvre complète 1952–1957* (1958), *Le Corbusier et son atelier rue de Sèvres 35, oeuvre complète 1957–1965* (1965); and the various theoretical works of Le Corbusier.

Monographs and articles devoted to Le Corbusier during his lifetime are generally distorted by a strong bias, either for or against the architect: F. DE PIERREFEU, *Le Corbusier et Pierre Jeanneret* (1932), the first book devoted to Le Corbusier; M. GAUTHIER, *Le Corbusier; ou, l'architecture au service de l'homme* (1944); S. PAPADAKIS (ed.), *Le Corbusier: Architect, Painter, Writer*, (1948), a series of essays by friends or former collaborators; F. CHOAY, *Le Corbusier* (1960), an attempt at a synthesis; and P. BLAKE, *Le Corbusier* (1964). In the works that have appeared since Le Corbusier's death: STANISLAUS VON MOOS, *Le Corbusier, Elemente einer Synthese* (1968; French trans., *Le Corbusier, l'architecte et son mythe*, 1971), a still very timid attempt to decipher Le Corbusier behind his myth; MAURICE BESSET, *Qui était Le Corbusier?* (1968; Eng. trans., *Who Was Le Corbusier?*, 1969), written by Le Corbusier's executor.

(F.C.)

Córdoba

Córdoba (Cordova) is the capital of Córdoba province in the Andalusia region of southern Spain; it is located on the Guadalquivir River, 82 miles northeast of Seville.

Early history (to 8th century AD). Córdoba was probably Carthaginian in origin, identified by some as the biblical city of Tarshish. Roman Corduba was one of the four judicial centres of the Roman province of Baetica. According to the Greek historians Polybius (*c.* 200–118 BC) and Strabo (1st century BC), it was established as a Roman colony in 152 BC by Marcus Claudius Marcellus, who made it his winter quarters during a Spanish campaign. The followers of Pompey the Great used it as their base during their civil war against Julius Caesar, who sacked the city in 45 BC and slaughtered 20,000 of its inhabitants. At the end of the Republic (late 1st century BC), it received the title of *colonia patricia* and became a flourishing commercial and cultural centre under the empire.

The Roman city

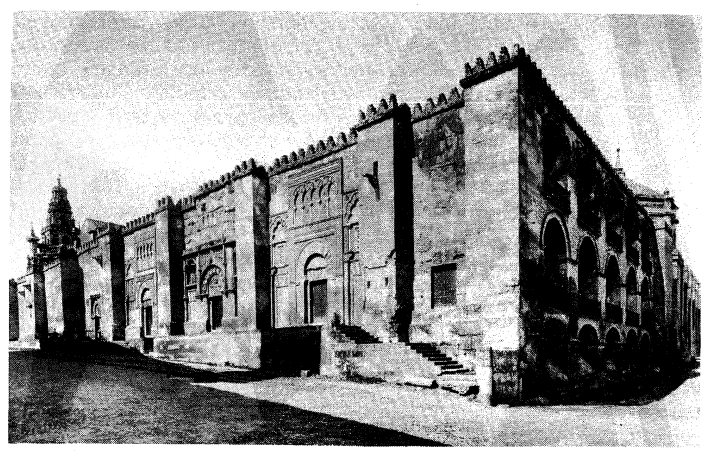
Geography contributed to Córdoba's prosperity. It stood in the middle of a rich agricultural district and served nearby mines. The Guadalquivir was then navigable up to the city itself, and Córdoba was connected by the Via Augusta to the cities of northern Spain and by other roads to Gades, Malaca, and Carthago Nova. Córdoba was the birthplace of both Senecas, the poet Lucan, and of Bishop Hosius, who influenced the conversion of the emperor Constantine. Though too little remains of Roman Córdoba to place all its main buildings with certainty, it is known that the Great Mosque was sited on the former temple to Janus.

Córdoba declined in importance under the Visigoths (ruled 6th–early 8th centuries), whose main centres of settlement were on the Meseta; but it was the centre of political revolt against King Agila (ruled 549–554) and of religious war (*c.* 570) in the struggle between Arians and Catholics.

Under Muslim rule (756–1236). Captured and largely destroyed by the Muslims in 711, Córdoba's recovery was impeded by tribal rivalries until ʿAbd ar-Raḥmān I, a member of the Umayyad family, accepted the leadership of the Spanish Muslims and made Córdoba his capital in 756. He founded the Great Mosque, which was enlarged by his successors and completed about 976 by Abū ʿĀmir al-Manṣūr. Though troubled by occasional revolt, Córdoba grew rapidly under Umayyad rule; and after ʿAbd ar-Raḥmān III proclaimed himself caliph of the West in 929, it became the largest and probably the most cultured city in Europe. Arab historians claim it covered an area 24 miles by 6; contained a population of 1,000,000; and had 260,000 buildings, including 80,000 shops, 3,000 mosques, baths, palaces, and a principal library of 400,000 volumes—figures that others consider exaggerated.

The Umayyads

Archivo Mas, Barcelona, Spain



Western facade of the Great Mosque at Córdoba.

As capital of the western caliphate, Córdoba was famous for its industries, architecture, art and literature, scholarship, and urbanity. Its woven silks and elaborate brocades, leather work, and jewelry were prized throughout Europe and the East, and its women copyists rivalled Christian monks in the production of religious works. The Umayyads blended Roman and Byzantine designs in their buildings and ransacked the Mediterranean for the remains of earlier civilizations to incorporate into their palaces. The most celebrated of these, ʿAbd ar-Raḥmān's palace city of Madīnat az-Zaḥrāʾ, largely destroyed in 1013, is said to have been roofed with silver and gold and to have had fountains of quicksilver. Education was actively promoted by the Umayyads and their successors the ʿĀmirids, and Córdoba became a centre for foreign scholars from East and West, noted for its expertise in Qurʾānic interpretation, Islāmic law, philology, music, medicine, and surgery. The refined manners of Baghdad, capital of the ʿAbbāsid caliphate, were popularized in

Córdoba by the singer Ziryāb (died 852), who established himself in Córdoba with his troupe and founded the Andalusian school of music and singing. Initially a certain tolerance was observed toward Córdoba's Christians, who were allowed to retain some churches and their own schools and libraries; this was curtailed by the later Umayyads who, under the pressure of the Christian reconquest, banned Latin and forced Christian children to attend Arabic schools. Arabic was the official language, but bilingualism in Arabic and a Hispano-Roman dialect was usual among the educated classes of both religions.

The ʿAbbādids, Almoravids, and Almohads

When the caliphate was dismembered by civil war early in the 11th century, Córdoba became the centre for a contest of power between the petty Muslim kingdoms (*taifas*) of Spain. It was severely sacked in 1013 and some of its most famous buildings were destroyed. After a brief interlude under the rule of the emir of Toledo, Córdoba was added to the domains of the ʿAbbādids of Seville during the reign of the poet-sultan al-Muʿtamid (ruled 1069–91). The successive occupations of the city by two North African sects—the al-Murābiṭūn (Almoravids) in 1091 and the al-Muwaḥḥidūn (Almohads) in 1172—restored internal stability, though they initiated a period of intense religious and artistic intolerance. Nevertheless, though Córdoba lost much of its importance as a political and military centre, poetry and scholarship continued to flourish there, and the city produced Ibn Ḥazm (994–1064), a profound jurist, theologian, and philologist known as the author of *The Ring of the Dove*. The philosophers Averroës (1126–98) and Maimonides (1135–1204) were both born and educated in Córdoba, though the former was expelled as a heretic and the latter fled to escape persecution by Almohad fanatics.

On the death of the last Almohad caliph in 1223, Córdoba again became a weak *taifa* state and fell to the Castillian king Ferdinand III in 1236.

Under Christian domination

History since 1236. Isolated from the sources of raw silk in the Muslim kingdom of Granada, the city lost its silk manufactures though not its leather industries. After the Christian reconquest (mid-13th century) Córdoba lost its political and cultural leadership, but the city remained important for more than two centuries as the main military base in the frontier warfare against Granada; many Castillian noble families and several distinguished soldiers settled there. The fall of Granada in 1492 left Córdoba a quiet provincial city of churches, monasteries, and aristocratic houses, renowned for its fine wines, horses, and women. The exotic poetry of Luis de Góngora y Argote (1561–1627) briefly revived Córdoba's poetic prestige in the 17th century.

The city was stormed and sacked by the French in 1808 for its part in fomenting the movement of independence. It was one of the first cities occupied by Francoist forces in the Spanish Civil War (1936–39).

Córdoba's Moorish character, its fine buildings and churches—especially the Great Mosque—have made it a popular tourist attraction. The city is also noted for its textile manufactures and for its brewing and distilling industries. In the late 1960s, its population was about 225,000.

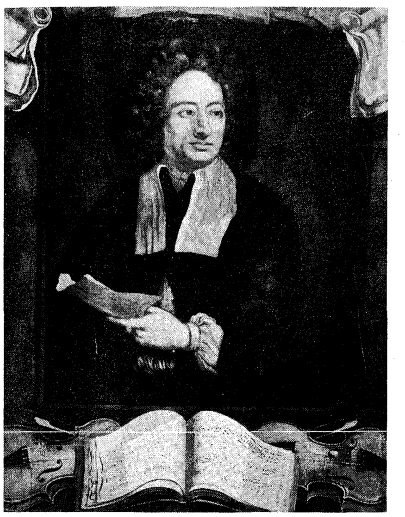
BIBLIOGRAPHY. J. DE LA TORRE, "Hallazgos arqueológicos junto a Córdoba," *Bol. de la Real Acad. de Hist.*, 79:419–427 (1921), is a useful introduction to Roman Córdoba. AL-MAKKARI, *The History of the Mohammedan Dynasties in Spain . . .*, trans. by P. DE GAYANGOS, 2 vol. (1840–43), is the best primary source of information for Córdoba in the Mohammedan period; W. MONTGOMERY WATT, *A History of Islamic Spain* (1965), provides a useful and contemporary summary. For Córdoba in the Christian period up to the mid-16th century, the best primary source of information is AMBROSIO DE MORALES, *Historia general y antigüedades de Córdoba*, 3 vol. (1574–1586). More modern impressions of the city, and a summary of its history, may be found in JAMES A. MICHENER, *Iberia*, pp. 155–180 (1968).

(K.Ga.)

Corelli, Arcangelo

One of the major violinist-composers of the late 17th century, Arcangelo Corelli was the founder of a style of violin playing and also of the concerto grosso—a type of musical composition in which there is interplay between a smaller and larger group of strings. His solo violin sonatas also became models for later developments in this field. Corelli's most important works in these genres are the *Concerti grossi*, Opus 6, and his *12 Sonatas for Violin and Violone or Harpsichord*, Opus 5.

By courtesy of the National Gallery of Ireland, Dublin



Corelli, portrait by Hugh Howard (1675–1737). In the National Gallery of Ireland, Dublin.

Corelli was born in Fusignano, near Ravenna, Italy, on February 17, 1653. His mother, Santa Raffini, having been left a widow five weeks before his birth, named him after his deceased father, Arcangelo. There are no documented details on his first years of study. It is thought that his first teacher was the curate of San Savino, a village on the outskirts of Fusignano. Later, he went to Faenza and Lugo, where he received his first elements of musical theory. Between 1666 and 1667 he studied with Giovanni Benvenuti, violinist of the chapel of San Petronio in Bologna. Benvenuti taught him the first principles of the violin; and another violinist, Leonardo Brugnoli, furthered his education.

After a four-year stay in Bologna, Corelli went to Rome. Reliable evidence on his activities is lacking for the first five years, but it is likely that he played the violin at the Teatro Tordinona. Also, it is possible that in 1677 he made a trip to Germany, returning to Rome in 1680.

First composition

On June 3, 1677, he sent his first composition, a *Sonata for Violin and Lute*, to Count Fabrizio Laderchi of Faenza.

By February 3, 1675, he was already third violinist in the orchestra of the chapel of San Luigi dei Francesi, Rome, and by the following year he was second violinist. In 1681 his *12 Trio Sonatas for Two Violins and Cello, with Organ Basso Continuo*, Opus 1, dedicated to Queen Christina of Sweden, who had a residence in Rome, were published. The following year he took the post of first violinist in the San Luigi dei Francesi orchestra, a position he held until 1685, the year in which his *12 Chamber Trio Sonatas for Two Violins, Violone and Violoncello or Harpsichord*, Opus 2, were published.

From September 1687 until November 1690, Corelli was musical director at the Palazzo Pamphili, where he both performed in and conducted important musical events. Corelli was particularly skilled as a conductor and may be considered one of the pioneers of modern orchestral direction. He was frequently called upon to organize and conduct special musical performances. Perhaps the most outstanding of these was the one sponsored by Queen Christina of Sweden for the British ambassador, who had been sent to Rome by King James II of England to attend the coronation of Pope Innocent XII. For this entertainment, Corelli conducted an orchestra of 150 strings. In 1690 he directed the performance of the oratorio *Santa Beatrice d'Este* by Giovanni Lulier, called *del violino*, also with a large number of players

(33 violins, 10 violas, 17 cellos). The same year he entered the service of Cardinal Pietro Ottoboni, in which he spent the rest of his life.

Entry into service of Cardinal Ottoboni

In 1689 Corelli's *12 Church Trio Sonatas for Two Violins and Archlute, with Organ Basso Continuo*, Opus 3, dedicated to Francesco II, duke of Modena (he had been the Modenesi Count, 1689–90) was published; and in 1694 his *12 Chamber Trio Sonatas for Two Violins and Violone or Harpsichord*, Opus 4, intended for the academy of Cardinal Ottoboni, also appeared.

It is probable that Corelli also taught at the German Institute in Rome and certain that in 1700 he occupied the post of first violinist and conductor for the concerts of the Palazzo della Cancelleria. Also in 1700 his *12 Sonatas for Violin and Violone or Harpsichord*, Opus 5, dedicated to Sophia Charlotte of Brandenburg, was published.

In 1702 Corelli went to Naples, where he probably played in the presence of the king and performed a composition by the Italian composer Alessandro Scarlatti. There is no exact documentation for this, just as there is none for his contacts with George Frideric Handel, who was in Rome between 1707 and 1708. In 1706, together with the Italian composer Bernardo Pasquini and Scarlatti, he was received into the Accademia dell' Arcadia and conducted a concert for the occasion.

Corelli died on January 8, 1713, having already retired from an active artistic life. He did not live to see the publication of his Opus 6, consisting of 12 concerti grossi, which was published in Amsterdam the year following his death.

Corelli was reserved and withdrawn, absorbed in his ideal as a musician. He lived in isolation with the exception of a period during which he and his favourite pupil, Matteo Fornari, were inseparable. He is said to have shown no interest in women; he certainly never married. Although he was averse to making any sort of personal display, he had amassed in his lifetime a valuable collection of paintings and violins.

MAJOR WORKS

INSTRUMENTAL MUSIC: *Concerti grossi con duoi violini e violoncello di concertino obligati, e duoi altri violini, viola, e basso di concerto grosso ad arbitrio che si potranno radoppiare*, op. 6 (published posthumously 1714).

CHAMBER MUSIC: *XII Suonate a tre, due violini e violoncello, col basso per l'organo*, op. 1 (published 1681); *XII Suonate da camera a tre, due violini, violoncello e violone o cembalo*, op. 2 (1685); *XII Suonate a tre, due violini e arcileuto col basso per l'organo*, op. 3 (1689); *XII Suonate da camera a tre, due violini e violone o cembalo*, op. 4 (1694); *XII Suonate a violino e violone o cembalo*, op. 5 (1700).

BIBLIOGRAPHY

Contemporary accounts: G.M. CRESCIMBENI, *Notizie istoriche degli Arcadi morti* (1720), supplies information on Corelli's early studies. G.B. MARTINI, *Serie cronologica dei Principi dell 'Accademia dei Filarmonici di Bologna* (1776, in manuscript), gives the earliest information on Corelli's birth, studies, and teachers. SIR JOHN HAWKINS, "The General History and Peculiar Character of the Works of Archangelo Corelli," in *Universal Magazine of Knowledge and Pleasure*, pp. 171–172 (April 1777), is a critical examination of Corelli's works, with information on his meeting with Handel in Rome.

Critical works: ALBERTO CAMETTI, "Arcangelo Corelli à Saint-Louis-des-Français à Rome," in *Revue musicale*, pp. 25–28 (January 1922); MARIO RINALDI, *Arcangelo Corelli* (1953, in Italian), presents the author's opinion that Corelli travelled in both France and Germany; MARC PINCHERLE, *Corelli et son temps* (1954), rejects Rinaldi's opinion that Corelli made a journey to France—of fundamental importance for the study of the composer's personality.

(G.Pan.)

Corinth

Corinth (Greek Kórinthos) is the name of both an ancient and a modern city of the Peloponnesus, in Greece. The remains of the ancient city lie about fifty miles west of Athens, at the eastern end of the Gulf of Corinth, on a terrace some 300 feet above sea level, and at the foot of Acrocorinth—a Gibraltar-like eminence rising 1,886 feet above sea level. The site has been inhabited since Neolithic times, well before 3000 BC.

History. During the 8th century BC Corinth began to develop as a commercial and manufacturing centre, exploiting its geographical position, and producing fine pottery and bronze. Its political influence was increased through territorial expansion in the vicinity and through colonies, conspicuously, Corcyra and Syracuse. Toward 600 BC the controlling family, the Bacchiads, was overthrown by Cypselus, who became a "tyrant" and was succeeded by his even more famous son Periander. Colonial expansion was extended along the Adriatic and into Macedonia, and contacts were made with the Near East and Egypt.

After the Greco-Persian Wars Corinth joined the Peloponnesian faction led by Sparta against Athens, and it was often the commercial rivalry of Corinth and Athens that generated crises. The culmination of this rivalry in the Peloponnesian War brought about the military defeat of Athens but also the eclipse of real political power for Corinth. Subsequently, Corinth was involved in most of the contemporary conflicts, but after 338 it figured chiefly as a pawn in the struggles of greater powers because of the strategic value of its citadel.

The mingling of these events with the advancing fortunes of Rome led to the destruction of Corinth in 146 by the Roman general Mummius; but in 44 BC Julius Caesar re-established the city as a colony of Rome (a base for Roman government in Greece), and once again it became one of the great cities of the Mediterranean. It flourished as a provincial centre under the Roman Empire and is known to readers of the New Testament for the letters addressed to its Christian community by the Apostle Paul. It enjoyed some prosperity under Byzantine rule but declined in the later Middle Ages. After the Turkish conquest in 1458, it was reduced to a country town.

Excavation and remains. The site has been explored by the American School of Classical Studies at Athens, in excavations begun in 1896 and continued since with brief interruptions.

The central area

The central area of the city, as known from the excavations, appears to have developed from the beginning around a low hill, on which stand seven columns of a temple built in the archaic style shortly after 550 BC. Below this to the east is a valley opening onto the lower natural terrace to the north, through which a road led to Lechaeum, the harbour on the Corinthian Gulf. Southeast of the temple, facing on this valley, there was a spring called Peirene's, which was improved as a fountain house in the 6th century and frequently enlarged and embellished in succeeding centuries. Just south of the temple another smaller spring in a branch of the valley came to have a central function in a small sanctuary not yet fully explored; a small apsidal temple associated with it seems to have been equipped for oracular or other cult performances. A little west of the archaic temple was a third archaic fountain called Glauce's. Probably as early as the 5th century BC, a peculiar stoa (columned portico) was built below the archaic temple facing the road to Lechaeum, and industrial establishments developed on the eastern side of the road. South of the temple was a race track laid out in what must have been an extensive, rather open area. Below the temple to the north a bathing establishment and a small stoa were built—perhaps part of a small gymnasium—and northwest of this a large theatre.

By the end of the 4th century there was an enormous stoa over 500 feet long, which might have served as a kind of hotel, defining the area south of the temple as a major commercial marketplace, or agora. Another smaller stoa immediately below the temple soon gave clearer definition to the north edge of this area, while the enclosure for a large temple precinct marked the western extremity.

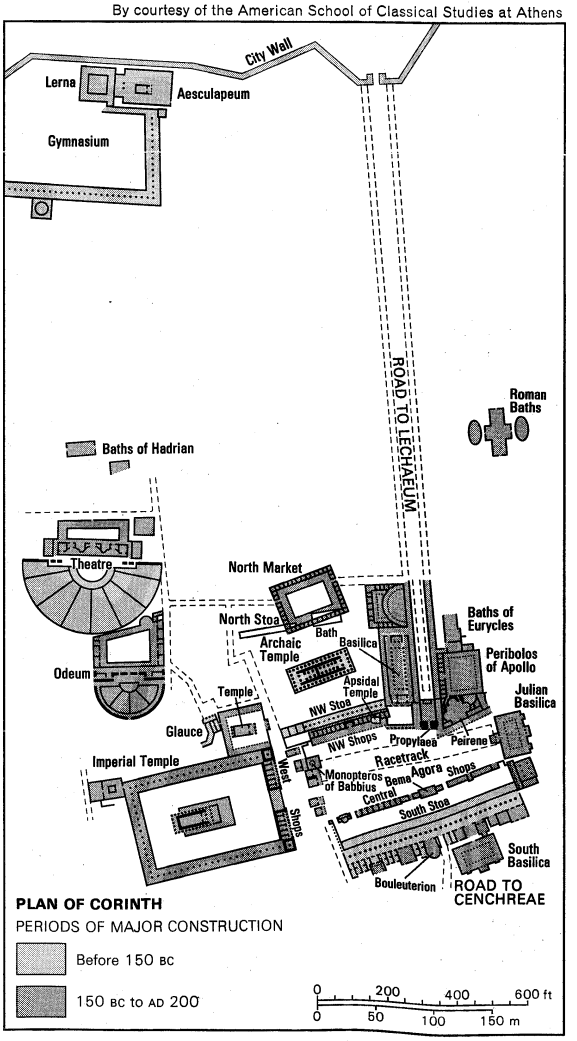
The general lines thus established by Hellenistic times were elaborated in detail during the first two centuries after Christ. The south stoa was gradually converted into a series of administrative offices; the open space of the agora was subdivided longitudinally by terracing and a



Ruins in the central area of Corinth. Left are the seven columns of the Temple of Apollo, *c.* 550 BC.
J. Allan Cash—Rapho Guillumette

row of shops with a *bēma* or "rostrum" in the middle; a basilica and a similar structure were built at the east end, and a row of temples emerged at the west. The entrance to the road to Lechaeum was adorned by a triumphal arch, and a large basilica and other buildings—religious and commercial—were built along the road itself. Between the fountain of Glauce and the theatre was built an odeum, or music hall, while the theatre itself was converted from its Greek form to accommodate Roman theatrical and gladiatorial spectacles.

By courtesy of the American School of Classical Studies at Athens



Central area of ancient Corinth

The surrounding parts of the city are known only in random spots. North of the theatre was a large gymnasium, partially excavated; just north of this was a temple and sanctuary of Asclepius, and further the city wall, which is about 600 yards north of the archaic temple. In connection with the sanctuary and gymnasium were elaborate arrangements of springs and cisterns—those near the Asclepeum possibly medicinal, those below the gymnasium perhaps for ordinary bathing. About 1,500 yards west of the archaic temple by the city wall were numerous early Greek pottery factories; almost as far to the east is a Roman amphitheatre of considerable size, excavated into the rock. Ancient writers speak of a luxurious suburb and resort area that lay close to the circuit wall in this region, some 1,500 yards southeast of the agora. Outlying areas

Heavily fortified and almost impregnable to assault from archaic times through the 17th century, the citadel on Acrocorinth, with its evidence of remodellings, is almost a museum of military architecture. The city itself was protected in Greek times by a circuit wall almost six miles in extent, but little of this has been preserved.

Corinth was served by two harbours: Lechaeum, less than two miles to the north, and Cenchreae, about six miles to the east. At Lechaeum are unexplored remains of the harbour works and an early Christian basilica of the 5th century, the largest in Greece, which has been entirely cleared. At Cenchreae some remains of the ancient harbour and associated buildings have been explored or cleared, although some of the most significant of these are submerged and inaccessible to view.

New Corinth. Located three miles northeast of Old Corinth, New Corinth was founded in 1858 after an earthquake that destroyed its predecessor. It is primarily a hub of communications between northern and southern Greece and is the primary point of export for local fruit, raisins, and tobacco. It is the chief town of the *nomós* of Corinth and the seat of an archbishop. As highway conditions improve, it is losing some of its advantages of location, being just 57 miles from Athens and only skirted by the finest new roads.

BIBLIOGRAPHY. For a condensed comprehensive account, see *The Urban Development of Ancient Corinth*, by H.S. ROBINSON (1965). For detailed reports of the excavations, see AMERICAN SCHOOL OF CLASSICAL STUDIES AT ATHENS, *Corinth: Results of Excavations* (1929 *et seq.*), several volumes by various authors; and preliminary reports of work in progress in the *American Journal of Archaeology* (1885–), and *Hesperia* (1932–).

(R.L.S.)

Cornales

The dogwood order (Cornales) of flowering plants encompasses 10 families, approximately 350 genera, and over 3,700 species. By far the largest family is the Umbelliferae or parsley family, also called the family Apiaceae, with about 275 genera and 2,850 species, about 80 percent of the species in the entire order; it is practically cosmopolitan in distribution but occurs chiefly in the north temperate regions. By contrast, the second largest family, the Araliaceae, is principally tropical. The family Cornaceae is mainly north temperate; the seven remaining smaller families are from North America or eastern Asia, mostly in warm or wet regions.

GENERAL FEATURES

Size range and diversity of structure. As befits an order with representatives scattered through both tropical and temperate regions, the members of the Cornales show considerable diversity in form. Most of the families are predominantly woody, principally consisting of shrubs but also with a fair representation of trees and a certain number of climbers, such as the ivy. Only the Umbelliferae contains any large representation of herbaceous (nonwoody) species; even within this family, the variation in life-forms is remarkable. At one extreme are low-growing, creeping species, such as the well-known

marsh pennywort of Europe (*Hydrocotyle vulgaris*) and allied species in North America and elsewhere. At the other extreme are large tropical trees or bushes such as the African *Stegonotaenia araliacea*. Even in temperate regions gigantic species of the family Umbelliferae occur—notably the giant hogweed of the Caucasus (*Heracleum mantegazzianum*), which can cause painful blistering in sunlight by producing photosensitization of the skin when touched. Spine-bearing species are found in the genus *Eryngium*, some of which are grown as decorative plants because of their bright amethyst colour. No parasitic or saprophytic plants—those that obtain nourishment from living or dead organisms, respectively—occur in the order.

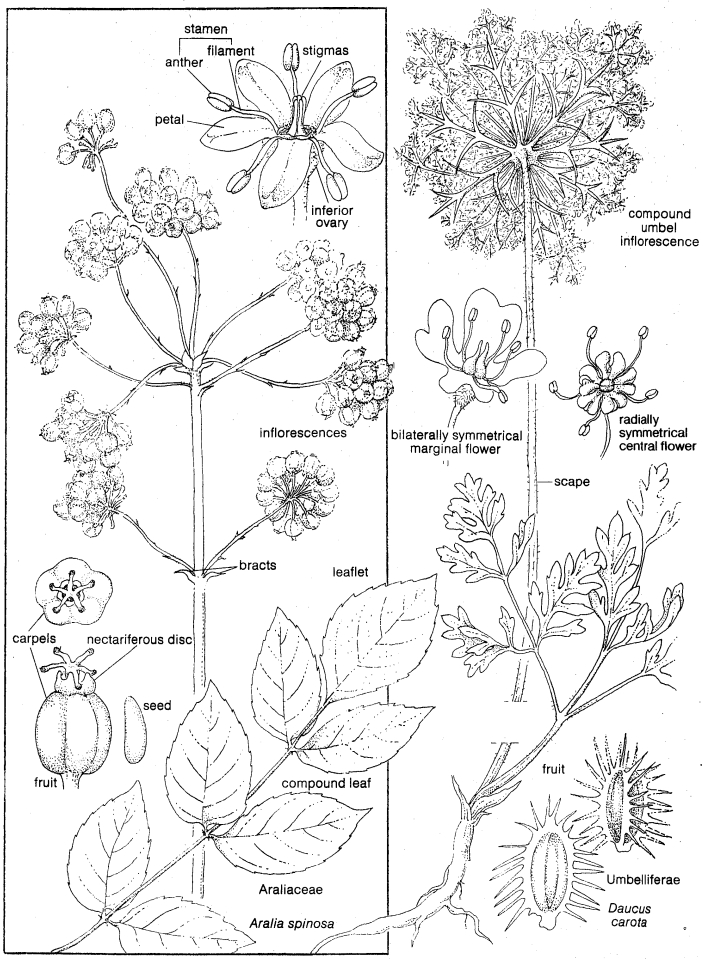


Figure 1: Representative plants from the two largest Cornales families.

Members of the order used as foods

Economic importance. Some members of the order are of widespread or local economic importance. Foodstuffs are important products of the family Umbelliferae, which yields such well-known vegetables as celery (*Apium graveolens*), carrots (*Daucus carota*), and parsnips (*Pastinaca sativa*), as well as the Andean root crop *Arracacia xanthorrhiza*. In various eastern Asiatic countries the young stems and leaves of several species of the Araliaceae are esteemed as vegetables. Plants of the family Umbelliferae also yield some well-known spices and flavourings, such as caraway seeds (*Carum carvi*). Some of these were used in biblical times; *e.g.*, coriander (*Coriandrum sativum*), aniseed (*Pimpinella anisum*), and cumin (*Cuminum cyminum*). The last species is also employed in perfume manufacture. A few medicinal plants also occur in the order. Dill (*Anethum graveolens*), gum ammoniac (*Dorema ammoniacum*), and asafetida (from several species of *Ferula*) are products of the family Umbelliferae. The bark of *Alangium salviifolium*, an emetic, has been used as an ipecac substitute, and that of one or two species of *Cornus* has been used as a quinine substitute. An interesting discovery has been that *Camptotheca acuminata* (family Nyssaceae) yields camptothecin, which is useful in the treatment of leukemia. Ginseng, *Panax schinseng* (family Araliaceae), one of the most famous Chinese drugs, is used to treat many diseases. North American ginseng is *P. quinquefolium*.

Many species of the Umbelliferae are poisonous, including the hemlock (*Conium maculatum*) used in ancient Greece as a means of capital punishment; hemlock water dropwort (*Oenanthe crocata*), which sometimes kills livestock; and cowbane (*Cicuta virosa*).

Wood products

Tetrapanax papyriferum (family Araliaceae) is the source of the so-called rice paper of commerce, used for, among other things, surgical dressings and artists' watercolour papers. The wood of some species of this family is used as timber. The wood of *Dendropanax arboreum* is strong and is used locally in general carpentry. The wood of *Didymopanax* species is less durable; that of *Kalopanax pictus* is very easily worked, pliable, and resonant. For this reason it has been used in China to manufacture drums. It is also suitable for furniture, in which form it has been called Japanese ash. Both *Cornus florida* (flowering dogwood) and *C. nuttallii* (western flowering dogwood) produce a close-grained wood, heavy and strong, that is suitable for turnery, wheel hubs, and similar products. The one species of the genus *Curtisia* (also of the family Cornaceae), *C. faginea*, the assegai, or Cape Lance tree, yields a valuable, red-brown durable timber that is close-grained, produces a good finish, and is valued for making furniture. Two species of *Griselinia* (*G. littoralis* and *G. lucida*) also produce heavy timber, but that of *G. littoralis* is rather brittle. The close-grained, heavy wood of *Alangium salviifolium* has been used to a limited extent in India, and some *Nyssa* species give tough but nondurable whitewoods.

Several species of *Cornus*, such as *Cornus mas* and *C. florida*, are grown as decorative shrubs because of their attractive, white petallike involucres, leaflike structures positioned just below the true flowers. *Cornus mas* also produces the cornelian cherry, which is eaten fresh or as preserves and is used in the manufacture of *vin de cornoulle*, a wine. *Aucuba japonica* is a popular garden evergreen. Several genera of the family Araliaceae are grown as decorative plants, including the common ivy (*Hedera helix*), of which there are also attractive variegated forms. One or two species of *Garrya*, notably *G. elliptica*, are grown as garden evergreens, which are attractive because of their glossy, dark-green foliage and reddish-purple flower clusters. *Nyssa sylvatica*, the sour gum tree, produces beautiful foliage tints in the fall, and *N. ogeche* yields the ogeechee lime fruits; its flowers are also a valuable source of commercial honey. One or two species of *Helwingia* are also occasionally seen in gardens.

NATURAL HISTORY

Seed dispersal. The Cornales is not among the more remarkable orders as far as seed dispersal is concerned, because most of the families produce drupes—stony-seeded fruits—or berries, and bird distribution is thus widespread. The fruit types, however, are of some interest. In some species of *Nyssa*, a genus mainly of swamp trees or shrubs, for example, the drupes may be distributed by water, but they do not germinate or establish seedlings while immersed. Seed distribution attains its greatest complexity with the family Umbelliferae. The characteristic fruit of this family is a dry compound fruit, known as a schizocarp, which splits down a septum, or wall, separating the adjacent segments of the fruit into two portions, called mericarps. These are at first anchored by a thin stalk (the carpophore), which usually divides to force the mericarps apart. In some genera the mericarps are strongly compressed and thin, with a broad, membranous wing for wind dispersal. In others, such as *Oenanthe*, the mericarps are frequently found in flood refuse and debris along riverbanks; a corky layer in the fruit presumably provides the buoyancy necessary for distribution by water. In still others the surface of the

mericarps is covered with hook- or grapple-shaped spines to facilitate distribution by attachment to animals. The efficiency of this last method is evidenced by the frequency with which an Australian species with grapple-shaped spines, *Daucus glochidiatus*, is found in European fields that have been fertilized with Australian wool refuse, or shoddy.

Pollination. Most families of the order produce species with a distinct nectar-secreting disk, and in correspondence with the exposed position of the nectar, pollination is usually accomplished by short-tongued insects, such as beetles, flies, wasps, and short-tongued bees. Higher bees visit plants that produce large quantities of concentrated nectar, such as are found in various genera of the family Umbelliferae.

Pollinating insects

Insects are attracted to members of the order by many-flowered open inflorescences (flower clusters) rather than by single, brightly coloured flowers. It has been suggested that the many tiny flowers of most species of the Umbelliferae, whose members have the greatest versatility of pollination mechanisms in the order, may appear to flicker to a passing insect. Some species of *Cornus* and certain genera of the family Umbelliferae, such as *Tordylium*, have much enlarged petals in the outer rows of flowers. The central flower of the umbel (a name given to the inflorescence; typical of the Umbelliferae) of the carrot (*Daucus carota*) is often purple. The southwest Asian *Artedia squamata* has enlarged outer petals coupled with a purple tuft of hairs in the centre of the umbel. All these variations presumably attract insects. In some inconspicuous species of the Umbelliferae, such as *Hydrocotyle*, self-fertilization apparently occurs. Self-fertilization also takes place in more conspicuous genera, particularly those with crowded heads, such as *Eryngium* and *Sanicula*, in which the elongating styles (part of the female reproductive organ) arch outward to bury themselves in the pollen of the anthers (part of the male reproductive organ); but additional insect pollination is more typical of these genera.

The distribution of the flower sexes in the inflorescences (or umbels) of the family Umbelliferae is very complex. In all cases, however, the purpose is to ensure cross-pollination by having the female organs develop in the first flowers to open and by having a corresponding reduction in the development of ones that open later. Wind pollination in the order Cornales occurs only in the genus *Garrya*.

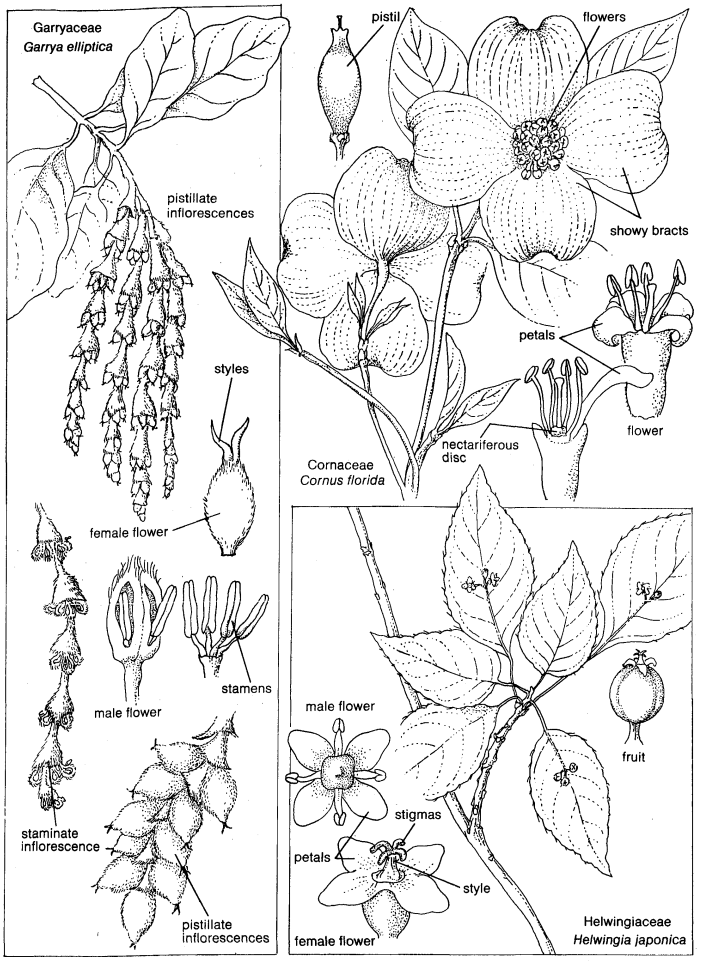


Figure 2: Representative plants from the central family (Cornaceae) and two of the smaller families of the order Cornales.
Drawing by M. Pahl

EVOLUTION AND PALEONTOLOGY

The fossil record. Fossil remains of woody families have been well investigated. *Nyssa* is one of the many genera generally considered to be remnants of a moist-climate forest belt that extended in the Tertiary Period (about 65,000,000 to 2,500,000 years ago) throughout almost all of the Northern Hemisphere. The distribution of *Nyssa* is characteristic of such persistent remnants of otherwise extinct floras—*i.e.*, it is centred on North America and eastern Asia. Paleobotanical evidence indicates that *Nyssa* apparently persisted in Europe into the early part of the Pleistocene Epoch (beginning about 2,500,000 years ago), when glaciation finally destroyed it, thus creating its present disjunct areas of occurrence. Tertiary remains are frequently found for the families Cornaceae (especially *Cornus*), Nyssaceae (*Nyssa*), Araliaceae, Alangiaceae, Umbelliferae, and Mastixiaceae. Records from the Cretaceous Period (136,000,000 to 65,000,000 years ago) are based upon fragmentary material and are taxonomically suspect. The Tertiary records range across the entire Northern Hemisphere from North America through Europe and Siberia to Japan and from all Tertiary divisions from the Paleocene Epoch (beginning 65,000,000 years ago) to the Pliocene (ending 2,500,000 years ago) and include fruits, pollen, wood, and leaves. The genus *Nyssa* has been particularly well investigated. Its fruits, readily recognized because they have a characteristic germination valve, sometimes occur in great quantities and show considerable variation in form, as in the brown-coal deposits at Brandon, Vermont, and in those of Europe and Japan. It has been found that the number of well-defined species of *Nyssa* was at least as great in the past as in recent times. The family Mastixiaceae, it appears, may have been considerably richer not only in species but also in genera—which may explain why the few members of this family today demonstrate such an interesting combination of features, some generally regarded as primitive and others considered more advanced.

Families represented in the fossil record

As might be expected of generally delicate plants, Tertiary remains of members of the Umbelliferae are more scanty, consisting mostly of fruits identified with such recent genera as *Oenanthe*, *Eryngium*, and *Chaerophyllum*.

Phylogeny. The order Cornales is considered to have arisen from the order Saxifragales, one of the more primitive orders of the postulated central line of descent from the order Magnoliales. Definite links with the Saxifragales are indicated by morphological resemblances between genera that represent the more primitive members of the Cornales and some genera belonging to the Saxifragales—notably in the families Hydrangeaceae and Escalloniaceae, such as *Broussaisia* and *Polyosma*. It has also been demonstrated that the wood anatomy of the family Cornaceae (in its widest sense; *i.e.*, including the Mastixiaceae, Toricelliaceae, and Helwingiaceae, here treated as separate families) closely resembles that of the wood of the family Hydrangeaceae of the Saxifragales—*e.g.*, that of *Philadelphus*. The chief difference between these two representatives of allied orders—the families Cornaceae and Hydrangeaceae—is in the parenchyma, a type of tissue composed of simple, thin-walled, undifferentiated cells. In both families the parenchyma tissue tends to be scanty or even absent, but, when well defined, it is paratracheal (*i.e.*, disposed around or close to the vessels) in the Hydrangeaceae and apotracheal (not asso-

ciated with the vessels) in all genera of the family Cornaceae with the exception of *Aucuba*.

There has been considerable phylogenetic speculation; yet, within certain limits, the authors of most general systems of classification have been in relatively good agreement. In addition to the order Saxifragales, the orders Sapindales, Rutales, Celastrales, and Rhamnales have received considerable support as close allies of the Cornales. All of them, including the Cornales, have progressively simplified flowers, the formation of an inferior or enclosed ovary, a tendency toward the production of flowers whose basic parts occur in sets of four, the formation of a nectar-producing disk surrounding the ovary, and a reduction in the number of ovules. The cumulative effect of this evidence is impressive, but it is probably also true that the many common characters indicate a common line of descent for all of the orders rather than lines of descent from one to the other; the most likely candidate as a common ancestor remains the order Saxifragales.

Related orders

Within the order, most workers agree that the Cornaceae is the basic and most primitive family and that the family Nyssaceae is derived from it. The remaining smaller families show distinct affinities with these two and, in some cases, also show links with the family Araliaceae.

Use of pollen in Cornales classification

As in many other plant orders, pollen morphology—form and structure of the pollen—has been of great assistance in the Cornales in helping to establish associations or divisions within the group. Pollen morphology has also provided evidence against unsatisfactory classification systems. The position of the family Alangiaceae, for example, has long been in doubt, and it has usually been compared with the Combretaceae. Evidence concerned with pollen is against this association, however. On the other hand, the pollen grains of *Alangium chinense* show some similarity to those of certain genera of the family Nyssaceae (*Nyssa*, *Camptotheca*) and of Cornaceae, thus associating a disputed link with an undisputed one (*i.e.*, that between Cornaceae and Nyssaceae). Evidence from pollen confirms the separation of *Davidia* as a family distinct from the Nyssaceae and *Toricellia* and *Mastixia* as genera not belonging to the Cornaceae. In the last two cases, however, the evidence is less conclusive, since the family Cornaceae is diverse with respect to pollen morphology. Further investigation along this and other lines of research may lead to the permanent isolation of additional families from the Cornaceae to accommodate such genera as *Aucuba* and *Kaliphora*. The distinctiveness of the family Garryaceae is no longer debated. The occurrence of pollen types of a more or less similar nature between Cornaceae and Garryaceae, and Garryaceae and certain Araliaceae provides a further confirmatory link within the order, as does the occurrence of common pollen types between the families Araliaceae and some members of the Umbelliferae.

With regard to relationships with other orders, pollen studies tend to confirm that the close association of the Cornales to the family Caprifoliaceae (order Dipsacales), specifically to the genera *Viburnum* and *Sambucus*, postulated by several authorities, may be well-founded. Below the family rank, further studies with pollen work will probably confirm other types of evidence that several natural sections occur within the genus *Cornus*.

CLASSIFICATION

Annotated classification.

ORDER CORNALES

Trees or shrubs, rarely herbs (except in the family Umbelliferae, then frequently herbs), rarely lianas. Leaves alternate or opposite, simple with entire (uncut, not toothed) margins or lobed in palmate or pinnate patterns (*i.e.*, the lobes radiate from a common point or from both sides of a central axis, respectively) with toothed segments. Stipules commonly absent or sometimes present and free (in the families Helwingiaceae and part of Umbelliferae). Inflorescences consist of cymes (flower clusters in which the central or topmost flower opens first and further flowers occur on lateral growing points), panicles (many-branched clusters), racemes (flower clusters with the flowers stalked and the lowest flowers opening first), or umbels (clusters radiating from a common point), more rarely spikes or heads or flowers solitary in the leaf axils (upper angles between leaf and stem) or on the leaf surface. Flowers usually small, radially symmetrical except sometimes bilaterally symmetrical in the enlarged marginal flowers in inflorescences of some species in the Umbelliferae; bisexual, polygamous (male, female, and bisexual flowers on one plant) or dioecious (male and female flowers on separate plants), commonly 4- or 5-merous (basic parts in 4s or 5s) but occasionally with up to 10 petals and sepals. Petals free, touching at the margins or overlapped, rarely absent. Sepals commonly much reduced, sometimes absent, the bases forming a tube fused around the ovary. Stamens in 1 or sometimes 2 or more series or whorls (very numerous in *Davidia*), equal in number to or up to 4 times as many as the petals. Nectar-producing disk present, attached to upper part of ovary. Styles 1 to several, free or fused together. Ovary inferior (enclosed within the basal parts of the sepals, petals, and stamens), of fused carpels. Ovules pendulous, usually solitary, and attached to the inner walls of the ovary. Fruit a berry, drupe (stone-seeded fruit), or a dry schizocarp (splitting fruit), generally formed of two segments (mericarps), which do not always separate at maturity. Embryo small to medium in size, endosperm copious to scanty. Ten families, about 350 genera, and more than 3,700 species with worldwide distribution, but chiefly found in north temperate regions.

Family Cornaceae (dogwood family)

Almost entirely trees and shrubs (creeping herbs only in the small genus *Chamaepericlymenum*). Leaves opposite or alternate, simple with entire margins, often furnished with stellate (branched in star patterns) or T-shaped hairs, stipules absent. Flowers small, unisexual or bisexual. Inflorescences in axillary or terminal umbels, corymbs, or in the form of dense heads with conspicuous involucres of leafy or petallike bracts. Petals and sepals 4 or 5, the sepals small or missing. Petals commonly touching at the margins (valvate) but occasionally overlapped (imbricate). Stamens equal in number to and alternate with petals. Ovary occasionally 1- but usually 2- to 4-locular (chambered), each locule containing 1 ovule. Fruits mostly drupaceous, but bony or crustaceous pyrenes (small, hard nutlets) also occur. Eight genera and 80 species. *Cynoxylon*, *Dendrobenthamia*, *Swida*, and *Afrocrania* are sometimes recognized as distinct from the genus *Cornus;* the number of genera is then correspondingly increased.

Family Garryaceae

Shrubs or small trees. Inflorescences catkin-like (pendulous elongate axes bearing numerous small flowers). Flowers dioecious; petals completely absent; sepals 4 or absent in the female flowers, very tiny, valvate. Ovary with 1 locule, 2 ovules. Stamens 4. Fruit a berry. Seeds with copious endosperm. One genus (*Garrya*) with 18 species, distributed only in the Northern Hemisphere of the New World.

Family Davidiaceae

Tree with deciduous, alternate, simple leaves without stipules. Flowers andromonoecious (*i.e.*, there are separate plants that contain male flowers only; others with bisexual flowers) crowded in terminal globose heads. Male flowers crowded together and each composed of 5 or 6 stamens only. Petals absent, sepals much reduced in size. Fertile flowers single in each head and placed asymmetrically, the ovary 6- to 10-locular, with a single columnar style bearing 6 to 9 stigmatic lobes. Fruit a drupe with a bony endocarp. Involucre of inflorescence of 2 very large, white, spreading bracts. One genus and species (*Davidia involucrata*), native to northern China but widely grown in gardens (commonly called the dove tree or handkerchief tree from the 2 large, white bracts).

Family Nyssaceae (Nyssa or gum family)

A somewhat nondescript group of trees or shrubs with simple, alternate, entire-margined or toothed leaves lacking stipules. Flowers very small, bisexual or more commonly unisexual. Male flowers in inflorescences varying from heads or spikes to racemes or umbels; female and bisexual flowers solitary or in few-flowered heads. Sepals 5 or more or absent. Petals usually 5 but occasionally from 4 to 8, imbricate; stamens 5 to 10 and typically arranged in 2 series if more than 5. Ovary constantly with 1 locule and 1 ovule. Style and stigma solitary. Seeds with scanty endosperm. Fruit a drupe in *Nyssa*, samara-like (*i.e.*, somewhat like the winged "keys" of maples) in *Camptotheca*. Two genera: *Nyssa*, with about 10 species in the Northern Hemisphere, and *Camptotheca*, with one species in China and Tibet.

Family Alangiaceae

Distinctive trees or shrubs with alternate, simple or lobed, often asymmetrical leaves lacking stipules. Flowers set on jointed pedicels (stalks) and disposed in axillary cymes. Sepals and petals 4 to 10, sepals generally more or less fused; petals generally free, or sometimes coherent below, very narrow and ribbonlike, valvate, frequently finally recurved, and

more or less hairy on the inner surface. Stamens as many as, and alternate with, the petals, or up to 4 times as many; also hairy within. Ovary 1- or 2-locular, with 1 locule, 1 ovule, and a single style. Fruits drupaceous and crowned by the persistent sepals and disk. One genus (*Alangium*) with about 20 species in the Old World tropics.

Family Toricelliaceae

Small trees with alternate, palmately lobed or toothed leaves, broadly sheathing at the base. Flowers white, dioecious, small; disposed in lax, pendulous thyrses (contracted, compact panicle-like clusters). Petals 5, induplicate-valvate (with inrolled edges that touch but do not overlap), with elongate, inflexed tips. Stamens generally 5. Ovary 4-locular. One genus (*Toricellia*) with 3 species in the Himalayas and China.

Family Mastixiaceae

Trees with entire, alternate or opposite, leaves lacking stipules. Flowers small, in terminal panicles. Petals 4 or 5, valvate, but inflexed at the tip and fringed or bidentate—notched, having 2 "teeth." Stamens 4 or 5. Ovary of 2 locules; ovules 2, epitropic. One genus (*Mastixia*) with about 25 species in Southeast Asia, east to the Solomon Islands.

Family Helwingiaceae

Shrubs with alternate or nearly opposite, simple, toothed leaves with stipules. Stipules frequently branched. Flowers dioecious (male and female flowers are on separate plants), borne in small umbels on the upper surface of the midribs of the leaves. Male flowers clustered, females solitary or few. All flowers with 3 to 5 valvate petals and 3 to 5 stamens, alternate with the petals, inserted around a flattened, angular disk. Sepals none. Ovary 3- or 4-locular, with a single ovule in each locule. Style short, dividing into 3 or 4 stigmas. Fruit drupaceous with 3 or 4 stones. One genus (*Helwingia*) with 4 species distributed in the Himalayas and eastern Asia.

Family Araliaceae (ivy or ginseng family)

Mostly trees or shrubs, some climbers with aerial roots. Leaves often large, compound, usually alternate, rarely opposite or whorled. Stipules small, usually more or less fused to the leafstalks. Inflorescences usually compound umbels but sometimes spikes or racemes, or rarely with flowers solitary. Flowers typically 5-merous—petals 5, sepals 5, very small, stamens 5. Ovary inferior with 1 ovule per locule. Locules variable in number. Styles usually directly on ovary (sessile). Fruit a berry, rarely a drupe. About 55 genera and 700 species distributed mostly in Southeast Asia and tropical America.

Family Apiaceae (Umbelliferae; parsley family)

Mostly herbs with pinnately divided leaves, rarely simple, or spinous, generally having sheathing leaf bases. Flowers usually bisexual (polygamous in some genera). Sepals and petals 5. Sepals frequently small or missing. Petals frequently with a strongly inflexed apical lobe. Inflorescence an umbel or compound umbel, or flowers sometimes condensed into a tight head. Ovary bilocular, with a single ovule in each locule, surmounted by a disk, on which the styles with variously thickened style bases or "stylopodia," are set. Fruit a dry schizocarp. About 275 genera and 2,850 species distributed throughout the world, but chiefly in north temperate regions.

Critical appraisal. As far as the broad outline of the classification is concerned, the evidence for the unity of this order is now impressive, and it is one of the few orders that has been given attention by many disciplines. Opinions differ as to whether the family Araliaceae should be separated as an independent order (Araliales) from the Cornales in the strict sense, largely because of the differing weight of evidence given by the study of anatomy, external morphology, and other evidence.

There is much need for further serological (study of plant proteins by means of effects similar to immune reactions in animals) and comparative biochemical studies on common species and groups of species in the order and particularly for general study in the developing countries of more obscure and atypical species, which, although largely ignored because they are of no economic value, often hold the key to plant relationships.

BIBLIOGRAPHY. J. HUTCHINSON, "Araliales," in *The Genera of Flowering Plants*, vol. 2, pp. 41–89 (1967), includes descriptions of the genera in the families Cornaceae (under which Mastixiaceae and Toricelliaceae are also treated), Alangiaceae, Garryaceae, Nyssaceae (including Davidiaceae), and Ariliaceae (including Helwingiaceae); S. BLOEMBERGEN, "A Revision of the Genus Alangium," *Bull. Jard. Bot. Buitenz.*, Series 3, 16:139–235 (1939), a good revision of the larger of the two genera in the family Alangiaceae, comprising all but three of the species in the family; H. HARMS, "Araliaceae," in A. ENGLER and K. PRANTL (eds.), *Die natürlichen Pflanzenfamilien*, 3:1–62 (1898), a good general view of the family, although out of date; R. VIGUIER, "Recherches anatomiques sur la classification des Araliacées," *Annls. Sci. Nat.*, Series 9, 4:1–207 (1906); and "Nouvelles recherches sur les Araliacées," *ibid.*, 9:305–405 (1909), two contributions forming a fairly extensive and detailed treatment of the anatomy of the family as it bears on taxonomy; W. WANGERIN, "Cornaceae," "Garryaceae," and "Nyssaceae," in A. ENGLER, *Das Pflanzenreich IV*, 229:1–110, 56a:1–17, and 220a:1–19 (1910), somewhat dated but still good treatments of these families; H. LI, "Davidia As the Type of a New Family Davidiaceae," *Lloydia*, 17:329–331 (1954), the case for treating *Davidia* as a separate family; A.S. HORNE, "The Structure and Affinities of *Davidia involucrata*, Baill," *Trans. Linn. Soc. Lond.*, 2nd Series, 7:303–326 (1909), a classic study on the internal and external morphology of this remarkable plant; F.A. HALLOCK, "The Relationship of Garrya," *Ann. Bot.*, 44: 771–812 (1930), a thorough and detailed study of *Garrya*, supporting Engler's position for the *Garryaceae;* R.H. EYDE, "Morphological and Paleobotanical Studies of the Nyssaceae, I. A Survey of the Modern Species and Their Fruits," *J. Arnold Arb.*, 44:1–59 (1963); and with E.S. BARGHOORN, ". . . of the Nyssaceae, II. The Fossil Record," *ibid.*, pp. 328–376 (1963), important studies involving *Mastixia* and *Davidia* as well as the family Nyssaceae in the strict sense; O. DRUDE, "Umbelliferae," in *Die natürlichen Pflanzenfamilien*, 3:63–250 (1898), an excellent review of the family, not yet superseded; H. WOLFF, "Umbelliferae (Saniculoideae, Apioideae, Ammineae)," in *Pflanzenreich IV*, vol. 228, pt. 43, 61, and 90 (1910–27), the standard work on these tribes, thorough and reliable.

(C.C.T.)

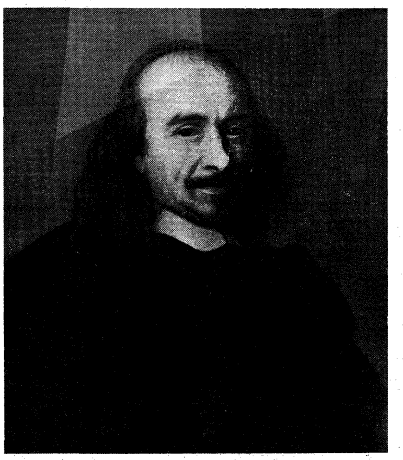
Corneille, Pierre

Not only did Pierre Corneille produce, for nearly 40 years in all, an astonishing variety of plays to entertain the French court and the Parisian middle class: he also prepared the way for a dramatic theatre that was the envy of Europe throughout the 17th century. His own contribution to this theatre, moreover, was that of master as much as of pioneer. Corneille's excellence as a playwright has long been held to lie in his ability to depict personal and moral forces in conflict. In play after play, dramatic situations lead to a finely balanced discussion of controversial issues. Willpower and self-mastery are glorified in many of his heroes, who display a heroic energy in meeting or mastering the dilemma that they face; but Corneille was less interested in exciting his audiences to pity and fear through visions of the limits of man's agony and endurance than he was in stirring them to admiration of his heroes. Thus, only a few of his plays deal in tragic emotion. Nevertheless, because his most famous work, *Le Cid*, anticipated the tragic intensity of plays by Jean Racine, his younger contemporary, Corneille has often been referred to as the "father" of French classical tragedy; and his contribution to the rise of comedy has, in comparison, often been overlooked. From a 20th-century vantage point, however, it is as a master of drama that he appears, rather than of tragedy in particular.

Early success. Pierre Corneille was born in Rouen, in northern France, on June 6, 1606, of a well-to-do, middle class Norman family. His grandfather, father, and an uncle were all lawyers; another uncle and a brother entered the church; his younger brother, Thomas, became a well-known poet and popular playwright. Pierre was educated at the Jesuit school in his home town, won two prizes for Latin verse composition, and became a licentiate in law. From 1628 to 1650 he held the position of king's counsellor in the local office of the department of waterways and forests.

Early comedies

Corneille's first play, written before he was 20 and apparently drawing upon a personal love experience, was an elegant and witty comedy, *Mélite*, first performed in Rouen in 1629. When it was repeated in Paris the following year, it built into a steady (and, according to Corneille, surprising) success. His next plays were comedies: *Clitandre* (performed 1631), *La Veuve* and *La Galerie du palais* (1632), *La Suivant* and *La Place royale* (1634), and *L'Illusion comique* (1635). His talent, meanwhile, had come to the attention of the cardinal de



Corneille, oil painting attributed to Charles Le Brun, 1647. In the Musée National de Versailles et des Trianons.
Cliche Musees Nationaux, Paris

Richelieu, France's great statesman, who included the playwright among a group known as *les cinq auteurs* ("society of the five authors"), which the Cardinal had formed to have plays written, the inspiration and outline of which were provided by himself. Corneille was temperamentally unsuited to this collective endeavour and irritated Richelieu by departing from his part (Act III) of the outline for *La Comédie des Tuileries* (1635). In the event, Corneille's contribution was artistically outstanding.

During these years, support had been growing for a new approach to tragedy that aimed at "regularity" through observance of what were called the "classical" unities. Deriving from Italy this doctrine of the unities demanded that there be unity of time (strictly, the play's events were to be limited to "the period between sunrise and sunset"), of place (the entire action was to take place in the one locus), and of action (subplots and the dramatic treatment of more than one situation were to be avoided). All this was based on a misunderstanding of Aristotle's *Poetics*, in which the philosopher attempted to give a critical definition of the nature of tragedy. The new theory was first put into dramatic practice in Jean Mairet's *Sophonisbe* (1634), a tragedy that enjoyed considerable success. Corneille, not directly involved in the call for regular tragedy of this kind, nevertheless responded to *Sophonisbe* by experimenting in the tragic form with *Médée* (1635). He then wrote *Le Cid* (performed early 1637), first issued as a tragicomedy, later as a tragedy.

Success of *Le Cid*

Le Cid, now commonly regarded as the most significant play in the history of French drama, proved an immense popular success. It sparked off a literary controversy, however, which was chiefly conducted by Corneille's rival dramatists, Mairet and Georges de Scudéry, and which resulted in a bitter pamphlet war. Richelieu, whose motives are not entirely clear, instructed the then recently instituted Académie Française to make a judgment on the play: the resulting document (*Les Sentiments de l'Académie française sur la tragi-comédie du Cid*, 1637), drafted in the main by Jean Chapelain, a critic who advocated "regular" tragedy, was worded tactfully and admitted the play's beauties but criticized *Le Cid* as dramatically implausible and morally defective. Richelieu used the judgment of the Académie as an excuse for suppressing public performances of the play.

Corneille, indeed, had not observed the dramatic unities in *Le Cid*. The play has nevertheless been generally regarded as the first flowering of French "classical" tragedy. For the best French drama of the "classical" period in the 17th century is properly characterized, not so much by rules—which are no more than a structural convention—as by emotional concentration on a moral dilemma and on a supreme moment of truth, when leading characters recognize the depth of their involvement in this dilemma. In *Le Cid*, Corneille rejected the discursive treatment of the subject given in his Spanish source (a long, florid, and violent play by Guillén de Castro y Bellvis, a 17th-century dramatist), concentrating instead on a conflict between passionate love and family loyalty, or honour. Thus *Le Cid* anticipated the "pure" tragedy of Racine, in whose work the "classical" concept of tragic intensity at the moment of self-realization found its most mature and perfect expression.

Major tragedies. Corneille seems to have taken to heart the criticisms levelled at *Le Cid*, and he wrote nothing for three years (though this time was also taken up with a lawsuit to prevent the creation of a legal office in Rouen on a par with his own). In 1640, however, appeared the Roman tragedy *Horace;* another, *Cinna*, appeared in 1641. In 1641 also Corneille married Marie de Lampérière, the daughter of a local magistrate, who was to bear him seven children to whom he was a devoted father. Corneille's brother Thomas married Marie's sister, and the two couples lived in extraordinary harmony, their households hardly separated; the brothers enjoyed literary amity and mutual assistance.

Corneille's "classical tetralogy"

Le Cid, *Horace*, *Cinna*, and *Polyeucte*, which appeared in 1643, are together known as Corneille's "classical tetralogy," and together represent perhaps his finest body of work for the theatre. *Horace* was based on an account by the Roman historian Livy of a legendary combat between members of the Horatii and Curiatii families, representing Rome and Alba; Corneille, however, concentrated on the murder by one of the patriots of his pacifist sister, the whole case afterward being argued before the king (a "duplicity" of action admitted by Corneille himself, who otherwise seems by now to have decided to follow the classical rules). *Cinna* was about a conspiracy against the first Roman emperor, Augustus, who checkmates his adversaries by granting them a political pardon instead of dealing them the expected violent fate, boasting that he has strength enough to be merciful. The hero of *Polyeucte* (which many critics have considered to be Corneille's finest work), on adopting Christianity seeks a martyr's death with almost militaristic fervour, choosing this as the path to *la gloire* ("glory") in another world, whereas his wife insists that the claims of marriage are as important as those of religion.

These four plays are charged with an energy peculiar to Corneille. Their arguments, presented elegantly, rhetorically, in the grand style, remain firm and sonorous. The alexandrine verse that he employed (though not exclusively) was used with astonishing flexibility as an instrument to convey all shades of meaning and expression: irony, anger, soliloquy, repartee, epigram. Corneille used language not so much to illumine character as to heighten the clash between concepts, hence the "sentences" in his poetry which are memorable even outside their dramatic context. Action here is reaction. These plays concern not so much what is done as what is resolved, felt, suffered. Their formal principle is symmetry: presentation, by a poet who was also a lawyer, of one side of the case then of the other, of one position followed by its opposite.

Contribution to comedy. The fame of his "classical tetralogy" has tended to obscure the enormous variety of Corneille's other drama, and his contribution to the development of French comedy has not always received its proper due. The Roman plays were followed by more tragedies: *La Mort de Pompée* (1643), *Rodogune* (1645), which was one of his greatest successes, *Théodore* (1646), which was his first taste of failure, and *Héraclius* (1647). But in 1643 Corneille had successfully turned to comedy with *Le Menteur*, following it with the less successful *La Suite du Menteur* (1643–44). Both were lively comedies of intrigue, adapted from Spanish models; and *Le Menteur* is the one outstanding French comedy before the plays of Molière, Corneille's young contemporary, who acknowledged its influence on his own work. *Le Menteur*, indeed, stands in relation to French classical comedy much as *Le Cid* does to tragedy.

In 1647, Corneille moved with his family to Paris and

was at last admitted to the Académie Française, having twice previously been rejected on the grounds of non-residence in the capital. *Don Sanche d'Aragon* (1649), *Andromède* (1650), a spectacular play in which stage machinery was very important, and *Nicomède* (1651) were all written during the political upheaval and civil war of the period known as the Fronde (1648–53), with *Don Sanche* in particular carrying contemporary political overtones. In 1651 or 1652 his play *Pertharite* seems to have been brutally received, and for the next eight years Corneille wrote nothing for the theatre, concentrating instead on a verse translation of St. Thomas à Kempis' *Imitatio Christi* (*Imitation of Christ*), which he completed in 1656, and also working at critical discourses on his plays that were to be included in a 1660 edition of his collected works.

Years of declining power. Corneille did not turn again to the theatre until 1659, when, with the encouragement of the statesman and patron of the arts Nicolas Fouquet, he presented *Oedipe*. For the next 14 years he wrote almost one play a year, including *Sertorius* (1662) and *Attila* (1667), both of which contain an amount of violent and surprising incident.

Last plays

Corneille's last plays, indeed, were closer in spirit to his works of the 1640s than to his classical tragedies. Their plots were endlessly complicated, their emotional climate close to that of tragicomedy. Other late plays include *La Toison d'or* (1660), his own *Sophonisbe* (1663), *Othon* (1664), *Agésilas* (1666), and *Pulchérie* (1672). In collaboration with Molière and Philippe Quinault he wrote *Psyché* (1671), a play employing music, incorporating ballet sequences, and striking a note of lyrical tenderness. A year earlier, however, he had presented *Tite et Bérénice*, in deliberate contest with a play on the same subject by Racine. Its failure indicated the public's growing preference for the younger playwright.

Corneille's final play was *Suréna* (1674), which showed an uncharacteristic delicacy and sentimental appeal. After this he was silent except for some beautiful verses, which appeared in 1676, thanking King Louis XIV for ordering the revival of his plays. Although not in desperate poverty, Corneille was by no means wealthy; and his situation was further embarrassed by the intermittent stoppage of a state pension that had been granted by Richelieu soon after the appearance of *Horace* in 1640. Corneille died in his house on the rue d'Argenteuil, Paris, on October 1, 1684, and was buried in the church of Saint-Roch. No monument marked his tomb until 1821.

Assessment. Corneille did not have to wait for "the next age" to do him justice. The cabal that had led the attack on *Le Cid* had no effect on the judgment of the public, and the great men of his time were his fervent admirers. Balzac praised him; Molière acknowledged him as his master and as the foremost of dramatists; Racine is said to have assured his son that Corneille made verses "a hundred times more beautiful" than his own. It was left to the 18th century, largely because of the criticisms of Voltaire, to exalt Racine at Corneille's expense; but the Romantic critics of the late 18th century began to restore Corneille to his true rank.

It cannot be denied, however, that Corneille signed much verse that is dull to mediocre. Molière acknowledged this fact by saying: "My friend Corneille has a familiar who inspires him with the finest verses in the world. But sometimes the familiar leaves him to shift for himself, and then he fares very badly." But the importance of his pioneer work in the development of French classical theatre cannot be denied; and, if a poet is to be judged by his best things, Corneille's place among the great dramatic poets is beyond question.

MAJOR WORKS

PLAYS: *Le Cid* (published 1637); *Horace* (1641); *Cinna, ou La Clemence d'Auguste* (1643); *Polyeucte martyr* (1643); *La Mort de Pompée* (1644)—all in English in *The Chief Plays of Corneille*, trans. by Lacy Lockert, 2nd ed. (1957). *Rodogune, princesse des Parthes* (1647; *Rodogune; or, The Rival Brothers*, trans. by S. Aspinwall, 1765); *Nicomède* (1651; *Nicomede*, trans. by J. Dancer, 1671).

BIBLIOGRAPHY. The standard edition of the dramatist's work, *Oeuvres de P. Corneille*, ed. by CHARLES MARTY-LAVEAUX, 13 vol. (1862–68), is still generally reliable, although a number of more recent editions—by MAURICE RAT (1962–66), and by JACQUES MAURENS (1968)—have profited from extensive work by modern scholars on dating and various historical aspects. A number of convenient one- or two-volume modern editions are available, notably *Oeuvres complètes*, ed. by ANDRE STEGMANN (1963). Many of the plays have also been published in critical editions (see list in ALEXANDRE CIORANESCU, *Bibliographie de la littérature française du 17e siècle*, 1965). Adequate modern translations into English have been provided by LACY LOCKERT, *The Chief Plays of Corneille*, 2nd ed. (1957) and *Moot Plays of Corneille* (1959); and more recently by SAMUEL SOLOMON, *Pierre Corneille: Seven Plays* (1969). Modern criticism of Corneille has begun to reverse the monotonous, reductionist view first set forth in Voltaire's *Commentaires sur Corneille* (1751) that had cast him chiefly as Racine's precursor in the perfecting of French classical tragedy. For a comprehensive sampling of the Corneille-Racine parallels across the centuries, see *Corneille and Racine: Parallels and Contrasts*, ed. by ROBERT J. NELSON (1966). A judicious corrective to the Voltairean bias informs both GEORGES MAY, *Tragédie cornélienne, tragédie racinienne* (1948); and OCTAVE NADAL'S seminal *Le Sentiment de l'amour dans l'oeuvre de Pierre Corneille* (1948). The two-volume study by ANDRE STEGMANN, *L'Héroïsme cornélien—genèse et signification* (1968), offers a learned, penetrating discussion of Corneille in the religious and political setting of his own time; while, in *Corneille et la dialectique du héros* (1963), SERGE DOUBROVSKY reviews the dramatist's intellectual and political outlook in a frankly modern perspective. HENRY CARRINGTON LANCASTER, *History of French Dramatic Literature in the Seventeenth Century*, 9 vol. (1929–42), is still useful on sources, influences, editions, and theatrical history. A more compact survey of the same aspects, with greater and more thoughtful attention to problems of interpretation, may be found in PHILIP YARROW, *Corneille* (1963). For a more specialized study of Corneille's themes, particularly the dramatist's conception of tragedy and his place in the history of the genre, see ROBERT J. NELSON, *Corneille: His Heroes and Their Worlds* (1963). Finally, brilliant if controversial reevaluations of the dramatist in the larger setting of 17th-century thought and letters are contained in E.B.O. BORGERHOFF, *The Freedom of French Classicism* (1950); and in W.G. MOORE, *French Classical Literature: An Essay* (1961). Bibliographies of the author and his works include AUGUSTE-EMILE PICOT, *Bibliographie cornélienne* . . . (1876); and PIERRE LE VERDIER and EDOUARD PELEY, *Additions à la bibliographie cornélienne* (1908), which are both still useful. Also, a number of more recent general bibliographies, such as *A Critical Bibliography of French Literature*, vol. 3, *The Seventeenth Century*, ed. by NATHAN EDELMAN (1961), contain extensive, updated sections on the dramatist.

(R.J.N.)

Corot, Jean-Baptiste-Camille

Camille Corot, as he liked to be called, was a 19th-century French painter of landscape and, to a lesser extent, of figure subjects. He began his long career as an exponent of the classical tradition of idealized Italian landscape, as exemplified by the 17th-century French masters Nicolas Poussin and Claude Lorrain, and he ended it as the contemporary and friend of the Impressionists, whose new and revolutionary landscape art he inspired and, in certain respects, anticipated.

Jean-Baptiste-Camille Corot was born of prosperous, bourgeois parents in Paris on July 16, 1796. His mother, who was Swiss born, had a fashionable milliner's shop, which Corot's father—a draper by trade—helped to manage. Camille was a poor scholar and even less adept when he tried to follow his father's trade. Finally, at the age of 25, he was given a small allowance by his father and allowed to become what he had always dreamed of being: a painter.

Like every young French artist, Corot spent much time studying the paintings in the Louvre, and he had some private instruction from Achille-Etna Michallon and Jean-Victor Bertin, who were both followers of the Neoclassical landscape painter Pierre-Henri de Valenciennes. From the beginning, however, Corot preferred to sit outdoors, sketching what he saw and learning by firsthand experience.

The Roman years: 1825–28

In the autumn of 1825 Corot went to Rome, and the three years that he spent there were the most important of his life. He painted the city and the Campagna, the

countryside around Rome; he made a trip to Naples and Ischia; and he returned to Paris by way of Venice. He was very happy. He told a friend in August 1826: "All I really want to do in life . . . is to paint landscapes. This firm resolve will stop me forming any serious attachments. That is to say, I shall not get married." He was as good as his word and never married. Female companionship played no part in his life, which was entirely devoted to painting.

By courtesy of the Gernsheim Collection, Humanities Research Center, the University of Texas at Austin



Corot, photographed by Eugène Dutilleux at Arras, France, 1871.

Back in France, Corot settled into a routine to which he kept for the whole of his life. He always spent the spring and summer months painting outside, making small oil sketches and drawings from nature. He acquired a mastery of tonal relationships that formed the basis of his art, for the balance and gradation of light and dark tones was always more important to him than the choice of colour. In the winter Corot would retire to his Paris studio to work on some much larger pictures, which he liked to have ready for exhibition at the annual Salon when it opened in May.

His first important work, "Le Pont de Narni," was shown at the Paris Salon in 1827, when he was still in Italy. In 1833 he exhibited a large landscape of the forest of Fontainebleau, which was awarded a second-class medal: this gave Corot the right to show his pictures without submission to the jury for their approval.

From May to October of 1834 Corot made his second visit to Italy. He painted views of Volterra, Florence, Pisa, Genoa, Venice, and the Italian lake district. He collected enough material in small sketches to last him the rest of his life, although he returned to Italy briefly in the summer of 1843, for the last time.

Apart from these visits to Italy, Corot had an exceptionally uneventful life. As a young man he had travelled widely in the provinces of France, sketching wherever he went. When he grew older he moved around less. Important trips were made to Avignon and the south of France in 1836, to Switzerland in 1842 and on several other occasions, to The Netherlands in 1854, and to London in 1862. His favorite regions of France were the forest of Fontainebleau, Brittany, the Normandy coast, the family property at Ville-d'Avray near Paris, and, later in life, Arras and Douai—in the north of France—where close friends lived.

Topographical landscapes

Throughout his life Corot liked occasionally to paint straightforward topographical landscapes, depicting buildings such as the cathedral at Chartres (1830) or the belfry at Douai (1871) exactly as they appeared to him. But the basic division in his work was between the sketch made from nature—small, direct, spontaneous—and the large finished picture done for the Salon. In the early 19th century the sketch was thought not to be suitable for public exhibition, and there were only a few connoisseur collectors who would buy such pictures. The finished landscapes were preferred. These were considered even more dignified and serious if they included a few small figures who could be identified with the heroic characters of legend, literature, or the Bible. Thus, Corot exhibited pictures with such titles as "Hagar in the Wilderness" (Salon 1835), "Diana Surprised by Actaeon" (Salon 1836), "Homer and the Shepherds" (Salon 1845), and "Christ in the Garden of Olives" (Salon 1849).

He was not really very interested in such subject matter, however, and, as tastes changed in the middle of the century, he bothered less about it. He was happy to introduce pretty young girls into the landscape, showing them as shepherdesses or sometimes, more daringly, as nude or half-dressed bathers, dipping their delicate feet in a pool. In the 1860s he invented a new kind of landscape, the "Souvenirs," in which he made compositions out of standardized elements—usually a lake with diaphanous trees painted in an overall silvery tonality—to evoke a mood of gentle melancholy. At the end of his life, he also painted a number of portraits and figure studies, especially of young women posed in his studio holding a flower or a musical instrument or looking at a landscape on the easel. These more private pictures Corot almost never exhibited.

During the 1830s Corot showed regularly at the Paris Salon and had some critical success. Yet he sold very few pictures and was glad of his father's allowance. Then, in 1840, the state purchased one of his exhibits, "The Little Shepherd" ("Le Petit Berger"), and, five years later, the poet and art critic Charles Baudelaire could write in his review of the 1845 Salon that "Corot stands at the head of the modern school of landscape." In 1846 he was made a member of the Legion of Honour, and, when his father died, in 1847, Corot was able to feel that he had justified the family's support of his ambition to be a painter.

Years of success

By the 1850s, collectors and dealers were eagerly seeking his pictures, and Corot henceforth had no material worries. He went on sending big pictures to the salons, where they fetched high prices. At the 1855 Paris Universal Exhibition he was awarded a first-class medal for painting, and Emperor Napoleon III bought a picture from him. In 1867 he was promoted to officer of the Legion of Honour. But taste was changing, and private collectors, especially in the United States, began to buy Corot's smaller, sketchier work. Though he was a prolific artist and painted over 3,000 pictures, demand outran supply, and Corot was much imitated and faked. In the 20th century, appreciation of Corot has shifted again to show a marked preference for the earlier, more naturalistic sketches over the later, more self-consciously poetic ones. The big Salon paintings have remained firmly out of fashion, though their qualities are not negligible.

Success made little difference to Corot, who was a man of extremely conservative habits. He always worked very hard because he loved his work, and this left him little time for other things. In old age he suffered from gout, but otherwise his health was good. He never read newspapers and took no interest in politics. The revolutions and changes of regime in his time passed him by. Altogether, he read little and showed no intellectual curiosity. He was very fond of music, especially Haydn, Mozart, Gluck, and Beethoven, and he regularly attended concerts when he was in Paris. He liked to talk about the harmonies of his painting, and his late work in particular aspires to the qualities of music. He kept the modern world firmly out of his pictures: the people in his landscapes are timeless peasant figures or characters from myth and literature. There is never a sign of the vast railway network that covered France in his lifetime or of the industrial and commercial development that transformed the country.

Corot enjoyed the company of fellow painters and was a close friend of the Barbizon group of artists, especially

Relations with the Barbizon painters and the Impressionists

Jean-François Millet, Théodore Rousseau, and Charles-François Daubigny. He used his money to give unostentatious help to less successful friends, such as the cartoonist Honoré Daumier. Without going out of his way to support them in public, Corot was sympathetic to younger painters. He gave lessons to the later Impressionists Camille Pissarro and Berthe Morisot and had many pupils and disciples. "Papa Corot" was universally loved for his unfailing kindness and generosity when he died, in Paris, on February 22, 1875.

Corot's place in the history of 19th-century painting is an assured one. He is no longer regarded as a major master, but his contribution was essential. If his larger pictures often look empty and rhetorical, his unpretentious paintings are a constant delight because of the freshness of their observation and the effortless ease of technique. When he started painting, the landscape sketch was regarded primarily as raw material for more considered work and was of no great artistic consequence in itself. Corot was one of the first to show that the sketch had qualities of vitality and spontaneity, a basic truth to nature that a more finished picture lacked. At the time of his death the sketch had triumphed, and any artificiality or contrivance in landscape painting was regarded with suspicion. Corot had helped to prepare the way for the Impressionist landscape painters, who learned much from him and who looked upon him with respect and veneration.

MAJOR WORKS

"Portrait de l'artiste à l'âge de vingt-neuf ans" ("Self-Portrait at the Age of 29"; 1825; Louvre, Paris); "Landscape: Le Petit Chaville, near Ville-d'Avray" (1823–25; Ashmolean Museum, Oxford); "View of the Forum from the Farnese Gardens" (1826; Louvre); "Old Man Sitting in Corot's Studio, Rome" (1826; Museum of Fine Arts, Boston, Massachusetts); "Le Pont de Narni" ("The Bridge at Narni"; 1826–27; National Gallery of Canada, Ottawa); "The Roman Campagna, with the Claudian Aqueduct" (1826–28; National Gallery, London); "Italian Monk Reading" (1827; Albright-Knox Art Gallery, Buffalo, New York); "Façade ouest de la Cathédrale de Chartres" (1830; Louvre); "View of Soissons" (1833; Rijksmuseum Kröller-Müller, Otterlo, The Netherlands); "View of Genoa" (1834; Art Institute of Chicago); "Hagar in the Wilderness" (1834–35; Metropolitan Museum of Art, New York); "Diana Surprised by Actaeon" (1836; Musée des Beaux-Arts, Bordeaux); "A View near Volterra" (1838; National Gallery of Art, Washington, D.C.); "The Little Shepherd" ("Le Petit Berger"; 1840; Musée de Metz, Metz, France); "Les Bords du Cousin" (*c.* 1840–45; Louvre, Paris); "Marietta 'L'Odalisque Romaine'" (1843; Musée du Petit Palais, Paris); "Homer and the Shepherds" (1845; Musée d'Art, Saint-Lô, France); "The Forest of Fontainebleau" (1846; Museum of Fine Arts, Boston); "La Danse des nymphes" (1850; Louvre); "Bretonnes à la fontaine" (early 1850s; Louvre); "Le Port de La Rochelle" ("The Harbor of La Rochelle"; 1851; Yale University Art Gallery, New Haven, Connecticut); "Entrée de village (environs de Beauvais du côte de Voinsinlieu)" (*c.* 1855–60; Louvre); "Portrait de Maurice Robert enfant" (1857; Louvre); "Macbeth and the Witches" (1859; Wallace Collection, London); "Jeune fille à sa toilette" (*c.* 1860–65; Louvre); "Souvenir de Mortefontaine" (1864; Louvre); "L'atelier" (1865–68; Louvre); "La Femme à la perle" (*c.* 1868–70, Louvre); "Le Coup de vent" ("The Gust of Wind"; *c.* 1865–70; Musée Saint-Denis, Reims, France); "Mademoiselle de Foudras" (1872; Glasgow Art Gallery and Museum); "La Charrette—Souvenir de Saintry" (1874; National Gallery, London); "La Dame en bleu" (1874; Louvre).

BIBLIOGRAPHY. The standard work is *L'Oeuvre de Corot*, 4 vol. (1905), a complete catalog by ALFRED ROBAUT, with a biographical introductory volume by ETIENNE MOREAU NELATON. There are also two supplementary volumes by ANDRE SCHOELLER and JEAN DIETERLE (1948 and 1956). *Corot, raconté par lui-même et par ses amis*, 2 vol., ed. by PIERRE COURTHION and PIERRE CAILLER (1946), is a useful collection of source material and letters. No serious study of Corot exists in English. The best monograph is GERMAIN BAZIN, *Corot* (1942), in French.

(Al.Bo.)

Corporation, Business

In terms of its size, influence, and visibility, the corporation has become the dominant business form in Western industrial societies. While corporations may be large or small, ranging from firms having hundreds of thousands of employees to neighbourhood businesses of very modest proportions, public attention has increasingly focussed on the several hundred giant companies that play a preponderant economic role in the United States, Canada, Japan, and the nations of western Europe. These firms not only occupy important positions in the economy; they have great social, political, and cultural influence as well. Both at home and abroad they affect the operations of governments, give shape to the landscapes of local communities and entire societies, and help to determine the values of ordinary individuals. Therefore, while in fact and in law corporate businesses are private enterprises, their activities have consequences that are public in character and as pervasive as those of many governments. No account will be taken here of the so-called public corporations that have arisen in recent decades in a number of countries. (Such government bodies, including the Tennessee Valley Authority and the Port of New York Authority in the United States and the National Coal Board in Britain are discussed in PUBLIC ENTERPRISES.)

Forms of business enterprises

Business enterprises customarily take one of three forms: individual proprietorships, partnerships, and corporations. In the first of these, a single person holds the entire firm as his personal property, and, in most instances, it is an operation that he manages on a day-to-day basis. Of the nearly 12,000,000 business establishments conducted in the United States in 1968, more than three-quarters were proprietorships and more than half of them had annual receipts of less than $10,000. A partnership may have from two to 50 or more members, as in the case of large law firms, brokerage houses, and advertising agencies. This form of business is owned by the partners themselves, although they may receive varying shares of the profits depending on their investment or contribution. Whenever a member leaves or a new member is added, the firm must be reconstituted as a new partnership.

The corporate arrangement is rather more complex. One way of regarding a corporation is as an association of assets, chartered under law, that is owned by individuals or institutions who have purchased shares representing fractions of the firm's holdings. By this token, a person owns part of a corporation only so long as he retains his shares in the company; and a corporation maintains its independent existence even if its stock changes hands, as it often does. The degree to which stockholders participate in the operation of their company varies, usually according to the firm's size. In 1968 the United States had more than 1,500,000 corporations, but more than half of these did an annual business of under $100,000. At the other end of the spectrum, 274,000 larger corporations accounted for more than 90 percent of all business conducted under this form; and, in 1970, the 500 largest industrial companies in the United States accounted for almost 75 percent of the profits earned in manufacturing. Some relatively small firms are corporations: a local grocery store or a dry-cleaning business or even a taxicab may be incorporated, with a proportion of its shares owned by a few friends or relatives of the man who actually runs it. But the focus of this article is on the small circle of companies commanding substantial assets, sales, and earnings.

EARLY HISTORY OF THE CORPORATION

The corporation is a relatively recent innovation. It has only been since the mid-19th century that incorporated businesses have risen to ascendancy over other modes of ownership. Thus, any attempt to trace the forerunners of the modern corporation should be distinguished from a general history of business or a chronicle of associated activity. Men have embarked on enterprises for profit and have joined together for collective purposes since the dawn of recorded history, but the early enterprises were related to the contemporary corporation only in the sense that new developments always embody some practices from the past. When a group of Athenian or

Phoenician merchants pooled their savings to build or charter a trading vessel, their organization was not a corporation but a partnership; ancient societies did not have laws of incorporation that specified the scope and standards of business activity.

Corporations in the Middle Ages

The corporate form itself developed in the early Middle Ages with the growth and codification of civil and canon law. Several centuries, however, would pass before business ownership was subsumed under this arrangement. The first corporations were towns, universities, and ecclesiastical orders. They differed from partnerships in that the organization existed independently of any particular membership; but they were not, like modern business corporations, the "property" of their participants. The holdings of a monastery, for example, belonged to the order itself; no individual owned shares in its assets. The same was true of the medieval guilds, which dominated many trades and occupations. As corporate bodies, they were chartered by government, and their business practices were regulated by public statutes; but each guild member was an individual proprietor who ran his own establishment, and, while many guilds had substantial properties, these were the historic accruals of the associations themselves. By the 15th century, the courts of England had agreed on the principle of "limited liability": *Si quid universitati debetur, singulis non debetur, nec quod debet universitas, singuli debent* ("If something is owed to the group, it is not owed to the individuals nor do the individuals owe what the group owes"). Originally applied to guilds and municipalities, this principle set limits on how much an alderman of the Liverpool Corporation, for example, might be called upon to pay if the city ran into debt or bankruptcy. As applied later on to stockholders in business corporations, it served to encourage investment because the most an individual could lose in the event of the firm's failure would be the actual amount he had originally paid for his shares.

The actual incorporation of business enterprises began in England during the Elizabethan era. This was a period when businessmen were beginning to accumulate substantial surpluses, and overseas exploration presented itself as an investment opportunity. This was also an age that gave overriding regulatory powers to the state, which sought to ensure that business activity was consonant with current mercantilist conceptions of national prosperity. Thus, the first joint-stock companies, while financed with private capital, were created by public charters setting down in detail the activities in which the enterprises might operate. In 1600, Queen Elizabeth I granted to a group of investors headed by the Earl of Sunderland the right to be "one body corporate," known as the Governor and Company of Merchants of London, trading into the East Indies. The East India Company was bestowed a trading monopoly in its territories and also was given the authority to make and enforce laws in the areas it entered. The East India Company, the Royal African Company, the Hudson's Bay Company, and similar incorporated firms were semipublic enterprises acting as arms of the state, as well as vehicles for private profit. The same principle held with the colonial charters on the American continent. In 1606 the crown vested in a syndicate of "loving and well-disposed Subjects" the right to develop Virginia as a royal domain, including the power to coin money and to maintain a military force. The same was done, in subsequent decades, for the "Governor and Company of the Massachusetts Bay in New England," and for William Penn's "Free Society of Traders" in Pennsylvania.

Much of North America's settlement was initially underwritten as a business venture. But, if British investors accepted the regulations inhering in their charters, American entrepreneurs were apt to regard such rules as repressive and unrealistic. The American Revolution was, in large measure, directed against the tenets of the mercantile system, raising serious questions about the whole idea of a direct tie between business enterprise and public policy. One result of the U.S. War of Independence, therefore, was to establish the premise that a corporation need not show that its activities advance a specific public purpose. Alexander Hamilton, the first secretary of the treasury and an admirer of Adam Smith, took the view that businessmen should be encouraged to explore their own avenues of enterprise. "To cherish and stimulate the activity of the human mind, by multiplying the objects of enterprise, is not among the least considerable of the expedients by which the wealth of a nation may be promoted," he wrote in 1791.

The corporate revolution did not occur overnight. For a long time, both in Europe and in the United States, the corporate form was regarded as a creature of government and monopoly. In the U.S. the new state legislatures granted charters principally to public-service companies intending to build or operate docks, bridges, turnpikes, canals, and waterworks, as well as to banks and insurance companies. Of the 335 companies receiving charters prior to 1800, only 13 were firms engaging in commerce or manufacturing. By 1811, however, New York had adopted a general act of incorporation, setting the precedent that businessmen had only to provide a summary description of their intentions for permission to launch an enterprise. By the 1840s and 1850s, the rest of the states had followed suit. In Great Britain after 1825, the statutes were gradually liberalized so that the former privilege of incorporating joint-stock companies became the right of any group complying with certain minimum conditions, and the principle of limited liability was extended to them. In France and Germany a similar development occurred.

19th-century statutes

DEVELOPMENT OF CORPORATE BUSINESS IN THE U.S.

Pools, trusts, and holding companies. In the United States the corporation took its present form in the latter part of the 19th century. The transcontinental railroad lines needed massive infusions of capital and depended on public stock issues. A great spur to incorporation was the Supreme Court case of *Santa Clara County* v. *Southern Pacific Railroad* in 1886 when the court ruled for the first time that a corporation should be construed as a "person" and was thus entitled to the protection of the Fourteenth Amendment of the Constitution, which declares: "nor shall any State deprive any person of life, liberty, or property, without due process of law." Under this interpretation, many laws seeking to govern corporate practices were declared to be violations of "due process" and hence unconstitutional.

The 1880s and 1890s witnessed the development of huge companies to serve an emerging national market; the creation of pools, trusts, and holding companies to keep prices high and competition limited; and the appearance of such industrial magnates as John D. Rockefeller, Andrew Carnegie, and James J. Hill. But many large firms remained private holdings. The Carnegie Steel Company, for example, was a partnership not open to public participation, and several of Rockefeller's Standard Oil affiliates were not incorporated. Until almost the turn of the century, corporations continued to be presided over by the generation of men who had originally founded them: the entrepreneurial presence had yet to be replaced by formal organization.

Growth by merger

A turn came within a surprisingly brief span of time, as investment banks began to serve as agents for industrial mergers. The half decade from 1899 to 1904 produced United States Steel, United States Rubber, American Can, and International Harvester, along with more than 2,500 other mergers—almost six times as many as had occurred in the preceding four years. By 1904, seven out of every ten of the country's production workers were employed by incorporated businesses. The pattern in the ensuing years was one of growth and expansion by the giant firms. Between 1919 and 1939, for example, the share of net corporate income earned by the 75 percent of the firms at the bottom of the industrial pyramid fell by one-half; even the larger corporations composing the next 20 percent saw their share of the total earnings fall by one-quarter. Between 1916 and 1956, while the nation's population rose by two-thirds, the number of corporations grew almost threefold.

The growth of conglomerate corporations. The giant firms continued to increase their sales and assets, both by expanding their markets and by absorbing smaller companies. Another period of mergers and acquisitions came in the 1960s. Whereas in the years 1948 through 1959 corporate mergers averaged only 428 per year, in the ensuing decade the average figure rose above 1,250. In 1969 alone, 2,307 such operations took place. Generally speaking, however, these absorptions did not involve companies in the same field. While there were exceptions, such as Atlantic's merger with Richfield in 1966 and its acquisition of Sinclair Oil in 1969, the general trend was toward diversification. Radio Corporation of America (RCA), for example, purchased the Hertz car rental agency; Columbia Broadcasting System bought Creative Playthings; and the Greyhound bus lines absorbed the Armour meat-packing company. One reason for this branching out was that the antitrust laws discouraged concentration in a single field. Another incentive to diversifying arose when a corporation foresaw limited growth in its original area. Most corporations aspire to expand their sales and assets at a rate surpassing the overall growth of the economy; if they find that demand for their major items cannot be increased, the alternative is to embark on new ventures, which is most easily done by buying up smaller companies that have growth possibilities. Thus, Liggett & Myers, the makers of Chesterfield and L&M cigarettes, anticipating a levelling off of tobacco consumption, proceeded to acquire Alpo dog food, J&B Scotch whisky, Brite watchbands, and Blue Lustre cleaning products.

Conglomerates

Diversification carried to the extreme brought into being the "conglomerate" corporation, which acquired and operated subsidiaries in fields often having no palpable relation to one another. Some of the more prominent conglomerates of the 1960s were Gulf & Western Industries, Transamerica, Ling-Temco-Vought, Boise Cascade, and Textron. Perhaps the most notable of all was the International Telephone and Telegraph Corporation. In a 15-year period from 1955 to 1970, ITT rose from the country's 80th largest corporation, with annual sales of $450,000,000, to the eighth-ranking firm, having sales in excess of $6,000,000,000—a rise of more than 1,200 percent. (During the same period, by comparison, General Electric's sales went up 181 percent and duPont's by 89 percent.) ITT achieved this by absorbing such companies as Sheraton Hotels, Avis car rentals, Bobbs-Merrill publishers, Levitt and Sons builders, Continental bakeries, and Smithfield hams, as well as firms in the cellulose, vending-machine, and fire-protection fields.

By the 1970s there were indications that the conglomerate wave had reached its crest, if only because of the general decline in stock prices. Since a conglomerate usually acquired other firms by offering its own shares in exchange for those of the smaller companies, a fall in the market value of its stock left it in a poor position to purchase new subsidiaries. Nevertheless, conglomerates had become a permanent part of the corporate scene, bringing with them financial and managerial techniques unenvisaged earlier.

Another path to growth for U.S. corporations has been through expansion abroad. Along with the export of U.S.-made goods to other countries, there was a marked tendency to move production closer to the market by establishing foreign subsidiaries. Some companies, including National Cash Register (NCR), Singer, F.W. Woolworth, and the Standard Oil Company of New Jersey and Standard Oil Company of California, received over half their revenues from sales outside the United States. The Chrysler Corporation bought Rootes Motors in Great Britain and Simca in France; Celanese and Dow Chemical acquired European manufacturing firms; and Jos. Schlitz Brewing Co. bought control of one of Belgium's largest breweries. By the end of the 1960s, more than 1,500 U.S.-owned companies in Great Britain produced 10 percent of that country's manufactured goods, dominating fields such as drugs, typewriters, and razor blades. One consequence was a large annual outflow of U.S. corporate capital to purchase foreign firms or to build new facilities abroad. The French journalist Jean-Jacques Servan-Schreiber wrote that U.S. industry in Europe had acquired the dimensions of a superpower. This kind of international expansion was not, of course, limited to United States business firms; many European and Japanese firms acquired assets in other countries, including the U.S.; but their activities were not as large and did not have the same effect on local sensitivities as those of U.S. corporations.

CHARACTERISTICS OF U.S. CORPORATIONS

In 1970, the 500 largest industrial corporations in the United States made almost two-thirds of all sales of manufacturing firms. One corporation, American Telephone & Telegraph, had over 750,000 employees, or nearly as many as the state governments of California, Illinois, Michigan, Ohio, Pennsylvania, and Florida combined. Standard Oil of New Jersey's sales exceeded $16,000,000,000, a figure larger than the taxes collected by New York, Georgia, Indiana, Massachusetts, Minnesota, New Jersey, and Texas taken together. Companies of this size inevitably have wide-ranging effects on the life around them. Many observers hold that large corporations are as important in modern society as governments. They are not only producers of goods and services but agents of innovative technology, participants in the political arena, and shapers of values and behaviour.

Dispersion of ownership

Separation of ownership and control. In order to buy the capital equipment necessary for new or expanded operations, corporations have traditionally gone to the investing public for funds. Almost every day, newspapers carry announcements that one or another company has created new shares and is offering them for sale. The result has been that the legal ownership of corporations has become widely dispersed. At the beginning of 1971, for example, General Electric and International Business Machines (IBM) each had over 500,000 shareholders. General Motors had nearly 1,500,000, and the American Telephone & Telegraph Company's list exceeded 3,000,000. While large blocks of shares may be held by wealthy individuals or institutions, the total amount of stock in these companies is so large that even a very wealthy person is not likely to own more than a small fraction of it.

The chief effect of this dispersion has been to give effective control of the companies to their salaried managers. While each company holds an annual meeting, which all stockholders may attend and vote, these gatherings, in fact, tend to ratify on-going policy as it has been determined by management. Even if sharp questions are occasionally asked, the presiding officers almost invariably hold enough proxies to override outside proposals. The only real recourse for a dissatisfied shareholder is to sell his stock and invest in a firm more to his liking. (If enough shareholders do this, of course, the price of the stock will fall quite markedly, perhaps impelling changes in management personnel or company policy.) Occasionally, there are "proxy battles," when attempts are made to persuade a majority of shareholders to vote against a firm's managers, but such struggles seldom involve the largest corporations. It is in the managers' interest to keep the stockholders happy, for, if the company's shares are regarded as a good buy, then it is easy to raise additional capital through a new stock issue.

Thus, if a corporation is doing well in sales and earnings, its executives have a relatively free hand. To be sure, there will be governmental regulations, negotiations with unions, and the constant need to woo and win the consumer. Most large firms manage to make a profit even in difficult years, just as most citizens manage to support themselves and have a little left over. If a company gets into trouble, its usual course is to agree to be merged into another corporation or to borrow money. In the latter case, the lending institution may insist on a new chief executive of its own choosing. If a corporation undergoes bankruptcy and receivership, the court may appoint someone to head the operation. But managerial autonomy is the rule. The salaried executives have the discretion and authority to decide what products and services they will put on the market, where they will locate

the plants and offices, how they will deal with employees, and whether and in what directions they will expand their spheres of operation. In these respects, corporate businesses are still very much "private" enterprises. Not only does their legal ownership rest with private individuals and institutions, but also the people who run them are presumed to have the right to do so. Government can only regulate business practices by passing legislation on an area-by-area basis; it has no generalized power to interfere with business operations.

Monopolistic competition

The problem of size. The sheer size of the largest U.S. companies has been a subject of controversy since the end of the 19th century. The enactment of the Sherman Anti-Trust Act in 1890 was an outcome of concern over the ability of a few giant firms to dominate markets and diminish competition. Despite this and other related statutes, no significant attempts have been made to break up large corporations. Some have been forced to divest themselves of one or another of their subsidiaries, however, as when Procter & Gamble had to rid itself of Clorox. Likewise, ITT was prevented from acquiring the American Broadcasting Corporation (ABC). If antitrust litigation has avoided broad-scale attacks on corporate bigness, one reason is the difficulty of showing that size implies monopoly or restraint of trade. Even in concentrated fields there are always two or more firms; and it can be argued that competition continues to prevail, albeit among giants. Alcoa, Reynolds Metals, and Kaiser share the aluminum market; but a customer can still choose among them, and they must vie with one another for his business. General Electric, Westinghouse, and Sylvania make 90 percent of the country's light bulbs; and Goodyear, Firestone, and Uniroyal produce almost that proportion of the tires. Many observers believe that the pricing patterns among the leading firms in such industries show a lack of competitiveness; but it has been argued that competition in quality and service may be as important as in price and that consumers often have the option to shift to the products of another industry—for example, from steel to aluminum or from glass to plastics.

Another consequence of large size may be to limit the entry of new firms. It seems unlikely that a new domestic competitor will arise in the U.S. automobile or copper or gypsum industries during the 20th century, because the large investment required would far exceed the potential returns. But an existing corporation may enter a new field of endeavour, as when Sears, Roebuck began selling insurance and IBM began to manufacture typewriters. And there have been cases in which large companies suffer declines. The combined sales of the major meat-packing companies, Swift, Armour, and Wilson, showed no increase between 1954 and 1964, although total industrial sales almost doubled and the nation's per capita meat consumption rose. A similar fate befell the major motion-picture studios, which lost out to competing forms of entertainment, mainly television, and saw themselves joined in their own area by independent production units. It can be argued, then, that there are still opportunities for entry. Between 1960 and 1968 the number of U.S. corporations rose from 1,100,000 to 1,500,000, while the number of proprietorships and partnerships remained relatively static. Some of the new corporations may conceivably rise to challenge the established giants. Among the 50 largest industrial corporations in 1970, 33 had been among the 50 largest 16 years before, indicating both a high degree of continuity and a measure of turnover. But five of the 17 that disappeared from the top had been absorbed by other corporations; their assets, if not their identities, remained in the top ranking.

Role of technology in corporate growth

The causes of vast corporate growth have found varying explanations. One school of thought, most prominently represented by the economist John Kenneth Galbraith, sees growth as stemming from the imperatives of modern technology. Only a large firm can employ the range of talent needed for research and development in areas such as aerospace or nuclear energy. And only companies of this stature have the capacity for innovating industrial processes and entering international markets. Just as government has had to grow in order to meet new responsibilities, so have corporations found that producing for the contemporary economy calls for the intricate interaction of executives, experts, and extensive staffs of employees. While there is certainly room for small firms, the kinds of goods and services that the public seems to want increasingly require the resources that only a large company can master.

Others hold that the optimum size of the efficient firm is substantially smaller than many people believe. George Romney, a former president of American Motors, contended that an automobile company could prosper and be profitable while producing only 200,000 cars a year. By this reasoning, most of the divisions of General Motors could be established as separate companies: Chevrolet does not gain in efficiency because of its corporate association with Pontiac and Buick. Some research has shown that profit rates in industries having a large number of smaller firms are just as high as in those in which a few big companies dominate a market. In this view, corporate expansion stems not from technological necessity but rather from an impulse to acquire or establish new subsidiaries or to branch out into new fields. The structures of most large corporations are really the equivalent of a congeries of semi-independent companies. In some cases, such as Procter & Gamble, these divisions compete against one another as if they were separately owned.

MANAGEMENT IN THE LARGE U.S. CORPORATION

Most large corporations are run by their top executives. Corporate officers have substantial power. In theory these men and women are hired to manage someone else's property; in practice, however, management regards the stockholders as simply one of several constituencies to which it must report at periodic intervals.

Management motivation

Managerial decision making. The guidelines governing management decisions cannot be reduced to a simple formula. Traditionally, economists have assumed that the goal of a business enterprise was to maximize its profits. The difficulty with this guideline is that no one can be sure which measures contribute to overall earnings. A company may spend several hundred thousand dollars on paintings and sculpture for its headquarters office or donate large amounts to charities. Perhaps such expenditures create "goodwill" and help to bring new customers. Executives often conduct their business over luxurious lunches and attend conventions at expensive resorts. It may be argued that lavishness of this sort more than repays itself in establishing contacts, exchanging information, or providing incentive to more effective performance. A case can be made that every activity of a company sooner or later, directly or indirectly, enhances its earnings, but it is best not to push this contention too far. A corporation is a social centre as well as a place of business, and employees at all levels intersperse their working hours with periods in which they waste time or simply enjoy themselves. If managements seek to maximize profits, then their efforts must be tempered by realistic expectations concerning dedication, discipline, and efficiency.

Most corporations turn over only about half of their earnings to stockholders as dividends. They plow the rest of their profits back into the purchase of new capital equipment and the expansion of operations. This is closely related to management autonomy; for, by retaining funds within the company, a corporation's executives maintain resources that enable them to move the firm into fields of their choosing. Some observers have argued that all earnings should be paid out to the stockholders and that a corporation desiring capital funds should be required to approach the investing public each time it wishes to embark on a new venture. Such a procedure would oblige management to make a persuasive case to a constituency outside its own boardrooms.

Whether or not managers seek to maximize profits, they undoubtedly try to maintain them at a high level. Their major motivation is to expand their operations faster than those of their competitors or at a rate exceeding the

growth of the population and the economy. The usual measure of an executive is his ability to augment his company's earnings, which he may do by helping to increase sales or productivity or by achieving savings in other ways. This, above all others, serves to distinguish the atmosphere of business from that of other fields. The purpose of a drug company is not to make pharmaceuticals or to improve the health of society: it exists, first and foremost, to make profits. If it found that it could make more money by manufacturing frozen orange juice, it would do so.

The modern executive. Much has been written about business executives as "organization men." According to this view, the typical corporation manager no longer displays the individualism of earlier generations of entrepreneurs. He seeks protection in committee-made decisions and tailors his personality to please his superiors; he aims to be a good "team" member, adopting the firm's values as his own. There is obviously a germ of truth in this portrayal. Only a minority of managers have unusual qualities of initiative and imagination, and the complexities of the corporate process call for continual conferences in which information is pooled and priorities are determined. Quite clearly, there are companies—and entire industries—that have discouraged innovative ideas. The real question, however, is whether corporations are willing to tolerate autonomy and adventuresomeness, especially among middle-level and younger managers. In many noteworthy cases, management has recognized and rewarded outstanding talents, even when accompanied by abrasive personalities. Certainly, most men and women who rise to presiding positions in large corporations have made some kind of individual record that sets them off from colleagues who simply go through the conventional paces. In not a few instances, this involves wagering one's career prospects on what seems a risk-laden proposition; in others, it consists of stirring subordinates to superior performance. Corporations will continue to attract their share of conformists and mediocrities, but most of these remain in middle management. The real danger is when too much caution and too little creativity emerge at the top.

The managerial elite

Management in the U.S. is an elite, but it is not a hereditary class. Most officers of corporations begin with college degrees, and only in exceptional instances have executives worked their way up from the shop floor. But most managers in the 1960s were the first in their family to experience higher education, and no more than one-third had attended Ivy League institutions. Thus, the typical executive tends to come from modest middle class surroundings; while not wholly "self-made," he must still depend on his own accomplishment if he is to rise in the corporate world. Those who become company presidents are apt to be Republican and Protestant; they are more likely to prefer a suburban home to a city apartment. And, while top executives in the large firms receive salaries and bonuses surpassing $250,000, few retire with a net worth of much over $1,000,000. Most reach the high-paying brackets only toward the close of their careers; and even stock options and deferred compensation do not give them the personal fortunes amassed by many self-employed businessmen. Only a handful can or do turn their positions over to their sons.

The impact of the corporation. While it is generally agreed that the power of corporations extends beyond the economic sphere, this influence is difficult to measure in any objective way. The processes of business entail at least some effort to ensure the sympathetic enactment and enforcement of legislation, since costs and earnings are affected by tax rates and government regulations. Companies and business groups retain lobbyists in the state and national capitals and use vehicles such as advertising to enlist support for policies they favour. While corporations may not legally contribute directly to candidates running for public office, their executives and stockholders may do so as individuals. Corporations may, however, make payments to lobbyists and contribute to committees working to pass or defeat legislative proposals. In practical terms, many lawmakers look upon corporations as part of their constituency, although, if their districts depend on local plants, these lawmakers may be concerned more with preserving jobs than with protecting company profits. In any case, corporations are central institutions in society; it would be unrealistic to expect them to remain aloof from the political process when faced with the prospect of being made to do things they would prefer not to do.

The decisions made by corporation managements ramify throughout society. In effect, corporations can decide which parts of the country will prosper and which will decline by choosing where to locate their plants and other installations. The giant companies not only decide what to produce but also help to instill in their customers a desire for the amenities they make available. To the extent that large firms provide employment, their personnel requirements determine the curriculums of schools and colleges. For these reasons, individuals' personalities are likely to reflect aspirations and dissatisfactions engendered by corporations. This does not mean that large business firms can influence the public in any way they choose; it is simply that they are the only institutions available to perform certain functions. Automobiles, typewriters, telephone service, frozen food, and electric toasters must come from corporate auspices if they are to be provided at all. Given this dependence, corporations tend to create an environment congenial to the conduct of their business.

Public attitudes

The majority of the U.S. population has come to accept the corporation's presence, if only because it plays so vital a role in the nation's life. Few citizens favour the removal of the productive process from private hands. Corporations are occasionally criticized, particularly in times of unemployment, during labour disagreements, and whenever citizens become exercised over the quality of consumer goods or the deterioration of the environment. While these attacks seldom take on political dimensions, they are evidence of some unease over the uses of concentrated power.

The social role of the corporation. Some corporation executives believe that their companies should act as "responsible" public institutions, holding power in trust for the community. Most companies engage in at least some public-service projects and make contributions to charities. A certain percentage of these donations can be deducted from a corporation's taxable income. In 1970, gifts by corporations amounted to approximately $900,000,000, or about 1 percent of total earnings. Most of this money went to private health, education, and welfare agencies, ranging from local hospital and charity funds to civil-rights groups and cultural institutions. Eastern Airlines, for example, contributed to the Metropolitan Opera company and McGraw-Hill to a high school in New York City.

These activities of corporations have been criticized. The Chicago economist Milton Friedman has asserted that corporations should reject the notion that they have public duties. Society as a whole will be better off, Friedman has suggested, if corporations maximize their profits, for this will expand employment, improve technology, raise living standards, and also provide individuals with more money to donate to causes of their own choosing. An accompaniment of this argument is that management has no right to withhold dividends; if stockholders wish to give gifts themselves, they should do so from their personal funds. On the other hand, some critics complain that large companies have been much too conservative in defining their responsibilities. Not only have most firms confined themselves to uncontroversial activities, but also they have sought to reap public-relations benefits from every dollar they donate. Very few, say the critics, have made a real effort to promote employees belonging to minority groups, provide day-care centres, or take on school dropouts and men released from prison. Companies have also been charged with abandoning the central cities, profiting from military contracts, misrepresenting their merchandise, and investing in foreign countries governed by repressive regimes. A perennial indictment has been that profits, prices, and executive compensation

are too high, while the wages and taxes paid by corporations are too low.

Consumer and environmental criticism

In the late 1960s a new school of critics emerged who stressed the social costs of the corporation. They charged that automobiles, pharmaceuticals, and other products were badly designed and dangerous to their users. The new consumer movement, one of the leading figures of which was Ralph Nader, was joined by environmental critics who pointed to the quantities of waste products released into streams or into the air. State and federal laws were passed in an effort to set higher standards of safety and to force companies to install anti-pollution devices. Not all of the critics understand that the costs of these measures would be passed on to the consumer. If a nuclear-power plant must have cooling towers so that it does not discharge heated water into an adjacent lake, the extra equipment results in higher electricity bills. Most companies were hesitant to take such steps on their own initiative, fearing that the need to raise their prices would reduce their sales. But society was already paying for the costs of traffic congestion, trash removal, and nutritional deficiencies. The prices charged by corporations were far from reflecting the total impact that the manufacture and consumption of their products had upon the life of the country. (A.Ha.)

THE LARGE BUSINESS CORPORATION IN EUROPE

International similarities

In many respects the corporation has become a social organization that transcends national differences. As modern business has grown more and more international in outlook, the enterprises by which it is carried on have everywhere come to resemble each other in structure and management. In most countries of western Europe, firms have shown a common tendency to grow larger since the 19th century. The giant firm has become particularly prominent in Great Britain, Germany, The Netherlands, and Belgium. It is less predominant in France, Italy, and Sweden, although its relative importance in France can be seen from the fact that in 1963 some 50 firms accounted for 65 percent of the output of the largest 500 firms. In Belgium in the 1950s, one firm, the Société Générale, was said to own very large proportions of the deposit-banking, insurance, metallurgical, coal, and electric-power industries. In Great Britain in 1955–56, companies comprising only 0.12 percent of all concerns received 45 percent of the gross income.

The tendency toward growth. The process of growth in size went on throughout most of the period from 1870 to 1970, although it was far from being continuous. There were periods of cumulative leaps and periods of reversal. In Germany a wave of amalgamations occurred in the 1920s in the heavy industries, spurred on by foreign investment. After the German defeat in World War II, the victorious Allies attempted to break up the large combines. These later tended to reappear, however, and in 1960 the 100 largest West German firms accounted for 40 percent of industrial output and employed one-third of the industrial labour force. In France, where a large stratum of small firms remained, government policies after World War II attempted to encourage the growth of large firms as well as certain forms of cooperation among them. A unique development was to be seen in Yugoslavia, where the reforms carried out under President Tito in the 1950s and 1960s created large Socialist enterprises nominally owned by the state but run by workers' councils, some of them resembling private Western corporations. One of the largest of these in 1971 was Energoinvest, a confederation of 41 enterprises with 22,000 employees producing electrical equipment, aluminum, various kinds of processing equipment, and a wide range of metals, minerals, and household appliances.

The tendency of large corporations to become larger did not necessarily mean that they became more dominant in their markets or controlled a larger proportion of total assets. In Great Britain and Germany, over a period of several decades, there was no evidence of a long-run trend toward greater monopoly by the largest companies. The degree of concentration in the German iron-and-steel, chemical, and electrical industries was not as great in the 1950s and 1960s as it had been in the 1930s; according to a study by the West German government, in the late 1930s, eight trusts had produced 95 percent of the German output of crude steel and 57 percent of the coal, while in 1958 the eight largest trusts produced 76 percent of the steel and 30 percent of the coal. The expansion of large corporations was partly a result of the growth of the economy as a whole; firms grew larger along with their industries. In part, also, it was the result of efforts by firms to diversify by expanding into a number of different industries. This latter trend was especially pronounced in the decades after World War II, and it produced some mammoth corporations, such as Michelin and Renault in France; Fiat, ENI, and Olivetti in Italy; Volvo in Sweden; Unilever and Courtaulds in Great Britain; and Philips' in The Netherlands, each of which was engaged in a number of different industries.

The business manager. As in the United States, the large European corporations were increasingly controlled by professional managers rather than by stockholders or owners. While relatively more of the large European corporations continued to be controlled by individuals or families, the number seemed to be steadily declining. The divorce of ownership from control, largely the result of the spread of stock ownership among the general public, did not go as far as it had in the United States. In Great Britain, nevertheless, according to a well-known study, the number of large companies in which the 20 largest stockholders owned 30 percent of the total voting stock declined markedly between 1936 and 1951. In other countries similar statistical evidence was not available, but the trend appeared to be the same. Even in companies that continued to be owned by a small group of stockholders or a family, the everyday running of affairs was often the responsibility of managers who acted independently.

Holdings of British managers

These managers were predominantly university-educated men of middle class origin. They seemed to be drawn somewhat more from the upper strata of the middle class than was the case in the United States. Studies made in the 1950s and 1960s showed that in Great Britain nearly half of them had fathers in the highest occupational groups. In France the top managers were recruited largely from the families of businessmen, civil servants, and members of the professions. In Sweden and Denmark, in the 1940s and 1950s, the upper middle class background of managers was even more pronounced. The democratic social climate was thus not reflected in the upper echelons of the large European corporations, and there were even signs that the inequality of selection was increasing rather than decreasing. (Ed.)

THE LARGE BUSINESS CORPORATION IN JAPAN

The family groups

Ownership and control. *The zaibatsu.* Before World War II the large Japanese companies in heavy industry and finance were controlled by holding companies. For the most part, these holding companies were family-owned businesses known as *zaibatsu*. They differed from Western cartels or trusts in two chief ways: usually they were organized around single families, and each *zaibatsu* controlled companies in a wide range of economic activities. The Mitsui, for example, controlled companies in banking, foreign trade, mining, insurance, textiles, sugar, food processing, machinery, and other fields. The four largest *zaibatsu* were Mitsui, Mitsubishi, Sumitomo, and Yasuda. They had all grown up after the Meiji restoration of 1868, expanding rapidly during World War I and afterward.

After the defeat of Japan in World War II, the Allied occupation authorities undertook to dissolve the *zaibatsu* on the ground that their close relations with the military establishment had been a factor in the drive to war. Their stock was put up for sale, and individual companies were freed from the control of the parent organizations. The amount of stock held by *zaibatsu* families in the holding companies at that time was as follows: Mitsui, 64 percent; Iwasaki (Mitsubishi), 50 percent; Sumitomo, 83 percent; and Yasuda, 90 percent. These families also had substantial holdings of the stock of individual companies

in their empires. Some indication of the influence of the *zaibatsu* organizations may be seen in the fact that the four largest of them accounted for about one-quarter of the total paid-up capital in Japanese industry; if the six next largest groups were included, the figure was over one-third.

An antitrust law was enacted, patterned after United States legislation, that prohibited the establishment of holding companies of the *zaibatsu* type, limited stock ownership by banking institutions, and restricted corporate mergers and acquisitions. Japan then entered its long postwar boom, during which the economy grew at a rate surpassing that of any other country. Large enterprises continued to increase their capital investments, but ownership was no longer in the hands of a few individual or institutional stockholders. Even in the cases of such successful postwar enterprises as Matsushita, Sony, and Honda, which were built up by a few talented entrepreneurs, the founders owned approximately 5 percent of the stock in their individual companies in 1970.

The role of the banks. During the period of rapid industrial growth after 1955, the corporations sought to diversify their activities and acquire sophisticated technology. Part of the intensified demand for capital was met through the capital markets; but the sale of securities to the public was far from enough, and enterprises became more and more dependent upon financing by the banks. The importance of banking institutions increased until they supplied as much as three-quarters of the total capital of industrial firms.

Among the big banks there arose a strong competition for new customers, leading them to encourage diversification among their business clientele. The approach of banks to diversification is known among Japanese economists as the one-set doctrine, meaning that each bank endeavours to obtain customers representing a complete set of industries. To the banks, which have invested so heavily in industry, the expansion of corporations offers a chance to increase their capital, with the result that the banks have taken an important part in Japanese industrial expansion.

The banks showed no inclination to use their financial power to take over control of management so long as the performance of the managers was adequate. Toward the end of the 1960s, the divorce of ownership and management had become almost complete among large Japanese corporations.

As for the corporations, they tended to avoid a concentration of control by any one bank or enterprise; stocks were generally held widely by a large number of shareholders, many of them being other corporations in the same group. An example is the giant Nippon Steel company, comparable in size to the United States Steel Corporation. At the end of September 1970 its shares were held by 570,000 persons, less than 1 percent of whom represented financial institutions; but these financial institutions owned 42 percent of the equity capital, while individuals, representing 99 percent of the shareholders, owned only 43 percent of the equity capital.

Another role of the banks in postwar Japanese economic history developed in the 1960s, when concern arose in Japanese business circles that foreign capital might seek to take control of their organizations. To avoid this potential danger, some large firms began to seek ties with banks and other institutional investors so as to keep a large proportion of their stock off the market.

Bantō and "salary men"

Professionalization of management. *The modern role of management.* The characteristics and aspirations of managers underwent a profound alteration in the postwar years. In the prewar industries controlled by the *zaibatsu* families, the management of operations had been delegated to professional managers employed for their skills and abilities. Many of these *bantō* had political connections and were influential in political and economic affairs, although their primary loyalty was to the families that employed them. At the end of World War II they were purged by the occupation authorities and replaced by new men promoted from among the middle-level managers, who felt allegiance to the corporation rather than to any particular master. Most managers in present-day Japanese corporations have risen through the ranks of a single company. Since they do not own substantial amounts of stock, they have come to be known as "salary men." In some of the newer corporations, however, such as Matsushita, Sony, and Honda, management has been retained in the hands of the founders.

There are three main legal organs in a corporation as determined by Japanese commercial law: the stockholders at general meetings, boards of directors, and the company auditors. In practice, however, all three of these have been reduced to a purely nominal status. Meetings of stockholders, as in other countries, are rituals enacted before an empty house, the real decisions having already been made in private negotiations among the large shareholders. The function of the auditors is limited to formal duties only; the actual accounting is carried out by independent certified public accountants, institutionalized after the war in accordance with the U.S. system. The boards of directors consist mainly of company men and rarely include outsiders. As the number of directors has increased with the growth of large companies, functions nominally in the province of the boards are actually carried out by "committees of senior executives," or *jomu-kai*, composed of senior members of the boards. The *jomu-kai* discuss and decide key issues, and, since the general stockholders and the auditors have no substantial influence, the *jomu-kai* are left as the top decision-making bodies. The chairman of the *jomu-kai* is usually also the president of the company, with all the corresponding power and authority.

The guiding objectives of management in the Japanese corporation are too complex to describe in simple terms. Generally, they do not necessarily seek to maximize profits, as did the prewar *zaibatsu*. Their major objectives seem rather to be the maximization of sales volume, the expansion of their corporations, and the increase of their own social prestige, as well as the avoidance of criticism from stockholders, trade unions, and consumers.

Employee advancement. The employees of large Japanese corporations tend to remain with the same employer throughout their working lives, advancing gradually on the basis of seniority. The employees who show superior ability and potential rapidly advance to higher ranks in the company, while those who show less than superior promise advance along standard lines. When the need arises to choose between a man and an organization, in most cases the latter will be protected. Most employees, however, welcoming the advantages of the corporation, devote their lives to the organization and accept their positions in their vertically structured companies.

Company cadets

The corporations recruit young men upon their graduation from the universities and train them as company cadets. Those among the cadets who demonstrate ability and a personality compatible with the organization are later selected as managers. Because of the seniority system, many are well past middle age before they achieve high status. This has led to occasional efforts to get around the seniority system by lifting promising young men out of their low-echelon positions; it has also caused many to urge that less emphasis should be placed upon educational background as a criterion for advancement. Such criticism of the tradition has been stimulated by the example of some of the new postwar corporations and of those owned by foreign capital.

The few men in the Japanese business world who have emerged as personalities are either founders of corporations, managers of family enterprises, or small businessmen. They share a strong inclination to make their own decisions and to minimize the role of directors and boards. But many of these individualists have competent managers to assist them, resulting in a peculiar form of leadership that the Japanese call the "three-legged race."

Industrial concentration. The concentration of ownership in Japanese industry increased in the decades after World War II. In 1967 a study of 586,000 companies by the Fair Trade Commission (excluding the fields of finance and insurance) showed that those having a capital

of less than 5,000,000 yen (360 yen = $1 U.S.; 864 yen = £1 sterling, on April 1, 1964) comprised 86 percent of the total number but had only 7.7 percent of the total capital. On the other hand, those companies with more than 10,000,000 yen capital, comprising 0.2 percent of the total, had 64 percent of the total capital. The 100 largest companies had 36 percent of the total capital. The degree of concentration was similar when total assets and profits were used as the basis for comparison.

The highest concentration was found in the iron-and-steel, shipbuilding, automobile, electrical-machinery, and domestic-appliance industries. Concentration was also high in fishing, chemicals, and shipping. It accelerated in the period from 1955 to 1965 and levelled off or even declined slightly in some areas between 1965 and 1970. The list of companies belonging to the 100 largest changed frequently from 1955 to 1965; for example, of those belonging to the list in 1963, some 39 were replaced by the following year.

Public attitudes toward corporations. By the 1960s and early 1970s the Japanese corporation was beginning to encounter criticism from most sectors of the public, particularly from the younger generation whose goodwill was so important to its future. The criticism was focussed on four kinds of problems: the pollution of air and water by large-scale industry; the destruction of the natural environment; infringement upon the rights of consumers; and the demoralizing effects of work in large organizations.

The problem of pollution

The most urgent problem seemed to be that of pollution. In earlier days, when the scale of industry was relatively small, firms were able to improvise their own solutions, as when, at the close of the 19th century, the Sumitomo copper refinery was removed to Shisaka Island in the Inland Sea, away from populated areas. Similarly, in the second decade of the 20th century, the Hitachi company undertook to construct a smokestack of extraordinary height at one of its mines in order to prevent smoke damage to the surrounding community. But, from the 1930s onward, little was done to counter the damaging effects of industrial expansion, with the result that urban areas began to suffer from air pollution on a large scale. Several new diseases were added to Japanese medical terminology. The Minamata disease was poisoning caused by a mercury compound in factory waste discharged into Minamata Bay in Kyushu and elsewhere. The "Itai-itai disease" was caused by cadmium discharged from refineries in Kamioka and Annaka. The local and national governments were slow to take effective measures against pollution caused by industrial wastes. In the early 1970s a movement was started by persons who had suffered some of its effects and were able to find public support for their cause. They brought legal action against certain companies; they also attempted to take their cases directly to the boardrooms by purchasing stock and attending stockholders' meetings. These efforts, coupled with a campaign by influential newspapers, were beginning to make pollution a political issue in Japan.

The phenomenal growth of Japanese industry also pressed upon the supply of land. The resulting increase in land values was felt in the housing market, as adequate housing became more difficult and expensive to obtain. Many critics argued that investment in social capital had not kept pace with the building of roads, railways, airports, and other land-consuming facilities.

Consumer groups were also beginning to protest against poor quality goods and improper selling practices. In response to their demands, the government's Fair Trade Commission undertook to enforce higher standards in the production and sale of durable goods, such as automobiles and television sets, as well as in soft goods, such as cosmetics, toothpaste, and processed foods.

The effect of large automated factories on those who work in them became the subject of much comment in the Japanese press. Increasing numbers of workers were reported to be alienated from their work and from life, particularly in steel and petrochemical plants, automobile assembling, and the production of home appliances. Some were leaving these industries for smaller ones in the hope of establishing a better working identity for themselves. (K.No.)

BIBLIOGRAPHY

Early history of the corporation, and the corporation in the United States: JOHN P. DAVIS, *Corporations: A Study of the Origin and Development of Great Business Combinations and Their Relation to the Authority of the State*, 2 vol. (1905, reprinted 1961), remains the definitive historical account of the corporate form in the medieval and mercantile periods; while EPHRAIM LIPSON, *The Economic History of England*, 12th ed., vol. 1 and 6th ed., vol. 2–3 (1959–61), provides a full exposition of the joint-stock companies and the chartered companies. LOUIS M. HACKER, *American Capitalism, Its Promise and Accomplishment* (1957), traces the rise of corporations in the U.S. in the 19th century. The divorce of ownership and management is analyzed in ADOLPH A. BERLE and GARDINER C. MEANS, *The Modern Corporation and Private Property*, rev. ed. (1968); and the implications of this development are treated in JOHN KENNETH GALBRAITH, *The New Industrial State*, 2nd ed. rev. (1971). WILBERT E. MOORE, *The Conduct of the Corporation* (1962), focusses on internal operations; while RICHARD J. BARBER, *The American Corporation: Its Power, Its Money, Its Politics* (1970), concentrates on external aspects. OSBORN ELLIOTT, *Men at the Top* (1959), examines the backgrounds and behaviour of company executives. JOSEPH LIVINGSTON, *The American Stockholder* (1958), evaluates the impact of the individuals and institutions that own corporate shares. MILTON FRIEDMAN, *Capitalism and Freedom* (1962), is a plea for non-interference by public agencies; while ESTES KEFAUVER, *In a Few Hands: Monopoly Power in America* (1965), supports further regulation and anti-trust action by government. EDWARD S. MASON (ed.), *The Corporation in Modern Society* (1960); and ANDREW HACKER (ed.), *The Corporation Take-Over* (1964), are collections containing both theoretical analyses and research findings.

Europe: Recent studies of the corporation in European countries are few. Most of them are discussed in MICHAEL M. POSTAN, *An Economic History of Western Europe, 1945–1964* (1967), in which several chapters are devoted to the organization and management of industry. See also P. SARGANT FLORENCE, *Ownership, Control and Success of Large Companies: An Analysis of English Industrial Structure and Policy, 1936–1951* (1961). French policies toward industry are studied in JOHN SHEAHAN, *Promotion and Control of Industry in Postwar France* (1963). An informal survey of managers is DAVID GRANICK, *The European Executive* (1962).

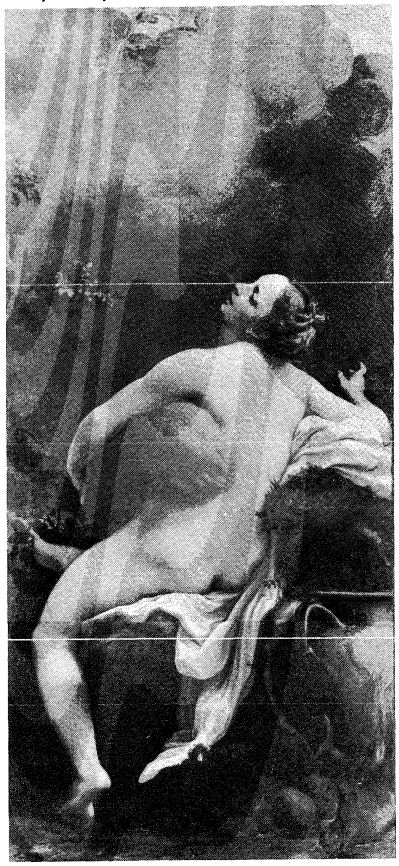
Japan: For the historical development of the Japanese economy, see WILLIAM W. LOCKWOOD, *The Economic Development of Japan*, expanded ed. (1968); and HENRY ROSOVSKY, *Capital Formation in Japan, 1868–1940* (1961), which cover the period from the Meiji restoration to World War II; and RYTARO KOMIYA (ed.), *Postwar Economic Growth in Japan* (1966; orig. pub. in Japanese, 1963), for the postwar period. For a statistical analysis of Japan's economic growth since the 1860s, LAWRENCE KLEIN and KAZUSHI OHKAWA (eds.), *Economic Growth: The Japanese Experience Since the Meiji Era* (1968), will be useful. The characteristics of entrepreneurs in the early Meiji period when modern capitalism rose in Japan are given in JOHANNES HIRSHMEIER, *The Origins of Entrepreneurship in Meiji Japan* (1964). KAZUO NODA, "The Postwar Japanese Executive," in RYTARO KOMIYA (ed.), *op. cit.*, explains the role of business executives in the post-World War II period. Some aspects of Japanese industrial labour are described in EZRA F. VOGEL, *Japan's New Middle Class* (1963); and in a comparative study by ARTHUR M. WHITEHILL, JR., and SHIN-ICHI TAKEZAWA, *The Other Worker: A Comparative Study of Industrial Relations in the United States and Japan* (1968). Regarding the Japanese industrial structure, JOE S. BAIN, *International Differences in Industrial Structure: Eight Nations in the 1950's* (1966), is recommended, although the data are rather outdated now. The present situation of Japan's bureaucratic establishment is described in MARSHALL E. DIMOCK, *The Japanese Technocracy* (1968), one of the few books covering this aspect of Japanese society. For a view of Japan's management system, see MICHAEL Y. YOSHINO, *Japan's Managerial System* (1968); THOMAS F.M. ADAMS and NORITAKE KOBAYASHI, *The World of Japanese Business* (1969); and ROBERT J. BALLON (ed.), *Doing Business in Japan*, 2nd ed. rev. (1968). In addition, JAMES C. ABEGGLEN, *The Japanese Factory* (1958); and SOLOMON B. LEVINE, *Industrial Relations in Postwar Japan* (1958), are considered semiclassics in the management field. For the future of Japan's standing in the world community, see HERMAN KAHN, *The Emerging Japanese Superstate: Challenge and Response* (1970), an optimistic view; and Z. BRZEZINSKI, *The Fragile Blossom* (1972), a pessimistic view.

(A.Ha./Ed./K.No.)

Correggio

One of the greatest masters of painting of the 16th century High Renaissance style in Italy, Antonio Allegri, universally known as Correggio, was in his late works a forerunner of the 17th-century Baroque style. In the 16th century he was already known as "the painter of the Graces," and, although his works were concentrated in Parma and in a few neighbouring Italian cities, he was early considered worthy of being named in the company of such illustrious contemporaries as the artists Raphael, Michelangelo, and Titian.

By courtesy of the Kunsthistorisches Museum, Vienna



"Jupiter and Io," oil painting by Correggio, c. 1530. In the Kunsthistorisches Museum, Vienna. 163.5 × 74 cm.

His father was Pellegrino Allegri, a tradesman living at Correggio, a small city in the territory of Modena. Antonio was born at Correggio, the town that gave him his name, probably not long before August 30, 1494, and died there on March 5, 1534. He was not, as it is often alleged, a self-taught artist. His early work refutes the theory, for it shows an educated knowledge of optics, perspective, architecture, sculpture, and anatomy. His initial instruction probably came from his uncle, Lorenzo Allegri, a painter of moderate ability, at Correggio. About 1503 he probably studied in Modena and then went to Mantua, arriving before the death in 1506 of the famed early Renaissance painter Andrea Mantegna. It has traditionally been said that he completed the decoration of Mantegna's family chapel in the Church of S. Andrea at Mantua after the artist's death. It seems certain the two round paintings, or tondi, of the "Entombment of Christ" and "Madonna and Saints" are by the young Correggio. Although his early works are pervaded with his knowledge of Mantegna's art, his artistic temperament was more akin to that of Leonardo da Vinci (1452–1519), who had a commanding influence upon almost all of the Renaissance painters of northern Italy. Where Mantegna uses tightly controlled line to define form, Correggio, like Leonardo, prefers chiaroscuro, or a subtle manipulation of light and shade creating softness of contour and an atmospheric effect. It is also fairly certain that early in his career he visited Rome and came under the influence of the Vatican frescoes of Michelangelo and Raphael.

Leaving Mantua, Correggio's time was divided between Parma and his hometown. His first documented painting, an altarpiece of the "Madonna of St. Francis," was commissioned for S. Francesco at Correggio in 1514. The best known works of his youth are a group of devotional pictures that became increasingly luscious in colour. They include the "Nativity" (Brera, Milan), "Adoration of the Kings," and "Christ Taking Leave of His Mother."

The Parma murals

Correggio's mature style emerged with his first commission for Parma, the ceiling of the abbess' parlour in the convent of S. Paolo, which was probably executed about 1518–19. Although there are echoes in this work of Mantegna's murals in the Castello at Mantua (1494), it was wholly original in conception. The abbess Giovanna de Piacenza secured for Correggio another important appointment, to decorate the dome of the Church of S. Giovanni Evangelista at Parma. The dome fresco of the "Ascension of Christ" (1520–23) was followed by the decoration of the apse of the same church, of which only the segment entitled "Coronation of the Virgin" survives (Galleria Nazionale, Parma), the remainder having been destroyed in 1587. This work was still in the High Renaissance tradition and owed much to Michelangelo.

The fresco of the "Assumption of the Virgin" in the dome of the Cathedral of Parma marks the culmination of Correggio's career as a mural painter. This fresco (a painting in plaster with water-soluble pigments) anticipates the Baroque style of dramatically illusionistic ceiling painting. The entire architectural surface is treated as a single pictorial unit of vast proportions, equating the dome of the church with the vault of heaven. The realistic way the figures in the clouds seem to protrude into the spectators' space is an audacious and astounding use for the time of foreshortening.

The remainder of Correggio's most famous works, the dates of few known with certainty, fall into three groups: the great altarpieces (and a few other large religious compositions); exquisite small works of private devotion; and a handful of mythological subjects of a lyrically sensuous character. Many of the altarpieces became so well known that they acquired nicknames. The "Adoration of the Shepherds" (*c.* 1530; Germäldegalerie, Dresden) is called "Night" ("La Notte"), and the "Madonna of St. Jerome" (Galleria Nazionale, Parma) is popularly known as "Day" ("Il Giorno"). The late altarpieces are generally characterized by an intimate and domestic mood sustained between idealized figures. This intimate and homely poetry also distinguishes the small devotional works, such as "The Madonna of the Basket" or "The Virgin Adoring the Child Jesus" (Uffizi, Florence), while the "Mystic Marriage of St. Catherine" is a visual essay in the mid-16th-century aesthetic of ideal feminine beauty. In these late works Correggio fully exploited the medium of oil painting. He was intrigued with the sensual beauty of paint texture and achieved his most remarkable effects in a series of mythological works, including the "Danae" (Borghese Gallery, Rome), "The Rape of Ganymede," and "Jupiter and Io" (both in the Kunsthistorisches Museum, Vienna). The sensuous character of the subject matter is enhanced by the quality of the paint, which seems to have been lightly breathed onto the canvas. These pictures carry the erotic to the limits it can go without becoming offensive or pornographic.

Sensual quality of paint

Although his influence can be detected in later Parmese painting, especially in the Mannerist style of Parmigianino (1503–40), Correggio had many imitators but no direct pupils who deserve mention. His decorative ideas were taken up by the Baroque painters of the 17th century, particularly in the ceiling painting of Giovanni Lanfranco (1582–1647), himself a native of Parma. Correggio became almost a tutelary deity of the French Rococo style, and his great altarpieces were among the works most abundantly copied by the travelling artists of the 18th century during their years of study in Italy.

MAJOR WORKS

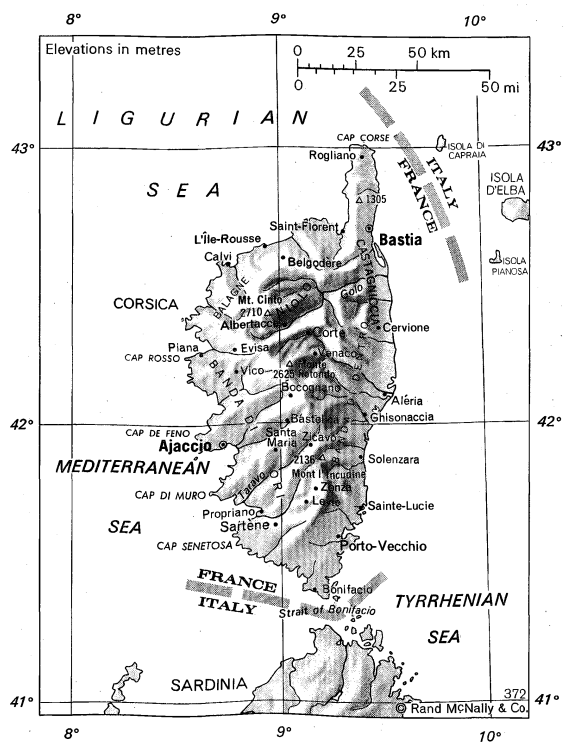
PAINTINGS: "The Mystic Marriage of St. Catherine" (1510–14; National Gallery of Art, Washington, D.C.; Detroit Institute of Arts); "Madonna of St. Francis" (1514; Gemäldegalerie, Dresden, East Germany); "Adoration of the Kings" (*c.* 1514; Brera, Milan); "Madonna" (Art Institute of Chicago); "Four Saints" (1514–17; Metropolitan Museum of Art, New York); "Christ Taking Leave of His Mother" (1514–17; National Gallery, London); "Campori Madonna" (1514–17; Galleria e Museo Estense, Modena); frescoes (1517–20; Camera di S. Paolo, Parma); frescoes (1520–23; cupola of S. Giovanni Evangelista, Parma); "Mystic Marriage of St. Catherine" (1520–26; Louvre, Paris); "The Sojourn in Egypt" (1520–26; Uffizi, Florence); "The Madonna of the Basket" (1520–26; National Gallery, London); "Noli Me Tangere" (1520–26; Prado, Madrid); "Ecce Homo" (1520–26; National Gallery, London); "Madonna of St. Sebastian" (1520–26; Gemäldegalerie, Dresden); "Madonna of St. Jerome," "Il Giorno" (1523; Galleria Nazionale, Parma); "Madonna of the Bowl" (*c.* 1525; Galleria Nazionale, Parma); "Holy Family with St. James" (*c.* 1525; Hampton Court, Middlesex); "Assumption of the Virgin" (1526–30; cupola of Cathedral, Parma); "The School of Love" (*c.* 1532; National Gallery, London); "The Virtues" (*c.* 1530s; Galleria Doria-Pamphili, Rome); "The Virtues" (*c.* 1530s; Louvre); "Leda" (*c.* 1530s; Staatliche Museen Preussischer Kulturbesitz, Berlin).

BIBLIOGRAPHY. C. RICCI, *Correggio* (Eng. trans. 1930), out of date but the fullest available account; A.E. POPHAM, *Correggio's Drawings* (1956), definitive and with catalog; SILVIA DE VITO BATTAGLIA, *Correggio Bibliografia* (1934), definitive bibliography up to 1934: for later works of consequence, see A. BEVILACQUA and A.C. QUINTAVALLE, *L'opera completa del Correggio* (1970), the only book that reproduces everything plausibly ascribed to Correggio (in Italian); E. PANOFSKY, *The Iconography of Correggio's Camera di San Paolo* (1961), the most stimulating book written about Correggio.

(E.K.W.)

Corsica

The fourth largest island in the Mediterranean (after Sicily, Sardinia, and Cyprus), Corsica lies just over 100 miles from the southeast of France, about half that distance westward from the mainland of Italy, and less than ten miles from its southern neighbour, Sardinia. Oval-shaped, it has an area of 3,352 square miles (8,681 square kilometres) and a maximum width of about 50 miles. It is characterized by an almost wholly mountainous territory, more than half of which lies above 1,300 feet.



CORSICA

Corsica (French Corse) became a part of France in 1769 after nearly four centuries of unconstructive colonial rule by the Genoese Republic. In 1789, at the request of its inhabitants, the island was made a French *département*. Yet political integration brought no rapid change in local conditions; living standards lagged well behind those of continental France until after 1957, when significant programs were put in hand to develop the island's most promising sources of revenue: agriculture and tourism.

History. The evolution of Corsica has been shaped by two factors: proximity to prosperous mainland areas and a rugged interior resistant to foreign penetration. The strategic value of Corsican ports was recognized from early times by the rival Mediterranean maritime powers, with the result that the island was often fought over and continually subjected to foreign overlords. Foreign control, however, was concentrated in coastal urban centres: the interior remained backward, while retaining a measure of independence that permitted the development of original social and political institutions.

Strategic importance

Greeks, Etruscans, and Carthaginians had battled for control of the eastern seaboard before the Roman conquest in 259 BC. The Romans built towns and implanted agricultural colonies on the coasts but exercised much less influence over the interior. A sequence of invasions and partial occupations between *c.* 450 and 1050—by Vandals, Lombards, and Saracens—impelled a reversion to the indigenous way of life: the towns were destroyed, their inhabitants driven inland, and the coastal agricultural lands were abandoned, as they remained until modern times.

During the long medieval period of political insecurity and of interrupted contacts with the European mainland, the Corsicans forged an enduring social structure based on the extended family and the rural community. The rural communities, virtually autonomous units, exploited the land through a collective system and elected their own magistrates. These, and other leaders, often assumed hereditary feudal privileges and came to constitute a class of nobles recurrently embroiled in internecine conflict. The struggles for the control of Corsica (between Pisa and Genoa till 1284 and between Genoa and Aragon from 1297 to 1434) were used by the conflicting forces in Corsican society to further their particular interests. In 1358 the nobles in the northeast of the island were overthrown in a popular revolution supported by the Giovannali, heretics in the Franciscan Third Order with communistic ideals. The revolutionary leaders then appealed to the Genoese who, at their request, assumed sovereignty of the island, and who, in their own defense, completed the ruin of the Corsican nobility by the early 16th century.

The rise and fall of the nobility

The Genoese centred their occupation in six coastal fortress towns; though they were corrupt and despotic rulers, they left the rural communities to administer themselves under their elected officers. Village affairs were settled in plenary assemblies: collective exploitation of the land subsisted alongside a gradual increase of private property. Corsican resentment against the Genoese regime facilitated a rapid conquest of the island by Henry II of France in 1553. Popular discontent increased after the island was returned to Genoa in 1559 and persisted throughout the 17th century, while Genoese trade monopolies antagonized a new class of relatively prosperous rural notables.

A national rebellion of 1729–69 was a spontaneous mass movement, directed by the notables, and as a result, the Genoese were driven from the interior almost immediately. Corsican experiments in self-government culminated in a written constitution of 1755. A General Diet, or national legislature, was instituted, with the majority of its members elected by universal suffrage. The Diet elected an executive council, which was responsible to it; Pasquale Paoli, a national leader and author of the constitution, became life-president of the council. The constitution operated until the French annexation in 1769.

After 1769, Corsica was a province of France, except

during a brief period that saw the establishment of the Anglo-Corsican kingdom (1794–96). The ascendancy of Napoleon put an end to separatist movements and consolidated links between continental France and his homeland. In the two World Wars the Corsicans proved themselves conspicuously loyal citizens.

The landscape. The Corsican landscape gives an overall impression of magnitude and untamed exuberance. About two-thirds of the island is composed of an ancient crystalline massif, subsequently worn down and then re-elevated. Dividing the island on a northwest to southeast axis, it rises to a cluster of 20 peaks exceeding 6,500 feet. Mont Cinto attains a height of 8,890 feet (2,710 metres). Lakes, scooped out by Ice Age glaciers, lie at the higher altitudes. The spiked and turreted mountain silhouettes are everywhere dramatic, with the rocks—especially the crimson, rose, and violet granites—varied and vivid. The mountains descend steeply in parallel ranges to the west, where the coast is cut into deep gulfs, with high cliffs and headlands. Rocky islets mark submerged former coastlines offshore. On the east the massif falls in broken escarpments to extensive alluvial plains, and a lagoon-indented coast has, in contrast, risen from the sea. In the northeast, a less spectacular younger formation of rocks has been folded against the ancient crystalline structure. Its rounded summits do not exceed 5,792 feet (1,766 metres). Erosion has cut a depression between the two mountain systems.

Dramatic terrain

The eastern and western watersheds are drained by seasonally torrential rivers that rise in the crystalline centre and cleave their way through impressive gorges in their upper reaches. Soil types, which show multiple local variations, have in common a high acidity caused by calcium deficiency. Lack of humus is general, owing to the scarcity of organic matter and its rapid distintegration in warm, light soils.

The typical sunny Mediterranean climate of the coasts —where the average temperature is 59.9° F (15.5° C) throughout the year, and 51.1° F (10.6° C) in winter—is modified, in the interior, by altitudinal factors. The climate is cooler from 2,000 feet, and, from 3,000 feet, has alpine characteristics: the average summer temperature in coastal Ajaccio is 70.2° F (21.2° C); it is only 64° F (18° C) at Albertacce, at 3,525 feet, where it falls to an average 36° F (2.2° C) in winter.

The prevailing winds are westerly (121 days a year) and northeasterly. Precipitation is heavy, averaging 35 inches (880 millimetres) a year, 80 percent of which falls in storms between November and April, when the island is exposed to cyclonic depressions that form in the Atlantic and the Gulf of Genoa. The area above 3,000 feet, about a third of the whole, receives over 39 inches (one metre) of rain, and is normally snow-covered from December until April.

Vegetation is luxuriant. The characteristic maquis proliferates at low and medium altitudes; it is composed of aromatic shrubs, together with holm oak and cork oak in the south. Chestnut forests appear between 1,600 and 2,600 feet; the trees were planted from the 16th century onward for human and animal nutrition. The laricio pine (*Pinus corsicanus*), which may grow to 160 feet, reigns in the higher forests, interspersed locally with beech and silver birch. Alders are found by the high mountain streams. Forests, in fact, cover 20 percent of the island. The Parc Naturel Régional of 370,000 acres was created in 1971 for the preservation of the natural environment. Boar and foxes are numerous, and about 150 strictly protected moufflons (wild, curly horned sheep) inhabit the crystalline heights.

Several Corsican regions may be characterized on the basis of distinctive local traditions and scenery. The Balagne is covered with ancient olive groves, the Castagniccia, marked by its planted chestnut forests, the Sartenais is a wild country of rocks and holm oak, and the Niolo plateau is the home of sturdy shepherds.

Regional development

Though the coastal plains were freed of malaria in 1943, their development is uneven. While some stretches of the seaboard are being built up to accommodate tourists, others are almost deserted. Only the east coast is continuously cultivated and inhabited: this achievement is largely due to the state-financed Somivac (Société pour la Mise en Valeur Agricole de la Corse), and to 280 or so French families repatriated from North Africa in the years following 1957. New houses are scattered in thriving farmlands where vines—and, with irrigation, citrus fruit—are cultivated for export.

The traditional rural communities are nucleated villages, mostly situated in a band lying between about 650 and 2,600 feet. Emigration has depleted their populations, and 255 out of the 364 Corsican communes had fewer than 300 inhabitants by the early 1970s. Ownership by the communes of a quarter of the island's territory is a survival of the ancient collective system; but there are few remnants of the traditional agriculture of the interior except for small family vegetable gardens and vineyards. Practically no cereals are grown, and olives and chestnuts are rarely harvested.

The coastal towns, Ajaccio and Bastia, each with over 40,000 inhabitants by the 1970s, have drawn to themselves more than a third of the island's population. Without noteworthy industries, they serve as commercial and distributing centres, Bastia for the east and north, Ajaccio—the administrative capital—for the west and south. Sartène, Porto-Vecchio, and Corte have each more than 5,000 inhabitants.

The people. French, the official language, is spoken by virtually all Corsicans, most of whom also use the Corsican dialect, which is thought to have evolved from Latin in the Middle Ages. The majority of the population is Corsican-born. The wide variety of physical types, ranging from tall, fair-haired people to smaller people with dark colouring, is a reminder that Corsica has been repeatedly invaded since prehistoric times. Continental French residents are most numerous in the towns. Since 1957 some 17,000 French citizens from North Africa have found homes in the island, followed by Arab, Spanish, and Portuguese workers constituting a fluctuating labour force of about 8,000. Nearly 10,000 Italians also make a living in Corsica, while other nationalities are represented in small numbers.

Roman Catholicism is the dominant religion; the island forms a single diocese of 150 parishes with a bishop resident in Ajaccio. French Protestants number about 100; a smaller Eastern Catholic congregation, at Cargèse, consists of descendants of Maniate Greeks who settled in Corsica in 1676.

The de facto population of Corsica in the early 1970s was a little over 200,000, about 80,000 less than it was at the end of the 19th century. The decline has been due to massive emigration to mainland France and the French colonies, provoked by the failure of the traditional agricultural economy to support increasing numbers: births have exceeded deaths throughout this period. Corsicans still emigrate at the rate of about 2,000 a year, although many Corsicans retire to the island after careers overseas. As a result, 16 percent of the population is over 64, compared with some 12 percent in France as a whole. The aging of the population has affected the birthrate, which is 1.61 percent, as compared with 1.74 percent for all France, although births still exceed deaths.

The role of emigration

In 1954 the population was down to about 170,000. The ensuing gain of some 50,000 inhabitants was due to the influx of French families from North Africa, accompanied by an economic revival that has attracted foreigners and reduced emigration. The process has been accompanied by a demographic shift from interior to coasts and country to town. The urban population, approximately a third of the whole in 1954, was about two-thirds by the 1970s, and there was a low population density in many rural areas. Corsica's population is still too small to ensure an expanding prosperity. Its increase depends on the economic inducements that can be offered to keep Corsicans in the island and attract non-Corsican residents.

The economy. In spite of striking recent progress, the Corsican standard of living, particularly in the interior, is still somewhat below that of continental France.

Corsica's exploitable natural resources are agricultural;

it contains various minerals, including anthracite and asbestos, but in such small deposits that mining has been abandoned. The only power resources are two medium hydroelectric plants and four small, fuel-powered plants.

The need to import fuel and equipment has prevented significant industrial development, but a canning factory, completed in 1970, is equipped to treat 33,000 tons of fruit and vegetables per annum. The principal Corsican exports are cheese (about 1,900 tons a year); citrus fruit; cork; cigarettes (from homegrown tobacco); and wines and liqueurs, which together represent almost half of the total export value. Imports include cereals, meat, and various fresh foods as well as manufactured goods.

Retirement pensions, state salaries, and social security benefits make up 46 percent of the island's revenue. Agriculture (including stockbreeding) contributes no more than 11 percent, while engaging about 16,000 people full or part time. Only the new farms on the coasts are reliably profitable; most farmers of the interior depend upon pensions or social aid. Traditional agriculture has been ruined by emigration, loss of lives in the two World Wars, failure of markets (olive oil has been undersold by peanut oil), the importation of low-priced cereals, archaic methods, and the division of properties by inheritance (only 34 percent of the Corsican farms exceed 25 acres). Stockbreeding, well adapted to the mountain areas, is in decline. Though the 1,200 owners of some 120,000 sheep make an adequate living selling ewes' milk to the cheese making centres established by the Sociétés de Roquefort, the pastures are critically impoverished by uncontrolled grazing, and the flocks dwindle year by year.

Decline of agriculture

As a tourist resort, Corsica has outstanding assets in its climate, its scenery, and its magnificent coastline. Yet tourism, handicapped by lack of capital, publicity, and experience, provides, at most, 16 percent of the island's revenue. Yearly numbers of holidaymakers increased from 50,000 in 1948 to about 500,000 in 1970; but there is room for many more. The majority come in summer and cluster on the seaboard, and more than three-quarters are French.

The state provides advice and subsidies, and has taken the lead in large-scale investment in the Corsican economy since 1957, through Somivac and through Setco (Société pour l'Équipement Touristique de la Corse). Management of the farms and hotels created by these organizations is in private hands. The example of the state has attracted private capital from overseas.

Taxation is the same as in continental France, except for certain concessions in indirect taxation. A proportion of taxes is allocated to a Fonds d'Expansion de la Corse (Corsican Expansion Fund). In 1968 this sum was equivalent to 10 percent of the normal budget of the *département.*

The small number of French salaried workers minimizes the influence of the trade unions, but increasingly useful functions are performed by 38 recently constituted agricultural cooperatives.

The French market for Corsican wines seems fully assured, and the export of citrus fruit is considered to be capable of a large extension. Agricultural prospects are none the less prejudiced by the small extent of land (about 22 percent of the whole) suited to mechanical cultivation; by a small local consumption, unevenly distributed through the year; and by high transport charges to the mainland. Stockbreeding could certainly provide a larger revenue if methods were modernized as planned. The possibilities of tourism appear almost unlimited, especially if steps are taken to exploit the mountain sites and improve facilities for outdoor and winter sports.

Transportation. There is reliable air and sea transport between Corsica and continental France, both operating with monopoly and state subsidy. Italian shipping lines run service to and from Sardinia and summer services between Bastia, Leghorn, and Genoa. Internal transport is provided by a state-subsidized railway and numerous privately owned road transport companies.

The network of 3,700 miles of tarred roads is adequate in extent—its upkeep engages 12 percent of the island's budget—while a railway links Ajaccio, Bastia, and Calvi, the principal ports, for air and sea travel. Plane and boat services to and from continental France run at a loss in winter but are barely sufficient during the peak summer tourist season.

Administration and social conditions. Corsica is administered by the same system as the *départements* of continental France. Ajaccio is the seat of the prefecture; there are four other *arrondissements.* In January 1970 Corsica was detached from the Région Marseilles-Cote d'Azur-Corse and constituted as a separate *région.*

Political framework

The *Conseil Général* is elected by 62 *cantons;* the 364 *communes* elect municipal administrations. A long tradition of local self-government makes elections acrimonious and sometimes turbulent; yet the Corsicans are moderate in their political views; the Communist party, for example, attracts the support of only about 6 percent of the electorate, in contrast to its strong overall showing in southern France and in Italy. At the start of the 1970s, the majority in the *Conseil Général* was radical, supported by a small Socialist group (SFIO [Section Française de l'Internationale Ouvrière]); the opposition was centrist. Party labels count for less than family allegiances to political power groups and their leaders; but in national elections voters feel less bound by these traditional loyalties; Corsica entered the 1970s with two radical senators, while the three deputies were UDR (Union Démocratique Républicaine) members of the Corsican political minority.

Justice is administered as it is in the continental *départements.* The capital of the judiciary is Bastia, seat of the courts of assizes and of appeal. The military command is stationed in Bastia, while in Ajaccio the army is represented by a small detachment, and the navy by a naval air base. The Foreign Legion, repatriated from North Africa, has a parachutist regiment at Calvi and training units at Corte and Bonifacio. There is a military air base on the east coast.

Educational facilities, under the direction of a *vice-recteur* of the Academy of Nice, had much improved by the early 1970s to meet the aspirations of a people who value professional qualifications: nearly 40,000 students are accommodated in more than 450 primary, and 41 secondary schools and in three technical colleges. A public demand for a university has so far been disallowed because of small numbers of prospective students. Up to 600 students a year enter universities in France.

Thanks to the national social security system and well-equipped medical services (250 doctors, 20 hospitals and clinics, a sanatorium, and a psychiatric hospital), public health is on a level with the national standard.

New homes have been built since 1963 at the rate of 2,000 a year, two-thirds of them in Ajaccio and Bastia, where there is still a demand for inexpensive housing. In the country houses are plentiful but underequipped.

Though Corsica is still underdeveloped in relation to its economic potential, there is no malnutrition and little sign of poverty: more cars per capita are, in fact, owned by the inhabitants than in France as a whole. Corsicans assess their position by middle class criteria; aim at making careers in the professions, the administrative services, or the armed forces; and leave most manual work to foreign labourers. The latter enter the island poor but, because of social security and wages that conform to the national scale, soon improve their conditions.

Cultural life and institutions. Cultural life has suffered from the emigration of the most gifted Corsicans, and there is little contemporary artistic or literary creation. Much intelligent effort is, nevertheless, devoted to the study and conservation of the insular cultural heritage. The *Conseil Général* votes yearly funds from its budget for archaeological research, which is entrusted to qualified experts, enthusiastically followed by the population, and publicized by the Association Archéologique de la Corse. The outstanding megalithic monuments discovered during the 1960s and the remarkable Greek ceramics excavated at Aléria attract many savants and tourists. Learned journals of a high quality are published, including the *Bulletin de la Société des Sciences historiques et*

naturelles de la Corse, founded in 1881. Works on Corsican history find a responsive public.

Corsica has many museums for its size: the Centre de la Préhistoire Corse (Sartène); Musée Jérôme Carcopino (Aléria); a Musée d'Ethnographie Corse (Bastia); a Musée d'Histoire Corse (Corte); and the Musée Pasquale Paoli at Paoli's birthplace, Morosaglia. Souvenirs of Napoleon Bonaparte and his family are exhibited in the Maison Bonaparte and the Musée Napoléonien in Ajaccio, where the Musée Fesch contains a fine collection of old masters assembled by Cardinal Joseph Fesch, Napoleon's step-uncle.

Traditional folk music is performed by groups in the towns and can still be heard in its rural setting, and traditional handcrafts are being revived. The much-studied Corsican dialect is the language of a literary periodical, *U Muntese*. The Maison Itinérante de la Culture de la Corse, founded in 1968, brings classical and modern theatre and concerts, with performers from overseas, to towns and villages and organizes art exhibitions and a summer music and drama festival. Two daily papers in Nice and Marseille carry several pages of Corsican news; and about a dozen papers and periodicals are published in Corsica. Radio and television programs are regularly broadcast from the island.

Problems and prospects. Since the 1960s, Corsica has been emerging with startling rapidity from an archaic, traditional pattern of life. But the transformation is incomplete and insecure: the impetus given by the return of French families from North Africa may be spent, and it is uncertain if the population will continue to increase or the standard of living to rise.

Questions crucial in relation to Corsica's future are whether the coastal monocultures will hold their own in the French market, and, indeed, in the Common Market as a whole, and whether the interior can be revived by a modernization of stockbreeding and a geographical extension of tourism. Continental France, meanwhile, still offers a living to almost every Corsican who emigrates, and Corsican problems appear much less acute when placed in a national rather than in an insular context.

BIBLIOGRAPHY. PAUL ARRIGHI (ed.), *Histoire de la Corse* (1971), is a modern treatment by a team of experts. The following studies illuminate specifically Corsican institutions: A. CASANOVA, "Essai sur la seigneurie banale en Corse," *Etudes Corses* (1959–60); PIERRE EMMANUELLI, *Recherches sur la Terra di Comune* (1958); FRANCIS POMPONI, *Essai sur les notables ruraux en Corse au XVIIème siècle* (1962); and DOROTHY CARRINGTON, "The Corsican Constitution of Pasquale Paoli, 1755–69," in the proceedings of the *International Commission for the History of Representative and Parliamentary Institutions* (1971). JANINE RENUCCI, *Corse traditionnelle et Corse nouvelle* (in prep.), is a comprehensive analysis of Corsican demography and economy. *Corse, Ile d'Elbe, Sardaigne*, 22nd ed. (1968, revised periodically), is a guide (Guide bleu) offering reliable general information.

(D.V.C.)

Cortés, Hernán

Hernán (or Hernando) Cortés, the Spaniard who conquered Mexico early in the 16th century, was—with Francisco Pizarro, conqueror of Peru—the greatest of the Spanish conquistadors. Cortés was 33 years old when he set out in February 1519 to do what two previous Spanish expeditions under Francisco Hernández de Córdoba (1517) and Juan de Grijalva (1518) had failed to do: colonize the mainland of America. By 1521 the Aztec Empire in Mexico had been destroyed; the Aztec priest-king Montezuma (Moctezuma II) was dead; the Aztec capital of Tenochtitlán was in ruins; and Cortés had founded New Spain, the first of the great Spanish American possessions. His conquest, achieved initially with barely 400 Spanish soldiers, represented the peak of his extraordinary career. It set the pattern for Pizarro's later conquest of the Incas of Peru and remains to this day one of the most fascinating and colourful campaigns in history. (For a detailed account of the Spanish conquest of Mexico, see MEXICO, HISTORY OF.)

Early years. Cortés was born in the Spanish town of Medellín in 1485. Thus, like Pizarro, who was vaguely related to him, he was reared in the Estremadura province of Spain at a time when the centuries-old crusade against the Moorish invader was drawing to a close. There, in harsh upland country, both men developed an exceptional hardiness; there, too, they recruited many of their toughest soldiers of fortune. Born to the sword and the saddle, the peasant world of their upbringing combined with the background of religious warfare to give them a blend of greed for gold and a religious fervour that was to carry them with fanatical courage into a new world.



Cortés, oil painting by an unknown artist, 1530. In the Hospital de Jesús, Mexico City.
By courtesy of the Hospital de Jesus, Mexico City

In the little town of Medellín, below its large castle, there are still some traces of the Cortés family home. He was the son of Martín Cortés de Monroy and of Doña Catalina Pizarro Altamarino—names of ancient lineage. "They had little wealth, but much honour," according to Cortés' secretary, Francisco López de Gómara, who tells how, at the age of 14, the young Hernán was sent to study at Salamanca, in west central Spain, "because he was very intelligent and clever in everything he did." Gómara went on to describe him as ruthless, haughty, mischievous, and quarrelsome, "a source of trouble to his parents." Certainly he was "much given to women," frustrated by provincial life, and excited by stories of the Indies Columbus had just discovered. He would have sailed with Nicolás de Ovando, royal governor of the Spanish Indies from 1501 to 1509, but was injured by a falling wall while escaping from the house of a married woman. He then set out for the east-coast port of Valencia with the idea of serving in the Italian wars, but instead he "wandered idly about for nearly a year." Clearly Spain's southern ports, with ships coming in full of the wealth and colour of the Indies, proved a greater attraction. He finally sailed for the island of Hispaniola (now Santo Domingo) in 1504. He was then 19.

Upbringing and study in Salamanca

Years in Hispaniola and Cuba. In Hispaniola, he became a farmer and notary to a town council; for the first six years or so, he seems to have been content to establish his position. He contracted syphilis and, as a result, missed the ill-fated expeditions of Diego de Nicuesa and Alonso de Ojeda, which sailed for the South American mainland in 1509. By 1511 he had recovered, and he sailed with Diego Velázquez to conquer Cuba. Velázquez

was appointed governor, and Cortés clerk to the treasurer. Cortés received a *repartimiento* (gift of land and Indian slaves), and the first house in the new capital of Santiago. He was now in a position of some power and the man to whom dissident elements in the colony began to turn for leadership. This, and his affair with Catalina Juárez, when Velázquez was courting her sister, led to a whole series of farcical episodes—breach of promise, imprisonment, escape into sanctuary, being put in irons on board a ship, a second escape (disguised as a servant boy), swimming the river to sanctuary. In the end, Cortés married Catalina and made his peace with Velázquez.

Appointment from Velázquez to lead an expedition to Mexico

Cortés was twice elected *alcalde* ("mayor") of the town of Santiago and was a man who "in all he did, in his presence, bearing, conversation, manner of eating and of dressing, gave signs of being a great lord." It was therefore to Cortés that Velázquez turned when, after news had come of the progress of Grijalva's efforts to establish a colony on the mainland, it was decided to send him help. An agreement appointing Cortés captain general of a new expedition was signed in October 1518. Experience of the rough and tumble of New World politics advised Cortés to move fast, before Velázquez changed his mind. He went about with a large armed following and had banners that proclaimed: "Brothers and Comrades, let us follow the sign of the Holy Cross in true faith, for under this sign we shall conquer." His sense of the dramatic, his long experience as an administrator, the knowledge gained from so many failed expeditions, above all his ability as a speaker gathered to him six ships and 300 men, all in less than a month. The reaction of Velázquez was predictable; his jealousy aroused, he resolved to place leadership of the expedition in other hands. Cortés, however, put hastily to sea to raise more men and ships in other Cuban ports.

The expedition to Mexico. When he finally sailed for the coast of Yucatán on February 18, 1519, he had 11 ships, 508 soldiers, about 100 sailors, and—most important—16 horses. He had also incurred the undying enmity of Velázquez. As a result, there was an element of disaffection within his fleet. The immediate solution was to send Diego de Ordaz, who commanded a Velázquez ship, away on a foraging expedition. On the mainland Cortés did what no other expedition leader had done: he exercised and disciplined his army, welding it into a cohesive force. But the ultimate expression of his determination to deal with disaffection was when he burned his ships. By that single action he committed himself and his entire force to survival only through conquest.

The key to that conquest lay in the political crisis within the Aztec Empire. The ability of Cortés as a leader is nowhere more apparent than in his quick grasp of the situation—a grasp that was ultimately to give him more than 200,000 Indian allies. He was also fortunate in being presented at the outset with a captured princess who spoke Nahuatl, the language of the Aztecs. She became his "tongue," his adviser on Indian affairs, and, later, his mistress. It was she who worked on the complicated enigmatic mind of Montezuma so subtly that he finally became the voluntary prisoner of her master. Any assessment of Cortés' two campaigns against the Aztecs must take account of his ability to attract and retain the loyalty of this extraordinary woman. She was given the title Doña Marina by the Spaniards, but, to the Aztecs, Cortés and his "tongue" were one—"Malinche."

The role of the captured princess Doña Marina

Spanish politics and envy, however, were to bedevil Cortés throughout his meteoric career. At the outset his problem concerned Velázquez, and he dealt with it by having his newly founded town of Veracruz on the coast of Mexico elect him captain and chief justice—a simple expedient that only a man trained in law and civic affairs would have appreciated. His power then rested on the civic rights of a town under the Spanish crown, not on his appointment by the Governor of Cuba. But this legal safeguard did not prevent Cortés from being threatened by a Spanish force from Cuba, led by Pánfilo Narváez, at a time (mid-1520) when he held the Aztec capital of Tenochtitlán by little more than the force of his personality. He defeated Narváez in a night attack, but his forced march to the coast, leaving his most reckless captain Pedro de Alvarado in control of Tenochtitlán, led directly to the *noche triste* (the nearly disastrous Spanish retreat from the city). Cortés, however, could face and contend with such disaster.

More insidious was the political attack Velázquez mounted in Spain through Bishop Juan Rodríguez de Fonseca and the Council of the Indies. Thus, in 1521, when Cortés had finally destroyed Tenochtitlán, conquering it street by street, canal by canal, and was absolute ruler of a huge territory that extended from the Caribbean to the Pacific, it was the politics and bureaucracy of Spain itself that began the slow, corrosive work of destroying him. Fully conscious of the vulnerability of a successful conqueror whose field of operations was almost 5,000 miles from the centre of political power, he countered with lengthy and detailed dispatches—five remarkable letters addressed to Charles V. His acceptance by the Indians and even his popularity as a relatively benign ruler was such that he could have established Mexico as an independent kingdom. Indeed, this is what the Council of the Indies feared. But his upbringing in a feudal world in which the king commanded absolute allegiance was against it.

The Honduran expedition and last years. In 1524, his restless urge to explore and conquer took him south to the jungles of Honduras. The two arduous years he spent on this disastrous expedition damaged his health and his position. His property was seized by the officials he had left in charge, and reports of the cruelty of their administration and the chaos it created aroused concern in Spain. His fifth letter to the Emperor attempts to justify his reckless behaviour and concludes with a bitter attack on "various and powerful rivals and enemies" who have "obscured the eyes of your Majesty." But it was his misfortune that he was not dealing simply with a king of Spain but with an emperor who ruled most of Europe and who had little time for distant colonies, except insofar as they contributed to his treasury. The Spanish bureaucrats sent out a commission of inquiry under Luis Ponce de León, and, when he died almost immediately, Cortés was accused of poisoning him and was forced to retire to his estate.

Reception in Spain

In 1528 Cortés sailed for Spain to plead his cause in person with the Emperor. He brought with him a great wealth of treasure and a magnificent entourage. He was received by Charles at his court at Toledo, confirmed as captain general (but not as governor), and created Marqués del Valle. He also remarried, into a ducal family. He returned to New Spain in 1530 to find the country in a state of anarchy and so many accusations made against him—even that he had murdered his first wife, Catalina, who had died that year—that, after reasserting his position and re-establishing some sort of order, he retired to his estates at Cuernavaca, about 30 miles south of Mexico City. There he concentrated on the building of his palace and on Pacific exploration.

Finally a viceroy was appointed, after which, in 1540, Cortés returned to Spain. By then he had become thoroughly disillusioned, his life made miserable by litigation. All the rest is anticlimax. "I am old, poor and in debt . . . again and again I have begged your Majesty. . . ." In the end he was permitted to return to Mexico, but he died before he had even reached Seville—on December 2, 1547.

BIBLIOGRAPHY. The three prime sources are Cortés himself, his secretary Gómara, and Bernal Díaz. *Hernando Cortés: Five Letters, 1519–1526*, trans. by J. BAYARD MORRIS (1928), covers the period of conquest and the immediate aftermath and is the most reliable source on facts, dates, and numbers of men involved in battle between these years. BERNAL DIAZ, *Historia verdadera de la conquista de la Nueva España* (first published in 1632; Eng. trans., *The Conquest of New Spain*, 1963), a classic work, covers much the same period, but the author began it when he was 70 years old, and though his memory is astonishing and very vivid, it cannot be compared for accuracy with the *Letters*, which were written at the time and on the spot. FRANCISCO LOPEZ DE GOMARA, *Istoria de la conquista de Mexico* (1553; Eng. trans., *Cortés: The Life of the Conqueror by His Secretary*, 1964), is the only real

source for Cortés' early life and is probably reasonably accurate. The works of later Spanish writers tend to be politically biassed and should all be treated with caution—and particularly the work of Las Casas, although he was contemporary with Cortés. The best academic source remains W.H. PRESCOTT, *History of the Conquest of Mexico*, 3 vol. (1843, with many later reprints), a work so thorough that no scholar or traveller has since been able to fault him other than in very minor details. Of the modern studies, SALVADOR DE MADARIAGA, *Hernán Cortéz, Conqueror of Mexico* (1941), is the most scholarly and most readable; HAMMOND INNES, *The Conquistadors* (1969), covering the whole sweep of Spain's New World involvement, demonstrates the duality of Cortés's conquest of the Aztecs and Pizarro's later conquest of the Incas.

(R.H.I.)

Cosmetics Industry

The cosmetics industry today includes in its scope not only preparations (other than soaps) designed to clean or beautify the face and body, such as are normally associated with the word cosmetics, but also perfumes, dentifrices, and other toilet preparations, some of which are treated in some countries as nonprescription drug products. In general, it may be said that the term is used to cover all products applied to the body that help the user to look or feel more attractive, acceptable, and desirable to others. Such products also deal with the abnormalities caused by normal daily stresses but not with those abnormalities that are the proper field of medicine.

The industry contains several major companies operating internationally, in addition to smaller local manufacturers and distributors. A similarly structured supply industry provides the special chemicals, perfume essences, and packing materials used by the industry. The size of the market tends to be underestimated. It is actually as large as the detergents market, although much more fragmented. In 1970 the world market at trade prices was $2,156,000,000 in the Western world, of which 14 percent was for dental products, 32 percent for hair and shaving products, 40 percent for cosmetics and skin products, and 4 percent for perfumes; of the total some 50 percent was in North America.

The market for cosmetic products

HISTORY

The use of cosmetics is very ancient. Evidence of the use of eye makeup and aromatic ointments has been found in Egyptian tombs dating to 3500 BC. Perfumes of natural origin were greatly prized and hence associated with priestly functions. Oils were used in bathing, possibly because of the drying Mediterranean climate, and this practice was evidently widespread in ancient Greece.

Early cosmetics

By the 1st century AD the Egyptian, Roman, Greek, and Middle Eastern cultures had developed cosmetics such as powders to whiten the skin; kohl to darken the eyelids, eyelashes, and eyebrows; rouge for the cheeks; abrasive products to clean the teeth; perfumed cleansers and unguents; and oils for the bath. The oils used were natural products such as those obtained from the almond and olive; the perfumes were floral or spicy, with natural gums employed as vehicles and fixatives.

Similar cosmetics were used throughout the centuries, developing on lines not very different from those employed today in the 20th century. Cosmetics for the face, dyes for the hair, perfumes for adornment, and environmental health and bath aids were all common in western Europe from the 13th century on. Less developed civilizations also applied the cosmetic art; facial decoration has been associated with both magic and war in cultures as far removed geographically as that of the North American Indian and the indigenous African. The African continent, particularly, developed the art of hairstyling, and the East contributed much to perfumery.

The significant technical developments that form the basis of the modern cosmetics industry, however, are much more recent in origin. In addition to new and improved products, improved packaging and promotional advertising played a role. Among important innovations of the last hundred years are the collapsible tube (1890s); chemicals in hair-waving (1920s); soapless shampoos and the cold permanent wave (1930s); the aerosol container (1940s); improved, less hazardous hair colorants and the fluoride toothpaste (1950s).

COSMETICS AND RELATED PREPARATIONS

Skin-care preparations. The skin consists of a series of outer layers called the epidermis and the underlying true skin, or dermis (see also SKIN, HUMAN). That skin creams and similar products are normally associated with the word cosmetics reflects a change in public attitudes. The basic step in facial care is cleansing, and soap and water is still one of the most effective means.

Cleansing creams and lotions. Cleansing creams and lotions are useful if heavy makeup is to be removed or if the skin is sensitive to soap. Their active ingredient is essentially oil, which acts as a solvent, presented in an emulsion (a mixture of liquids in which one is suspended in droplets in another) with water. Consistency ranges from the thick cold-cream types, made from an emulsion of water suspended in oil, to the relatively thin oil-in-water emulsions called cleansing milks. Cold cream, one of the oldest beauty aids, originally consisted of water beaten into relatively hard mixtures of such natural fats as lard or almond oil. But modern preparations use mineral oil with a suitable emulsifier to aid and maintain the uniform dispersion of oil in water; and their cooling effect is produced by the delayed evaporation of the water.

Emollients. Emollient, or softening, creams; night creams; and nourishing creams are heavier derivatives of the basic cold cream. They are usually formulated to encourage a massaging action in application and often leave a thick film on the face overnight, minimizing water loss during that period when natural metabolism is at work, possibly aided by the creams' special ingredients, such as humectants (moisturizers) and herbal extracts. Among such ingredients, the hormones (substances naturally secreted by the endocrine glands and affecting metabolism) are almost unique in having some proved physiological action. At the low levels at which they must be used for safety, however, their effect is minimal.

Hand creams and lotions. Hand creams and lotions are used chiefly to prevent or reduce the roughness and dryness arising from exposure not only to household detergents but also to wind and dry atmospheres. Like the analogous facial products, they act largely by replacing lost water and laying down an oil film to reduce subsequent moisture loss while natural processes repair the damage. To avoid a grease layer on the hands, such creams are generally creamy, oil-in-water emulsions.

Other skin-care products. Many more specialized products are used for skin care, and particular mention should be made of the products called tonics and fresheners—dilute alcoholic lotions containing some astringent and, especially for use by children and young adults, some germicide usually to deal with the secondary infection of acne. Many products claim some skin-lightening effect, but this is only appreciable when they contain those chemicals that have a real action on the melanin (a dark brown pigment in the skin), such as hydroquinone or related chemicals.

Makeup preparations. *Foundations, face powder, and rouge.* The classic foundation is vanishing cream, essentially an oil-in-water emulsion that contains approximately 15 percent stearic acid (a solid fatty acid), a small part of which is saponified, converting the stearic acid to the crystalline form that provides the accepted sheen on the product. Such creams spread easily and leave no oily finish, although they provide an even, adherent base. Hence, lacking pigment, the cream "vanishes." Variations on this theme are provided by oilier products for dry skin with nonionic emulsifiers and containing colouring matter. Liquid foundations, also oil-in-water emulsions but with less oil, are more heavily pigmented and can replace the classic combination of foundation cream and face powder. Face powder dusted on top of a foundation provides a peach-skin appearance. Many ingredients are needed to provide the characteristics of good powder. Talc helps it slip, or spread easily; chalk or kaolin give it moisture-absorbing qualities; magnesium stearate helps it adhere; zinc oxide and titanium dioxide permit it to

Vanishing cream

cover the skin more thoroughly; and various pigments add colour.

Heightened colour and texture can be provided with rouge, in compressed powder or in cream form, used for highlighting the cheekbones; the more modern version is the blusher, used to blend more colour in the face. Make-up in compressed cake, cream, or gel form and compressed powders are popular as a more or less complete facial makeup and are particularly useful for carrying in the handbag.

Eye makeup. Eye makeup, usually considered indispensable to a complete maquillage (full makeup), includes mascara to emphasize the eyelashes, either in cream form or as a soap-based block; eye shadow for the eyelids, available in many shades; and eyebrow pencils and eyeliner to pick out the edges of the lids. Because eye cosmetics are used adjacent to a very sensitive area, innocuity of ingredients is essential.

Lipstick and lipsalve. Lipstick is an almost universal cosmetic since, together with the eyes, the mouth is a leading feature, and can be attractively coloured, textured, and seemingly changed in size to meet the needs of beauty and fashion. Lipstick has a fatty base that is firm in itself and yet spreads easily when applied. The colour may be provided by pigments—usually reds but also titanium dioxide, a white compound that gives brightness and cover—or by stains that dye the surface of the lips. The stains are always eosin, a red crystalline fluorescent dye, or derivatives of fluorescein, a yellow granular or red crystalline dye. Being only slightly soluble in the wax base, these dyes need a further solvent to render them effective. Because lipsticks are placed on a sensitive surface and ultimately ingested, they are made to the highest safety specifications, considerably restricting the materials available and especially the colours.

Products similar to lipsticks but with little or no colour are also sold as lipsalves to lay a thickish protective coating on the lips against the ravages of wind, sun, and extremely low humidities, or to provide a glossy finish.

Suntan preparations. Further protection by suntan preparations is needed when tanning of the face or body is to be achieved from natural sources without further damage. Products corresponding to most of the cosmetic range, such as vanishing creams, general-purpose creams, and cleansing milks, together with oils and foams, are used to apply the necessary film that also contains a specific sunscreen ingredient, such as ethyl paradimethylaminobenzoate, designed to transmit enough of the sun's spectrum to produce tanning but not the short ultraviolet rays that burn.

Sunscreens

Hair preparations. *Shampoos.* Shampoos must clean the hair adequately and leave it in manageable condition. The use of soap as the sole detergent risks scum production; the basis of retail shampoo is usually an anionic (negatively charged) detergent of good solubility, insensitivity to hard water, and with high grease-removal powers, such as triethanolamine lauryl sulfate in a concentration of about 15 percent. Perfume and additions to control viscosity are essential to an acceptable product, but shampoos may also contain ingredients to meet the special requirements for dry or greasy hair or to combat dandruff. Such products are usually in an opaque or cream form. The firmer varieties can be packed in tubes or jars; other types are bottled.

The job of shampooing is basically that of removing greases, such as sebum and hairdressing, that bind particulate dirt to hair and scalp. This is an easy task for detergents when used in warm water, and the choice of a shampoo usually is based on the secondary requirement of leaving the hair not only clean but as the user requires; *e.g.*, more manageable or freer of dandruff.

Hairdressings, sets, and sprays. Hairdressings, sets, and sprays are useful in grooming when the hair is to be fixed in a desired style, possibly with added gloss and certainly without loss of natural lustre. The aerosol spray is widely accepted by women and permits significant choice of permanence from a very light hold to a virtual varnishing occasionally used for special purposes. Hair sprays are composed of a suitable resin that may be natural (*e.g.*, shellac) or synthetic (*e.g.*, polyvinylpyrrolidone). They also include perfume and possibly a plasticizer, a chemical substance imparting workability, dissolved in a volatile solvent such as alcohol, and delivered in fine-spray form by the aerosol propellant (see below). Hair sprays for use by men were gaining acceptance in the early 1970s.

Before the arrival of the aerosol can, hair was set in position by use of fairly simple gum solutions, usually tragacanth, and these are still employed because of their effectiveness and cheapness; use of "gominas" by men is particularly popular in South America.

Oily dressings, either as brilliantines that are nearly 100-percent oily matter, or as emulsions, usually of water-in-oil, provide an acceptable gloss, particularly on dark hair, and some measure of fixation. The emulsion products also provide the water necessary to make the hair initially more capable of being reformed into the required position. Brilliantines, which cannot do this, are most used by both sexes in hot climates where the hair is straight, as in India, or short and not in need of restyling, as in West Africa. Locally used oils are usually natural (*e.g.*, coconut oil in India), but heavy brilliantines and pomades are based on petroleum oils and jellies. In tropical climates particularly, these brilliantines are also used as a means of positive perfuming.

Brilliantines and emulsions

Alcoholic lotions, especially popular in Europe and South America, provide a fresh stimulus to the scalp and are excellent vehicles for treatment ingredients and perfumes.

Increased attention to the hair results in more work being done on it by combing, setting, bleaching, and colouring; and this, together with environmental pollution and the consequent need for frequent shampooing, makes "out-of-condition" hair a common complaint. This condition relates largely to surface damage to the scales, which can be treated with suitable surface-active compositions normally made up as emulsions for after-shampoo treatments and sold as hair conditioners or strengtheners.

Permanent waving. Permanent waving can be effected in various ways by altering the shape of the hair into the desired configuration under the influence of heat or chemicals that cause an irreversible change. Straight-haired people frequently would like a wave; others with tightly curled hair may want the tightness released to make the hair more controllable. The process is essentially the same. Heat can be used, as in traditional curling processes or hot-comb straightening, but the more common process is to use ammonium thioglycollate to release the bonds in hair by chemical means, permitting it to be reset in its new form by oxidation. This process is also used in milder form to add a small amount of wave to give the hair more lift and body.

Hair colorants and bleaches. Hair colorants and bleaches are used to add "life" to dull or mousy-coloured hair or to cover gray. Dyes were once harsh and drastic and their use confined to theatrical performers and others whose professions required youthful hair colour. Today both permanent colouring, remaining until it grows out, and semipermanent colour, removable after a small number of shampoos, are available for home use. Permanent dyes act by letting a small, colourless, dye precursor (a substance that becomes a dye when polymerized) penetrate the hair and then polymerizing it by oxidation in place, usually with hydrogen peroxide, to a large coloured molecule. Semipermanent types dye directly by incorporating a small-molecule dye of the textile-disperse type in a special shampoo base; some products combine both principles. There are also strictly temporary coloured lacquers and setting lotions that put a coloured film on the hair.

Perfumes and other fragrance products. Perfumes are used in practically every cosmetic and toiletry item, sometimes subsidiary to the main function of the product but often solely to provide fragrance. The fragrance must be formulated for its specific use, and its chemical stability in the particular system must be considered. Various ingredients are used to obtain so-called top

notes, providing the more volatile, immediate odours; middle notes, adding the full, solid character; and base notes, responsible for less volatile, persistent odours. The ingredients making up the fragrance may be of vegetable, animal, or chemical origin. A simple essence, such as for a toilet soap, may have some 20 ingredients and a fine handkerchief perfume, more than 100 (see also OILS, FATS, AND WAXES).

The simplest fragrances are the toilet waters and eau de colognes, which are aqueous alcoholic solutions of low persistence, containing a fairly small proportion of the fragrance-producing ingredients, designed to be reasonably fugitive and fresh in impact. Handkerchief perfumes are more concentrated and essentially alcohol solutions of the essence; the creation of these perfumes is an art form as well as a science, and its leading practitioners are recognized as artists of the highest calibre. Solids and semisolids can also be used as fragrance vehicles. Products of this type include cologne sticks, in which the alcohol is gelled with soap.

HYGIENE AND OTHER PREPARATIONS

Dentifrices. Regular brushing of the teeth is effective in combatting gum problems and limiting bacterial attack on teeth. Dentifrices assist these processes by providing an abrasive, which may be very mild or quite hard, and a detergent; in toothpastes, these are put together with a substance serving as a vehicle and a suitable flavour. Typical formulae include abrasive (*e.g.*, calcium phosphate dihydrate, chalk, alumina), 40 percent; glycerine, 20 percent; detergent (*e.g.*, sodium lauryl sulfate), 1.5 percent; thickener (*e.g.*, sodium carboxymethyl cellulose), 1 percent; flavour, 1 percent; and water.

Fluorides added to toothpaste

A significant advance in oral hygiene was made when, following the discovery of the effect of the fluoride ion in water in reducing dental decay (caries), it was shown to be possible also to reduce the incidence of caries by topical treatment with fluorides (salt of hydrofluoric acid) or fluorophosphates in dentifrice formulations. The efficiency of such products depends very much on the assiduity of use and the skill of the formulator in preserving the activity of the fluorine compound.

Mouthwashes. Mouthwashes are sometimes considered helpful in oral hygiene, although their widespread use is largely confined to the United States. These are usually based on a mild germicidal action, plus some astringency and appropriate flavour.

Antiperspirants and deodorants. A rising consciousness of personal hygiene has been brought about in much of the world by a general improvement in living standards, population density in the cities, and increased promotion and advertising of products intended to improve the impact of the individual on the senses of others. Antiperspirants, deodorants, and deodorant soaps have made especially rapid strides since the introduction of aerosol sprays. The first antiperspirants were dilute alcoholic solutions of aluminum chlorhydrate applied with a sponge, later developed into cream and stick formulae but all suffering from the unaesthetic nature of the operation. Many other materials have been tried, but aluminum derivatives remain the safest and most effective. Deodorants, essentially mild germicides with a good perfume, combat the odours developed by bacterial or enzyme activity from the initially odourless sweat. The combination of an antiperspirant to reduce the excretion of moisture, a deodorant to reduce odour, and perfume is the attraction of the popular sprays and roll-ball applicators.

Bath preparations. Bath preparations have risen rapidly in popularity, with extended use of germicidal soaps, bath oils, and preparations that add perfume and a pleasant feel to the skin. Bath crystals, in the form of the original bath salts (usually perfumed and coloured sodium sesquicarbonate) or as compressed tablets supply an additional benefit in softening the water and reducing scum from the soap. Bubble baths also provide a touch of luxury with their abundant foam, produced by the action of the stream of water from the tap, whipping it up with the aid of a high content of surfactant (a substance aiding in wetting and dispersion) and foam stabilizer. Other foam baths contain specific additives to promote a feeling of well-being.

Talcums, dusting powders, and other after-bath preparations. In hot climates, extensive use is made of talcum and dusting powders as after-bath aids to freshness. The pleasant feel of the powdered skin is also appreciated in temperate climates. In the tropics, alcoholic lotions and colognes are used as much for normal hygiene purposes as for perfume. The difference in the scale of use is vividly displayed in the sizes of the unit containers sold in the tropics compared with those for the same class of product in Europe; *e.g.*, for powders, 16 ounces as against four ounces. These products may also include a small amount of germicide to assist in deodorizing perspiration.

Depilatories. Depilatories are used to remove hair, particularly from the legs and underarms. Removal of underarm hair also permits more effective use of antiperspirant deodorants. The active ingredient in such products is usually calcium thioglycollate made into a cream or paste. This reacts with hair in the same way as do waving lotions, by breaking chemical bonds in the structure. Instead of controlling and then reversing the process, however, as in waving, the depilatory action is allowed to go to completion, weakening the hair so that it can be removed at skin level with minimal effort. Other methods of depilation include application of waxes that set around the hair, pulling the hairs out of the follicle when the wax is pulled off; and, of course, shaving.

Miscellaneous preparations. *Shaving preparations.* Shaving preparations are designed to make the task of shaving easy and pleasant. The beard must be moistened to soften it, and full saturation of a hair with water takes some three to four minutes, so that the shaving soap, cream, or foam must have a consistency or lather to hold water in contact this long. Hence, shaving soaps, sticks, and lather creams produce a tight, slow-to-collapse foam; brushless cream and aerosol creams perform the same task without lathering. The brushless cream is similar to a foundation vanishing cream in being an oil-in-water emulsion containing some 15–20 percent of oily matter, essentially stearic acid. The lather products are always based on stearic-acid soaps, to produce close lather, saponified with a mixture of soda and potash to provide adequate solubility or speed of lather. Aerosol foams are produced by the expansion of propellant in a cream.

Other shaving aids include preshave lotions and powders to prepare the face for dry shaving with electric shavers and after-shave lotions, creams, and powders to provide a smooth, soothed finish to the operation. After-shave lotions are alcoholic, similar to colognes but less strongly perfumed, and may include an additional astringent.

Manicure preparations. The nails, particularly of the hands, also receive care and adornment, and many specialized manicure preparations exist. Chief among these is the nail lacquer, or varnish, and its associated remover. Nail polish is basically a solution of nitrocellulose lacquer, with a plasticizer to make it spread, and appropriate colouring matter. Removers are made from such solvents as acetone and ethyl acetate.

Baby preparations. Preparations used for babies are technically similar to the corresponding adult products, but in a milder and less strongly perfumed form. Avoidance of irritant material is essential, and many adult users buy baby products for the implied characteristics of mildness.

TECHNICAL ASPECTS

Manufacturing. The manufacture of cosmetics and toilet preparations does not generally require heavy capital investment, nor is it excessively labour intensive, and the added value of the resulting products is high in relation to the direct manufacturing costs. Most of the processes are basically those of mixing and blending. Because of the variety of shades, types, and packs necessary to provide the wide choice to the consumer that is intrinsic to the industry, it is usual to employ the batch process, in which a specific quantity is completely pro-

Batch processing

cessed at one time (unlike continuous processing). This type of processing demands ready convertibility and cleaning; cosmetic factories have standards of good housekeeping approaching and often equalling those of the pharmaceutical industry.

Packaging is of the highest importance because the functionality of cosmetic products is indissolubly linked with the unit pack. The most extreme instances of functional packaging are probably the aerosol and the roll-ball, but every product needs its appropriate pack, which, especially for cosmetics carried in the handbag or used at the dressing table, usually requires design of high aesthetic quality, in addition to functional and protective qualities. Fully automated packing lines are used by larger manufacturers for their main lines, particularly for aerosol hair sprays, toothpastes, and shampoos.

Aerosols, since their introduction in 1943 for insecticidal sprays for use by the armed forces of the United States, have shown a phenomenal growth in use for such products as hairsprays, room deodorants, insecticides, and antiperspirants and deodorants or both.

In addition to the good housekeeping required in the plant to produce a clean, uncontaminated product, care must be taken to maintain such conditions throughout the storage and use of the product. Adequate preservatives must always be included in such products if they are not self-preserving.

Evaluation and testing. *Safety testing.* Legislation on cosmetics manufacturing varies among countries, but, in general, no manufacturer will ever deliberately subject his customers to unknown risks. Products must always be tested before sale. The use of chemicals or other ingredients of known hazard is limited or excluded, and products containing such ingredients are subject to labelling regulations.

New products or ingredients are usually screened first on animals. The main tests are for: (1) primary irritation, (2) sensitization, (3) eye damage, (4) acute toxicity, and (5) chronic toxicity. Primary irritation is determined by patching (applying to a small area) on rats, mice, or guinea pigs. Sensitization, a reaction that may be peculiar to the individual, differs from a primary irritation in that the subject becomes sensitized to the chemical, or one closely related to it, by a prior use that may provoke no reaction, while subsequent use provokes an allergic response, involving inflammation, swelling, or reddening not necessarily at the site of application.

Allergic response

For shampoos and other products likely to reach the eyes, animal tests are used to determine the possibility of eye damage. The product is dropped into a rabbit's eye and the nature of any damage or inflammation and rate of recovery after specified periods of time is observed. It is accepted that soap may sting a bit if it goes into the eyes, but there must be no permanent damage; some products of high functionality may carry a warning.

Products destined or likely to be swallowed are assessed for toxicity, employing established standards based on the size of the dose versus the weight of the test animal. There is little difficulty in measuring acute toxicity, although chronic toxicity—the effect of ingesting small amounts over a long time—is more difficult to measure. Some functional products, such as cold permanent-waving solution, are required to carry labels stating "not to be drunk" or "not for internal use."

Assuming that potential hazards have been adequately assessed on prior evidence or animal tests, it is still necessary to confirm that undesired human reactions do not take place. Because of its idiosyncratic nature, reactions produced by sensitization may remain undetected prior to marketing, even when large test panels are used. Skin tests can at least assess the real level of any primary irritation and can determine if widespread sensitization is a risk. Such tests involve application of the product to the skin, usually the forearms, of volunteers, and covering the patch to intensify the reaction. The final challenge is made after a suitable period has elapsed from the initial patch applications.

Despite having previously satisfied himself of the product's safety by pre-market testing, the manufacturer keeps particularly close watch on any consumer complaints during early sales, especially during a test-market operation in a restricted area, by, for instance, alerting hospitals and clinics to report any unusual occurrences. A sensitization that only occurs to one person in 50,000 might never be picked up in human pretests but could be recognized in selling to a population in the millions.

Performance and preference testing. The actual functionality of a product can be tested only by human subjects in use. Performance claims must be verified before they are advertised. Testing may range from relatively simple panel or salon tests to, for example, three-year clinical trials on caries-reducing toothpastes.

Modern preference testing has reached high levels of sophistication. In such highly subjective matters as perfume and flavours statistically based techniques, largely originating in the food industry, are used to establish straightforward preference between like items. In products judged largely on function, there may be an optimum or a maximum performance—for example, a permanent wave can give too tight a curl as well as one that is too loose. In attributes difficult to quantify, mathematical techniques can now handle rank decisions, arranged in order of preference and based on paired comparisons. Modern consumer testing is a substantial part of cosmetics research and development, in addition to actual market research employed to establish facts about the nature of the market as a whole.

LEGAL CONSIDERATIONS

The duty of care is recognized in most legal codes. The duty of the manufacturer to take care not to harm the people at large is reinforced in most countries by more detailed legislation specific to various types of product, controlling some or all of the ingredients, labelling, packaging, and advertising. Some governments maintain regulatory bodies responsible for counselling and enforcement. The most highly developed system is that in the United States, where the Food and Drug Administration (FDA) and the Federal Trade Commission (FTC) between them regulate the fields of drugs, cosmetics, and food products. The FDA requires that industry use certified dye-stuffs for particular uses, and the launching of new toiletry items with therapeutic claims requires heavily documented proof of performance and safety, such toiletries being treated as drugs.

By contrast, the United Kingdom has no specific cosmetic legislation, relying largely on the common law but reinforcing it by the Pharmacy and Poisons Act with negative and restricted lists common to the drug industry and with general laws on claims and performance treated in The Medicines Act 1968 and The Trade Descriptions Act. Unless they make substantial curative claims, toiletries are not treated as drugs.

In the early 1970s the European Economic Community was engaged in harmonizing the separate laws of the member countries, and it seemed likely that legislation would emerge treating toiletries separately from drugs and foodstuffs and incorporating negative and restricted lists of ingredients. Similar patterns apply to many other countries, but there are variable patterns of control at the sales point. In France, for example, a physiologically active product must obtain a "visa" and can be sold only through pharmacies.

Most countries call for special labelling on products that, in order to achieve their performance standards, require ingredients of possible but acceptable hazard—*e.g.*, certain hair colorants that instruct the buyer to perform a patch test before making general application and certain antidandruff shampoos that instruct users to keep the product away from the eyes.

BIBLIOGRAPHY. Comprehensive single-volume reference books in English include R.G. HARRY, *Modern Cosmeticology*, rev. by J.B. WILKINSON *et al.*, 5th ed. (1962); and M.S. BALSAM and EDWARD SAGARIN (eds.), *Cosmetics: Science and Technology*, 2nd ed., 2 vol. (1972), containing a helpful section on the history of cosmetics. M.G. DE NAVARRE (ed.), *Chemistry and Manufacture of Cosmetics*, 2nd ed., 2 vol. (1962), is a comprehensive work. H.W. HIBBOT (ed.), *Handbook of Cos-*

metic Science (1963), produced for students of the Diploma Course of the Society of Cosmetic Chemists of Great Britain, highly recommended as an elementary text. J. STEPHAN JELLINEK, *Kosmetologie*, 2nd rev. ed. (1967; Eng. trans., *Formulation and Function of Cosmetics*, 1970), is a valuable work presenting background and formulation aspects in relation to function in use.

F.V. WELLS and I.I. LUBOWE, *Cosmetics and the Skin* (1964), links the views of cosmetic scientist and dermatologist. In French, EDWIN SIDI and CHARLES ZVIAK, *Problèmes capillaires* (1966), does the same task for hair products and offers excellent chapters on hair coloration. A. HERZKA (ed.), *International Encyclopaedia of Aerosol Packaging* (1966), is one of the most comprehensive texts specializing in aerosols.

Leading trade journals published in English include (all issued monthly): *Cosmetics and Perfumery* (U.S.); *Drug and Cosmetic Industry* (U.S.); *Soap, Perfumery and Cosmetics* (U.K.); and *Manufacturing Chemist* (U.K.). There are similar journals in most countries. An excellent source of original scientific information is the monthly *Journal of the Society of Cosmetic Chemists* (U.S., U.K., and Germany).

(J.B.Wi.)

Cosmic Rays

In the early 1900s investigators discovered that the Earth is continually bombarded by highly energetic very fast-moving particles that are now known to have originated far beyond the atmosphere. These incoming particles are known as "primary" cosmic rays. Interactions between these particles and the atoms and ions of the air, the Earth's magnetic field, recording instruments, and the ground produce "secondary" cosmic rays with much lower energy, which differ from the primaries in composition. The primary cosmic rays consist of atomic nuclei—mostly protons (nuclei of hydrogen atoms) with appreciable numbers of nuclei of other kinds of atoms—and some electrons. Secondary cosmic rays consist mainly of subatomic particles that are short-lived (they change within fractions of a second into other types of particles); they cannot have come far and are thus known to have been produced within the atmosphere.

For many years cosmic radiation, as the only source of high-energy particles, was of critical importance in the discovery of new elementary particles; the unravelling of the complicated chain of interactions initiated by high-energy cosmic rays in the atmosphere was accomplished by the efforts of several generations of scientists. An observer situated within Earth's atmosphere has only an obstructed view of the primary cosmic rays, but the modifying effect of the atmosphere was deduced in some detail by the comparison of data from balloons, mountain observing stations, and others nearer sea level and underground. Cosmic-ray studies, particularly those regarding details of the composition and energy distribution of the primary radiation, have become part of astronomy and astrophysics and have produced results that have contributed to the understanding of supernovae (exploding stars, now regarded as the most probable ultimate sources of the rays), of nuclear synthesis in stars, and of the properties of the Galaxy, the Sun, and the solar system. Investigations of the radioactive elements produced by cosmic-ray bombardment of the Earth, of meteorites, and of the lunar surface have added much information about the history of these objects. The isotopic composition, particularly of the radioactive isotopes, of cosmic rays promises to reveal much about the nature of the cosmic-ray sources and the particle interactions in interstellar space; cosmic-ray neutrinos and solar neutrinos are being studied currently (1970), and it is hoped that they will provide further clues toward new fundamental discoveries.

Significance of cosmic-ray studies to astronomy

HISTORICAL SURVEY

Discovery of cosmic rays. Around 1900, early investigators of natural radioactivity using ionization chambers in which the ionization of a gas by fast, charged particles is measured, found that there is residual ionization even in the absence of all radioactive material. Since such radiation could not be eliminated by heavy shielding, workers hypothesized that the phenomenon was likely to be due to a penetrating radiation of unknown origin. In 1911–12, balloon flights that carried detection instruments to altitudes of about 5,000 metres led to the discovery that the intensity of the radiation drastically increases with altitude. From this it was concluded that the radiation originated outside, or in, the upper atmosphere. From the fact that the radiation persisted day and night with the same intensity, it was concluded that the Sun was not its direct source.

The first decades of analysis. In the 1920s progress came from the work of several investigators. One of these, Robert A. Millikan, was the first to show that the atmosphere acts only as a mass of absorbing material, and is not the producer of primary cosmic rays. He showed that the radiation is absorbed in the water of mountain lakes to the same degree as under an equivalent weight of air at lower altitudes, consistent with an extraterrestrial origin of the cosmic rays. Since at that time the most penetrating radiations known were the gamma rays (electromagnetic radiation of very short wave length), it was assumed that primary cosmic radiation consisted of these gamma rays. Later, it was discovered that the intensity of the rays is less near the Earth's Equator than it is toward the polar regions. This phenomenon can be adequately explained if the radiation is assumed to consist of electrically charged particles that, moving through the magnetic field of the Earth, encounter greater difficulties in reaching Earth near the Equator where the lines of magnetic force are nearly parallel to the surface than they do at the poles where the lines emerge almost vertically.

In the late 1920s, two new research techniques became available. The first cloud chamber pictures (see below) of cosmic rays were made and the use of multiple Geiger counters to record simultaneous pulses from different counters proved that very fast and, therefore, very energetic charged particles must be a dominant constituent of cosmic radiation at sea level. (The extremely short time between the arrival of signals showed that the speed of the particles was enormous.)

During the 1930s scientists developed their understanding of the complicated phenomena that occur in Earth's atmosphere under cosmic-ray bombardment. Ground and balloon observations were extended, and the absorption of cosmic radiation deep underground was measured. Two components of secondary cosmic rays were distinguished. A "hard" component, capable of penetrating deep underground and today known to consist mainly of muons (see below), was separated from a "soft" component, now known to consist of electrons and photons, which could be absorbed in a few centimetres of lead. High-altitude balloon flights established that the total cosmic-ray intensity reaches a maximum at around 12 kilometres above sea level and declines at higher altitudes, showing that many of the particles must be produced well within the atmosphere and demonstrating the secondary nature of much of the cosmic radiation within the atmosphere. The same decade saw the discovery of extensive air showers, large numbers of secondary particles clearly produced by a single primary particle of extremely high energy (greater than 10^{16} electron volts). (An electron volt, symbol eV, is a unit of energy, namely the energy acquired by an electron or proton falling through a difference in potential of one volt. The units kiloelectron volt (keV), megaelectron volt (MeV), gigaelectron volt (GeV) represent 10^3, 10^6, and 10^9 eV, respectively.)

Secondary particles

In the early 1940s it was demonstrated that most primary particles are protons—positively charged hydrogen nuclei. In the second half of the decade photographic emulsions sensitive to nuclear particles were produced; the scintillation counter (see below) was invented; sophisticated electronics systems emerged; and high-altitude balloons became reliable tools for cosmic-ray research. The nuclear emulsion technique, in which ionized tracks formed in the emulsion as a result of cosmic-ray passage through thick photographic emulsion are made visible, led the way to many discoveries, the most important of which was the detection of the pi-meson, or π-meson, and the muon. With the discovery that many π^+, π^- and π^0 mesons are created in high-energy nuclear collisions,

the missing link necessary to account for the large flux of photons and electrons in atmospheric cosmic rays (see below) was found. A large number of other unstable particles were subsequently discovered, most of them through the use of nuclear emulsions, and some in cloud chambers developed to a high degree of sophistication.

Developments after 1950. Unmanned spacecraft, used since the late 1950s, have made possible the investigation of primary cosmic radiation free of terrestrial influences; with the development of solid state electronics the scientist was able to put small, sophisticated instruments into orbit around the Earth and the Sun. There followed the discovery of the Van Allen radiation belts, the configuration of the outer geomagnetic field (the magnetic field that surrounds the Earth), the solar wind, and interplanetary magnetic fields, all phenomena closely connected with cosmic-ray research.

Galactic cosmic rays

In the 1950s cosmic-ray research rapidly evolved into a branch of astrophysics, joining with other branches of astronomy in probing galactic phenomena. It became clear that high-energy particles play an important role in the physics of most astronomical objects (planets, satellites, stars and galaxies, etc.), as well as in interstellar space. The acceleration of particles to high energies, so difficult on Earth, appears to occur readily and frequently in some astronomical environments. The distribution of energy in galactic cosmic rays contains information about the acceleration mechanisms. The elemental composition of the particles carries information about the objects from which they originate, may reveal the degree to which they interact with the interstellar medium, and, indirectly, provides evidence about the length of time the particles spend in the Galaxy.

METHODS OF DETECTION

Principles and types of detecting instruments. *Counters.* The Geiger counter, the oldest of the cosmic-ray-detection instruments, produces an electrical pulse whenever an ionizing particle passes through its gas-filled chamber. Though it played a vital role in early investigations of cosmic rays, it has now been largely superseded by other detectors. Both the scintillation counter and the Cerenkov counter make use of the light produced by the passage of charged particles through transparent materials, and require light-sensing devices, generally photomultipliers—electron tubes that convert weak light signals to electronic signals and amplify them. Scintillation counters work because some transparent materials used in them emit light in proportion to the energy transferred from the ionizing particles, and the output of the photomultiplier yields a measure of this energy. If a particle is stopped within the scintillator, the electric pulse from the photomultiplier is a direct measure of its energy. If a particle penetrates through a slab of scintillation material, its energy loss is proportional to the square of its charge and inversely proportional to the square of its velocity. Measurements of the energy loss and of the total energy together yield the charge as well as the energy of the particle. The method of simultaneous observation of the energy loss and total energy of a cosmic-ray particle is the basis of most composition measurements using particle counters.

The operation of the Cerenkov counter is based on the emission of an electromagnetic-shock wave by a charged particle moving through a transparent medium with a velocity faster than the speed of light in the medium. A mechanical analogue to this process is the bow wave emitted by a boat moving through water at a speed greater than the velocity of the water waves. The Cerenkov light created by the particle moves closely along the path of the particle and can therefore be used to determine the direction of incidence. Many studies of cosmic-ray electrons and gamma-rays rely on the counter's ability to discriminate against particles of low velocity.

The semiconductor detector, a counter device that has found wide application in satellite-borne cosmic-ray detectors, consists of a thin wafer of silicon, made into a solid-state electronic diode. When an ionizing particle traverses the device, the size of the resulting output pulse is proportional to the particle's energy loss by ionization; this type of counter can therefore replace a scintillation counter and photomultiplier. Its small-weight and low-voltage requirements make it extremely useful for space experiments, the only drawback being the limited range of sizes of semiconductors available.

Visual detectors. Foremost in this group is the nuclear emulsion, a supersensitive photographic emulsion in which charged particles produce a track of developable grains, whose density is proportional to the ionization-energy loss of the particle. A measure of the energy loss can therefore be obtained by counting the number of grains per unit length of the track. If the particle should be stopped within the emulsion, its penetration distance, called "range," can also be obtained. The range of a particle in matter is proportional to its mass, inversely proportional to the square of its charge, and depends also on its initial velocity. A further important parameter of the particle track that can be measured in the emulsion is the scattering; *i.e.*, the random deviations of the trajectory from a straight line. Determination of the scattering factors provides a method of identification of mass, charge, and energy of the particle. The mass identification by this method led to the first determination of the contribution of nuclei heavier than hydrogen to the primary cosmic rays. Interaction of the cosmic-ray particle with the constituents of the emulsion can often be observed as a "star" in the emulsion; *i.e.*, a spot from which secondary particles are emitted. These secondary particles are frequently nucleons, but if the interaction takes place at sufficiently high energy, unstable mesons are created. The first observation of the creation of π-mesons, decaying into muons, then further decaying into electrons, was a triumph of the emulsion technique. Subsequently many other unstable particles were discovered with the help of nuclear emulsions, as were, most recently, primary cosmic-ray nuclei of elements between iron and uranium.

Particle tracks

The cloud chamber, a visual detector, the first for charged particles, consists of a vessel containing a gas and saturated vapour (vapour on the point of condensation). The vessel is suddenly expanded, supersaturating the vapour so that droplet formation takes place along the ionizing particle's track, which can then be photographed. Importantly, the cloud chamber may be triggered by counters (see above), giving assurance that it will be expanded whenever it is traversed by cosmic-ray particles. The tracks of cosmic-ray-produced muons and positrons were first observed in cloud chambers, and several unstable elementary particles were discovered. The insertion of plates of heavy material in the chamber allowed observation of nuclear interactions, of electron-photon showers, and of meson showers.

Use of spark chambers

Spark chambers also have applications in cosmic-ray research. These chambers consist of layers of metal plates usually in a mixture of neon and helium gas. After the passage of a charged particle through the assembly, which can be indicated by the triggering of an appropriate counter arrangement similar to that mentioned in the previous paragraph, a fast high-voltage pulse is put on alternate plates, causing sparks between the plates where the ionizing particle has left a track of gas ions. The set of sparks reproduces the particle track and can be photographed. In a digitized spark chamber the metal plates are replaced by grids of wires and sensors that indicate which of the wires has been fired by the spark. This method eliminates the need for an optical system and cameras and yields the information immediately in digital (directly counted) form. It is used in spark chambers on unmanned spacecraft from which the retrieval of film is not possible. Spark chambers employed on high-altitude balloons made possible the discovery of primary cosmic-ray positrons (particles with the mass of an electron but with a positive electric charge).

In 1962 it was found that ionizing particles leave tracks that can be made visible by chemical etching in many insulating solid materials. The importance of the track-detection method to cosmic-ray investigations is twofold. First, it yields information on the composition, energy distribution, and intensity of cosmic rays from past ep-

ochs through the study of old particle tracks in natural crystals from meteorites and from the lunar surface. Second, it is a powerful tool with which to investigate the rare contemporary very heavy cosmic rays through exposure of large areas of plastic sheets on balloons and spacecraft. It is probably the least expensive large-area particle detector. The first observations on such particle tracks were made in natural crystals from meteorites that had been subjected in space to a long bombardment by cosmic rays; these observations yielded the first indication that nuclei of elements much heavier than iron are present in primary cosmic rays. It was soon found that mica, plastics, and glasses are all capable of registering cosmic-ray tracks. Sheets of special plastics sent up in balloons are now used to investigate the heavy nuclei of primary cosmic rays. The properties of the track and subsequent etch pattern depend on the energy loss of the particles, hence distinction between particles of different nuclear charges is possible by methods similar to those employed with nuclear emulsions.

Observational methods. *Investigations at the Earth's surface.* These are restricted to the study of effects caused by primary particles with energies in excess of several times 10^9 electron volts, because atmospheric absorption prevents significant observations of lower energy primary particles. Continuous monitoring over long periods is possible at the surface and can provide a long-term record of cosmic-ray intensity but without discrimination between the various primary components. Recording ionization chambers were first installed in the 1930s and several are still operating. The secondary meson component (the mesons produced as secondary particles after collisions between primary cosmic rays and other particles in the Earth's atmosphere or underground; see above) can be monitored with Geiger counters and scintillation counter telescopes, often placed underground to eliminate effects from lower-energy particles. Lower primary energies can be investigated indirectly with neutron monitors, which record secondary neutrons from the atmosphere. Networks of such monitors exist throughout the world. Some evidence on the energy of primary cosmic-ray particles is obtained by comparing the monitor responses at different geomagnetic latitudes (see below). Cosmic-ray monitors led to the discovery of substantial cosmic-ray intensity variations associated with solar phenomena.

There is about one chance in a thousand that a very high-energy primary cosmic-ray proton will reach a mountain altitude detector without first undergoing a collision with an atom in the atmosphere. Since such high-energy particles cannot be produced by artificial accelerators, high-altitude mountain laboratories are the only sites where they are investigated.

Measurement of air showers

To investigate individual, very rare, cosmic-ray particles with energies exceeding 10^{15} eV would require detectors of unmanageable size. Such particles, however, produce extensive air showers; that is, large numbers of electrons and photons that simultaneously reach points at mountain altitudes or sea level. Studies of large air showers, therefore, extend knowledge of the cosmic-ray spectrum to the highest energies. To investigate the structure and size of the shower, some observers use detectors distributed over many square miles of the surface of the Earth that record the simultaneous arrival of the particles and measure their density. Measurements of extensive air showers have demonstrated the existence of individual cosmic-ray particles with the incredible energy of 10^{19} eV (that is, almost the amount of energy required to raise a mass of one kilogram one metre) concentrated in one nuclear particle. A further important observation available through air-shower measurements is the incidence of high-energy particles in more or less equal numbers from all directions (see below).

Observations from high-altitude balloons, satellites, and space probes. High-altitude balloons have been extensively used in the past decades, and all known primary cosmic-ray components have been discovered with balloon-borne apparatus that carry instruments to a height where there are only two to three grams per square centimetre (g/cm²)—that is, practically no residual atmosphere above the apparatus. The unit gram per square centimetre is used as a measure of the material traversed. The average mass of the entire atmosphere per square centimetre in the vertical direction above a point at sea level is 1,030 grams; this mass decreases by a factor of almost 2 for each 5 kilometres (16,000 feet) of elevation. At 16 kilometres (10 miles), the residual atmosphere is 100 g/cm² and a balloon floating under 2 to 3 g/cm² is 130,000 to 140,000 feet high. The shortcomings of balloon experiments are: (a) the limited time of exposure, normally 10 to 15 hours; (b) the presence of a layer of atmosphere above it in which various undesirable processes can occur. These include the production of secondary particles that may be confused with the primaries and the slowing down or stoppage of very-low-energy primaries. Heavy primary particles may undergo collision and subsequent spallation (flaking or breaking) resulting in a changed chemical composition. The advantages of the balloon technique are: (a) ability to carry relatively heavy payloads; (b) recoverability of the payload permitting reuse of the instrument, an indispensable requirement for all visual detectors; (c) relatively low cost.

Complete freedom from terrestrial influences is provided by Earth satellites in highly eccentric orbits and deep space probes. Instruments on such spacecraft are the source of the most important and reliable information on several aspects of cosmic radiation. Notable among many scientific advances made with satellites are the following: (1) An extension of the knowledge of the primary cosmic-ray energy spectrum to as low an energy as several times 10^6 electron volts, which allows detailed investigations of the solar influence on the cosmic ray flux (rate of flow past a unit of area at the measuring point), an effect that is most pronounced at low energies and provides evidence about electromagnetic conditions in interplanetary space; (2) discovery of the frequent emission of energetic particles by the Sun, associated with other solar phenomena; (3) detailed determination of the chemical composition of primary cosmic rays for elements from hydrogen through iron, which reveal the character of the cosmic-ray sources and also contain clues regarding cosmic-ray propagation in interstellar space; (4) study of the isotopic composition of primary cosmic rays including hydrogen–deuterium, helium isotopes of atomic weights 3 and 4, and beryllium isotopes of atomic weights 7 and 9 or 10; (5) the spatial distribution of the intensity of cosmic rays, both solar and galactic, in the solar system. The intensity of galactic cosmic rays in interplanetary space depends on the distance from the Sun. Space probes flown in the future to Jupiter and beyond may reveal the extent of the volume of space under solar influence and lead to a measurement of the cosmic-ray intensity in interstellar (galactic) space, free of the Sun's influence.

Data from space probes

PROPERTIES OF GALACTIC COSMIC RAYS

Composition. By far the most abundant nuclei in galactic cosmic rays are protons (about 83 percent). Next are helium nuclei with an abundance of around 16 percent. The remainder are heavier nuclei from the entire periodic table of the elements. Figure 1 shows the relative abundance of some low-energy nuclei (measured at 150 MeV/nucleon) from helium to the iron group. This distribution of the elements exhibits some striking features. With some notable exceptions, it greatly resembles the estimated cosmic-abundance distribution of the elements, indicating that cosmic rays originate from objects containing the element distribution consistent with evolution through thermonuclear processes. A characteristic feature of this distribution is the considerably larger abundance of elements with even-charge number compared to those with odd charge, a feature that is clearly displayed in Figure 1 and which is a consequence of the greater stability of atomic nuclei with even charge. The element distribution also demonstrates that cosmic-ray particles have undergone interactions in their travel between source and observer. This is most clearly seen in the relatively high cosmic-ray abundance of the elements

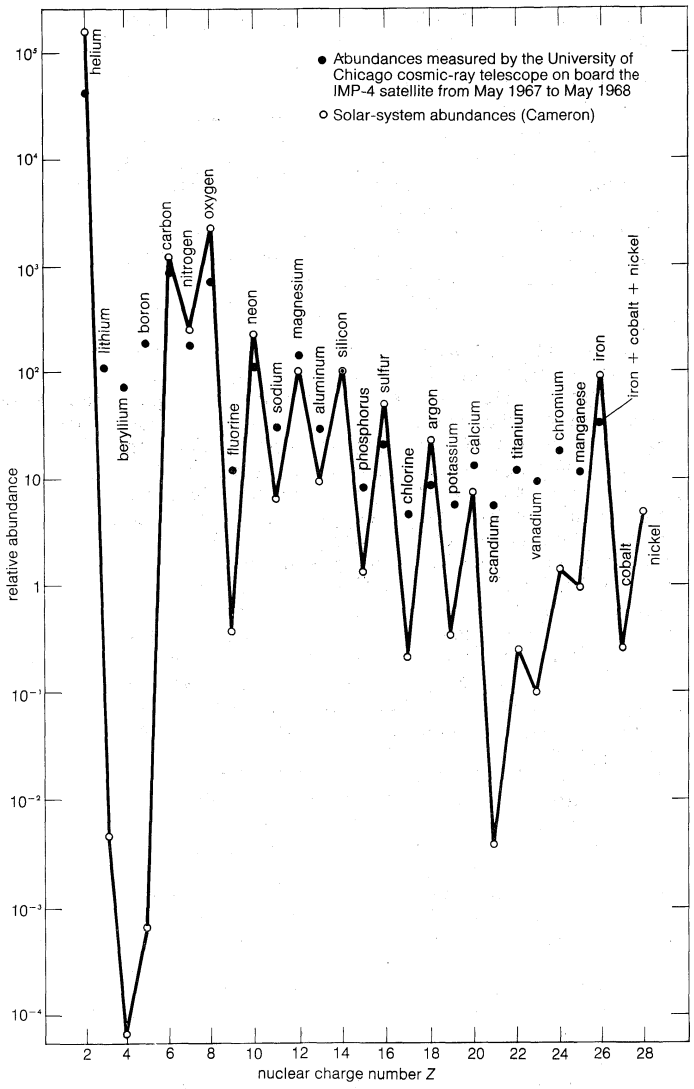


Figure 1: Comparison of the abundance of elements in the galactic cosmic rays at 150 megaelectron volts per nucleon with the "solar system abundances." These abundances are compared with that of silicon, which is given a value of 100. By courtesy of J.A. Simpson and M. Garcia Munoz

Production of lithium, boron, and beryllium in interstellar space

lithium, beryllium, and boron, which are practically absent in stellar objects. These elements must therefore have been produced in interstellar space, and there exists good evidence that they are the product of collisions and subsequent spallation (ejection of the light nuclei), mostly of carbon and oxygen, with particles present in the tenuous interstellar gas. Their abundance has led to the estimate that cosmic rays on the average encounter 3 to 5 g/cm^2 of interstellar matter.

Until 1967 no nucleus heavier than iron was known to exist in the cosmic radiation. The discovery in meteoritic crystals of particle tracks produced by primaries heavier than iron, and subsequent observations of cosmic rays of high nuclear charge, changed this situation. The frequency of these very heavy nuclei is extremely low; for each nucleus with a nuclear charge in excess of 31, 10^4 nuclei in the iron group (nuclear charge around 26) are found. Their presence, however, is important in determining the nature of the source of the cosmic rays. It is most likely that they originate in violent events, such as supernova explosions, in the Galaxy. Though there is interest in whether elements heavier than uranium may be present in the cosmic rays, this has not been proved.

Cosmic-ray electrons and positrons

Cosmic-ray electrons were discovered long after most of the other nuclear components, and the first evidence for their existence came from observation of an emission of radio waves distributed over a wide range of frequencies and having apparently the entire Galaxy as their source. These radio waves were interpreted as being due to high-energy electrons moving in galactic magnetic fields. Electrically charged particles that move on curved paths due to the presence of magnetic fields emit radio waves that become particularly strong if the particle velocity closely approaches the speed of light. This emission is called synchrotron radiation. This evidence for the presence of electrons in cosmic radiation was confirmed in 1961 with the discovery of primary cosmic-ray electrons having an intensity about 1 to 2 percent that of the protons. Later, it was shown that the electron component contains an admixture of around 10 percent positively charged particles (positrons). Positrons (see above) are the antiparticles of electrons, and in the laboratory they quickly combine with electrons in mutual annihilation. In interstellar space, where the density is low, as long as the positron has a high velocity it runs little risk of annihilation and survives for a long time. Cosmic-ray positrons originate in collisions between high-energy cosmic-ray protons and interstellar hydrogen nuclei. This source provides about equal numbers of electrons and positrons, but the observed large electron excess in the composition of cosmic rays may be explained as due to electrons of supernova origin.

Detailed information on the composition of the primary cosmic radiation exists for energies of up to 10^{10} eV per nucleon; beyond that, knowledge is fragmentary.

Energy distribution. One striking feature of the cosmic-ray energy distribution is the similarity of the energy distribution of all nuclear components, suggesting an acceleration mechanism that does not discriminate between them. Figure 2 shows measurements for the two most abundant components, protons and helium nuclei. Below

From *Annual Review of Astronomy and Astrophysics* (1969)

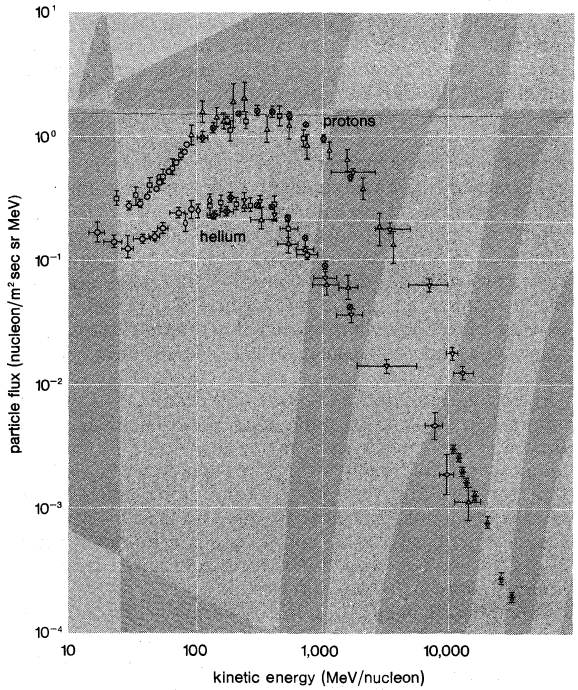


Figure 2: Flux of primary cosmic-ray protons and helium nuclei as a function of their kinetic energy. The different symbols indicate measurements by different workers, and the flux is measured in nucleons per square metre per second per steradian per megaelectron volt.

1,000 MeV/nucleon these distributions do not represent the true interstellar energy spectrum but include the effects of solar modulation. Above that energy, solar modulation has little influence (see below).

The energy spectrum of the total cosmic radiation is known to much higher energies than those of individual components. It follows a "power law" where the flux of particles of energy greater than E is proportional to $E^{-\gamma}$ with $\gamma = 1.6$ up to an energy of 10^{16} eV. Between 10^{16} eV and 10^{18} eV the spectrum becomes slightly steeper

($\gamma = 2.2$) and beyond 10^{18} eV it appears to resume its former slope up to the highest energies, though there is considerable uncertainty about this. Only two particles with energies above 10^{19} eV arrive in a year at every square kilometre of surface, and extensive air-shower measurements are the only tool to observe these particles, which in spite of their scarcity are of great importance as their very existence contains important clues concerning cosmic-ray origins.

Isotropy. An obvious way to identify the source of cosmic rays might be to determine the direction from which the radiation comes. This, however, is not possible since these electrically charged particles are deflected by magnetic fields. In the galactic field, a proton of 1 GeV would move on a circle or spiral with radius of about 10^{12} cm, a length that is extremely small compared to the thickness of the galactic disc (~400 light-years or 10^{20} cm). These particles move along tightly wound spirals around the magnetic field lines. As a result, low-energy cosmic rays arrive from all directions in the Galaxy and give no directional information regarding their origin. Small anisotropies in distribution are found, which have their origin in features of the solar system.

Very high energies

The situation becomes different at higher energies. The path of a proton with 10^{17} eV energy or above has a radius of curvature comparable to or larger than the thickness of the galactic disc. At such energies, if the particles originate in the disc, more of them should appear to come from the general direction of the galactic centre. All experimental results, however, have so far shown an isotropic distribution of incidence. This forces acceptance of the notion that very high energy cosmic rays are likely to be of extragalactic origin.

Energy density. From the observed flux of energy carried by the cosmic rays (~ 10^{-2} erg per square centimetre per second) the energy density of cosmic rays in nearby interstellar space can be calculated to be about 10^{-12} erg per cubic centimetre (erg/cm^3). The amount of energy impinging on the Earth in the form of cosmic rays is therefore almost the same as that of starlight (with the exception of the Sun), a coincidence that seems to have no particular physical significance. More interesting is a comparison with the energy density of the galactic magnetic fields that prevent cosmic rays from leaving the Galaxy along straight-line paths. That their energy density is found to be of the same magnitude as that of the cosmic ray is probably not due to chance. If the fields are capable of holding the cosmic rays in the Galaxy, their energy density must somewhat exceed that of the cosmic rays. Otherwise, the hot "cosmic-ray gas" would blow the field apart.

From the estimate (see above) that cosmic rays traverse about 3 to 5 g/cm^2 of interstellar material in their lifetime and that the density of gas in interstellar space is about 1 to 2 hydrogen atoms per cm^3, it can be deduced that the dwelling time of a cosmic-ray particle in the Galaxy is about 10^6 years on the average. To maintain the energy density of cosmic rays, a continuous input of 5×10^{-26} erg/cm^3-second is therefore required from cosmic-ray sources in the Galaxy.

This estimate is based on the assumption that most cosmic rays originate in the Galaxy. There is an alternate view that cosmic rays exist throughout extragalactic space with a density similar to that observed near the Earth. While such a proposal can not be ruled out experimentally, it encounters considerable difficulties. To maintain a flux of cosmic rays throughout the universe with an energy density of 10^{-12} erg/cm^3 requires the generation of cosmic rays at a rate of 5×10^{-30} erg/cm^3-sec. Since as a consequence of the expansion of the universe intergalactic cosmic rays have to fill an ever-increasing volume of space, their intensity would continuously decline unless they are supplied at the enormous rate mentioned above; a rate very much larger than the energy emitted by extragalactic objects in any form. The present point of view on this question may be summarized as follows: The bulk of cosmic radiation is most likely accelerated within the Galaxy. The observed isotropy of cosmic rays of highest energy indicates, however, that a small fraction of the cosmic radiation has its origin elsewhere.

PROPERTIES OF SOLAR COSMIC RAYS

The Sun is the only stellar object whose emission of cosmic-ray particles can be directly observed at the Earth. In 1942 ground-based cosmic-ray detectors for the first time showed a large, temporary increase in cosmic-ray intensity following a major solar flare. This type of event can be observed once every three to four years, with varying degree of intensity, and is always associated with a solar flare. Clearly, high-energy particles are accelerated and emitted from the solar surface in the flare process.

Solar flare energies

While ground-based detectors are capable of observing the rare events when solar particle energies exceed several GeV, detectors on satellites and space probes permit a study of solar cosmic rays at lower energy and provide evidence that low-energy particle acceleration and emission frequently accompanies a solar flare. Depending on the general level of solar activity, there are one or two solar-flare events per month in which the particles reach energies up to a few hundred MeV.

Solar particle emission offers the possibility of studying in some detail particle-acceleration processes on astronomical scales. Though the detailed mechanism of acceleration is by no means fully understood, it is clear that the large concentration of energy at the flare site must be found in magnetic fields. Altogether, energies of the order of 10^{32} erg can be released in a short time in the form of various radiations. The rapidly changing magnetic fields, including electric fields, are likely to cause the particle acceleration. In a typical solar-flare event, the cosmic-ray intensity observed at the Earth increases to a maximum within 30 minutes to an hour following the visual eruption on the solar surface. The maximum intensity of the solar particles can be many times the value of the normal galactic cosmic-ray flux. In some of the largest events the flux of particles with energy in excess of several GeV has been observed to increase by a factor of 30, and it is not uncommon to see an increase by a factor of 10^3 for particles of about 100 MeV. The occurrence of such high-intensity particle fluxes is of concern for manned space flight and for airplanes travelling at extremely high altitudes. Without the protection of Earth's atmosphere, passengers in such vehicles may be exposed to excessive radiation.

While the maximum intensity of flare particles is reached in an hour or less, the decay in intensity occurs slowly over a day or two, indicating that the particles in interplanetary space must be temporarily stored since there are good reasons to believe that their emission at the Sun occurred within a time span of 10 to 30 minutes. Magnetic fields within the solar system prevent the particles from leaving along straight paths and force them to diffuse slowly out of the interplanetary medium. Further, the particles, while initially coming mainly from the direction of the Sun, will, after reaching maximum intensity, impinge nearly isotropically upon the Earth.

Energy distribution of solar cosmic rays

The energy distribution of solar cosmic rays is considerably more complex than that of galactic particles. Solar cosmic rays show a much steeper energy spectrum, which further steepens as the flare event progresses. This indicates differences in the rate of diffusion in the interplanetary medium for particles of different energy and qualitatively shows, as would be expected, that high-energy particles can leave the solar system more readily than those of lower energy.

By far the most abundant component of solar-flare particles are protons. There exists, however, a significant admixture of heavier nuclei, intensively studied because their composition provides direct evidence regarding the composition of solar matter. This composition closely resembles that of the visible layers of the Sun as determined spectroscopically. The solar abundances of the noble gases, helium and neon, were first determined through solar-particle studies since it is not possible to obtain estimates for them spectroscopically. The composition of solar particles is distinctly different from that of

galactic cosmic rays. Recent evidence has indicated that electrons, also, are frequently ejected during flare events.

Aside from individual flare events, the Sun contributes energetic particles on a more continuous basis. Some of its active regions emit streams of particles with low energy, around 10 to 50 MeV, for several months. This continuous solar production makes it difficult to distinguish between particles of solar and of galactic origin at low energies.

Neutrinos

Extremely interesting solar components are the neutrinos (particles that, since they have no mass or charge, can penetrate undisturbed through large amounts of matter). As solar energy is provided by nuclear reactions, a copious flux of neutrinos must be steadily emitted by the Sun. Attempts have been made to detect this neutrino flux at the Earth. So far, however, the effect of solar neutrinos has not been detected although the sensitivity of the apparatus used should be adequate to do so.

Attempts to detect the emission of neutrons from the Sun have not yet been successful.

THE ORIGIN OF COSMIC RAYS

The Sun is only a minor contributor to the total cosmic-ray flux observed at Earth. Early hypotheses that cosmic radiation was a local solar phenomenon were soon abandoned. As major sources of the high-energy particles, the following possibilities remain: (1) acceleration of cosmic rays in interstellar space, (2) acceleration by stars, (3) violent phenomena in the Galaxy (which appears to be the most promising choice), (4) extragalactic origin.

Interstellar origin. In 1949 Enrico Fermi (best known for his development of controlled nuclear fission) first suggested an ingenious mechanism in which cosmic rays are accelerated in interstellar space by bouncing from moving clouds of ionized gas containing irregular magnetic fields. In this way the particles may gain energy in head-on collisions as a ping-pong ball gains energy from moving paddles. This mechanism can explain the energy distribution of the cosmic rays but is too inefficient to account for all the observed cosmic rays and the required rate of production.

Stellar origin. The flare phenomenon is not restricted to the sun. Stars exist in which flares occur on a much larger scale, but while flare stars are likely to contribute some cosmic rays, they cannot provide all because the energy required is too great.

Supernova origin. The most likely sources of the bulk of the cosmic rays are violent phenomena in the Galaxy, in particular, supernova explosions. The most famous remnant of a supernova is the Crab Nebula, a continuously expanding nebulosity found today where a supernova explosion was observed by Chinese astronomers in the year 1054 AD. Many interesting phenomena observed in the Crab may have a direct bearing on its potential as a high-energy particle source. Estimates of the energy of electrons present in the nebula reach up to 10^{11} or even 10^{13} eV. There is no direct evidence for the acceleration of atomic nuclei in supernova remnants, but it is likely that such nuclei are accelerated with high efficiency. A strong source producing the high-energy electrons is needed in the supernova remnant, to replenish particles that rapidly lose their energy by synchrotron emission. Calculation of the time for an electron to lose about half of its energy by synchrotron radiation reveals that electrons of 10^{11} eV energy live about 100 years if the magnetic field in the Crab is 10^{-3} gauss and the time decreases rapidly for particles of increasing energy. Since these times are considerably below the age of the supernova, electrons must therefore be accelerated in the supernova shell long after the initial explosion. The recent discovery of a pulsar at the centre of the Crab Nebula has revealed a potentially powerful source of energy for particle acceleration. Many if not all supernova remnants may contain pulsars. The properties of pulsars suggest that efficient particle acceleration may take place.

The energy output per supernova, if supernova explosions occur at the rate of one or two per century, can just about suffice to provide the energy input of 5×10^{-26} erg/cm^3-sec required to maintain the galactic cosmic-ray density. It is estimated that on the average, 10^{51} ergs of energy are released by each supernova explosion, an energy production of 10^{42} ergs/sec, or, taking the volume of the galactic disc as 3×10^{66} cm^3, an energy production per unit volume of about 10^{-25} erg/cm^3-sec, which is approximately 10 times larger than the requirement for cosmic rays. If these supernova remnants are the major source of cosmic-ray particles, particle acceleration must be quite efficient.

Extragalactic origin. In spite of its appeal, the supernova hypothesis cannot answer all problems of cosmic-ray origin. For example, a supernova origin cannot account for the isotropic distribution of very high energy particles, 10^{18} to 10^{20} eV. This situation led to the assumption that the sources of very high energy cosmic rays are possibly extragalactic objects. It is likely that 1 percent to 10 percent of cosmic radiation originates outside the Galaxy, in objects where incredibly high fluxes of energetic electrons are present, as shown by observations of the synchrotron radiation (see above).

Possible sources of very high energy cosmic rays

INTERACTIONS OF COSMIC RAYS

Interstellar medium. A cosmic-ray particle spends an average of 10^6 years in the galactic disc before being expelled into extragalactic space. During this time it traverses 3 to 5 g/cm^2 of very dilute interstellar matter, mostly hydrogen, while following a complicated trajectory, spiralling around magnetic lines of force. The cosmic ray nuclei occasionally collide with the nuclei of the interstellar gas. For example, since cosmic-ray protons can, on the average, traverse 60 g/cm^2 hydrogen before colliding, 6 percent of them will undergo a nuclear collision before they can escape from the galactic disc. If the energy of the colliding proton is greater than about 1 GeV, positive pi-mesons (π^+-mesons), negative pi-mesons (π^--mesons), and neutral pi-mesons (π^0-mesons) are created in the collision process. The $\pi^\pm$ mesons decay into muons and neutrinos, and the negative and positive muons, in turn, decay into electrons or positrons, respectively. This latter decay is also accompanied by the emission of neutrinos and antineutrinos. The high-energy electrons and positrons survive for a long time and contribute to the primary cosmic-ray electron component. The collision process appears to be the source of all cosmic-ray positrons observed near Earth. The π^0-mesons decay very rapidly into two gamma rays (γ-rays). The cosmic rays and this γ-ray source should be similarly distributed.

Nuclei heavier than protons have larger cross sections for collision than do protons and the heavier the nucleus, the more probable is a collision. The collision, aside from producing mesons, leads to spallation (see above) of the arriving nucleus after which the fragments proceed with almost undiminished speed. The energy spectrum of the cosmic radiation is not much influenced by the collision process but the chemical composition, as discussed above and shown in Figure 1, certainly is. Most elements, from calcium through manganese, in cosmic rays appear to be produced by the spallation of iron-group nuclei. The chemical composition points to characteristics of the cosmic-ray source and also reveals the importance of nuclear interactions in interstellar space. Magnetic fields in the Galaxy determine the trajectory of the nuclear particles but do not change their energy unless they reach extremely high energies.

Spallation

Electrons, on the other hand, are not much influenced by interstellar matter. Magnetic fields and photons, however, cause large energy losses to cosmic-ray electrons and positrons, as discussed above. The electron energy spectrum observed near Earth therefore will differ from the source spectrum, in ways dependent on the interstellar photon density, on the lifetime of individual cosmic-ray particles in the galactic disc, and on the distribution of cosmic-ray sources. Knowledge of the electron spectrum can provide information on these important quantities and extensive studies have therefore been undertaken.

Cosmic-ray modulation by the Sun. Before approaching Earth, the cosmic radiation is modified by solar influ-

ences. The Sun continuously emits very hot highly ionized gas (about one million tons per second) which moves away from the Sun with an average speed of 400 km/sec. This "solar wind" carries with it solar magnetic fields whose lines of force therefore tend to point radially away from the Sun. Since the Sun rotates, the lines of force are bent into an Archimedes spiral. Superimposed on this general field configuration are local irregularities that also move outward with the solar wind speed. Galactic cosmic radiation is scattered by the irregularities and at the same time continuously swept out by the wind. The net result is a depression of the cosmic-ray intensity in interplanetary space, especially of the low-energy portion of the spectrum. This cosmic-ray modulation manifests itself most clearly in an 11-year intensity variation of cosmic rays, closely correlated to the 11-year activity cycle of the Sun. Cosmic-ray intensity is high in the years of low solar activity, because the solar wind is not blowing so strongly, and low in years of enhanced solar activity. Figure 3 shows this behaviour of the cosmic-ray intensity as measured by a neutron monitor from 1952 to 1969, years of minimum solar activity indicated by arrows. The properties of interplanetary magnetic fields

Solar wind

From J.A. Simpson and J.R. Wang, *Astrophysical Journal*, vol. 161, p. 268 (copyright 1970), University of Chicago Press

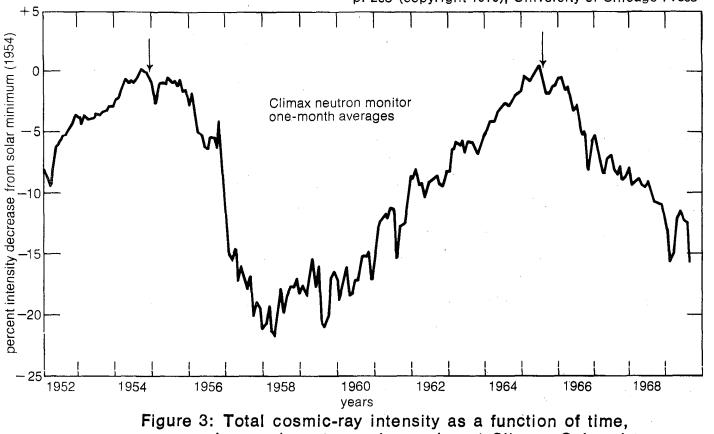


Figure 3: Total cosmic-ray intensity as a function of time, measured over about two solar cycles at Climax, Colorado (altitude 11,800 feet). The ordinate is expressed as a change from the intensity at solar minimum, 1954.

and the solar wind that lead to this modulation were first deduced indirectly through interpretation of cosmic-ray data and related evidence. More recently they have been directly observed using instruments on satellites and deep space probes. The modification of the cosmic-ray energy spectra as observed at the Earth can also be seen in Figure 2. An extrapolation out of the interplanetary medium is necessary in order to picture the interstellar spectra.

Aside from the 11-year cycle intensity variation of cosmic rays other short-term fluctuations are observed. All can be traced to changes of the interplanetary medium induced by the solar wind. Active regions on the Sun occasionally greatly enhance the solar wind.

The geomagnetic field. All ground-based cosmic-ray measurements, balloon observations, and near-Earth satellite work are affected by the geomagnetic field. The first hint of this influence came when the cosmic-ray intensity was found to be lower in the vicinity of the Earth's Equator than at higher latitudes.

Depending on their charge, their direction with respect to the geomagnetic field, and their momentum, cosmic-ray particles are to some degree deflected in the Earth's field. The observed particle flux therefore depends on the location of the observer on the surface of the Earth, and for any given location it depends also on the direction of incidence. Under the simplifying assumption that the Earth's magnetic field may be represented as due to a magnetic dipole at the Earth's centre, the orbits of particles of any momentum or direction of incidence can be calculated. This type of calculation, first carried out in studies of the aurora borealis, has been applied to cosmic-ray particles. The calculations are complex in detail, particularly when realistic models of the geomagnetic field are used and when deformations of the field under the impact of the solar wind are taken into account. Modern computers are used. The most important effects of the geomagnetic field may be summarized as follows. For each geomagnetic latitude (measured from the magnetic poles) there exists a cut-off energy below which vertically incident primary cosmic-ray particles cannot reach the top of the atmosphere. For protons this energy is about 14 GeV at the geomagnetic equator; 4 to 5 GeV in Texas; less than 1 GeV in Minnesota, and of course zero at the magnetic pole. The Earth therefore serves as a magnetic spectrometer for the determination of the energy spectrum of primary cosmic-ray particles and of solar-flare particles. Changes of cosmic-ray spectrum under solar modulation were first discovered using airplanes equipped with cosmic ray detectors that could rapidly move between places of widely different latitude.

Many studies of low-energy cosmic rays with balloon-borne apparatus are carried out at high geomagnetic latitudes, for example, at Ft. Churchill, Manitoba.

PARTICLES DISCOVERED IN SECONDARY RAYS

All known types of elementary particles occur in secondary cosmic rays. They are produced by interactions between extremely fast-moving primary particles and nuclei in the air and by subsequent nuclear interactions or radioactive decay of the initial products. Many particles are found in nature only when created by cosmic rays, because of their instability against radioactive decay and the concentration of energy required to create their masses. Many were first seen among cosmic rays, because high energies were available in cosmic rays before powerful man-made accelerators were built, and because of the exploratory rather than selective nature of cosmic-ray experiments.

Several of the elementary particles first discovered among cosmic rays are:

Production of elementary particles

Positrons. Positrons (see above) are often produced in cosmic rays by pair production, a process in which a high-energy photon interacts with the nuclear field of a nucleus and is absorbed, while an electron and a positron are created, sharing the energy of the photon. A second means of production is in the decay of positive muons, discussed below. Almost as numerous as electrons, positrons constitute about 10 percent of the charged cosmic rays at sea level, and 20 percent to 30 percent at mountain or airplane altitudes.

Muons. The existence of positive and negative muons (formerly called mu-mesons) was proven between 1933 and 1938 by the combined weight of evidence from many experimenters. Muons have the longest ranges of any particles except neutrinos; they constitute three-fourths of the charged cosmic rays at sea level, and are the only charged cosmic rays that penetrate more than a few metres into the ground. Some have been detected after passage through as much as two miles of rock. Some decay in the atmosphere, producing most of the positrons and electrons among the cosmic rays at sea level. For muons of 3 GeV energy, however, the average range before decay is 18 km, and many reach sea level even though they are produced at high altitudes.

Positive and negative π-mesons. Positive and negative π-mesons were discovered in 1947, when a π^+-meson came to rest in a nuclear emulsion and underwent radioactive decay. In the air, both π^+ and π^- mesons decay before being captured. The mean lifetime is 2.56×10^{-8} sec, and the decay products are a muon and a neutrino.

Abundant $\pi^\pm$-mesons are created in cosmic rays by collision of primary and secondary protons and neutrons and other $\pi^\pm$-mesons with atmospheric nuclei; a few are created by the nuclear absorption of photons. When produced in the air, $\pi^\pm$-mesons soon decay, unless their energies are exceptionally high, in which case they are mostly absorbed by interacting with nuclei. As a result, $\pi^\pm$-mesons constitute less than 1 percent of charged cosmic rays at sea level, and their decay is the principal source of the muons appearing in sea-level cosmic radiation.

Neutral π-mesons. The existence of the neutral π-meson was predicted two years before its experimental discovery. Its very short lifetime is followed by decay into two photons, which in turn create electrons and positrons by pair-production. The theory explained two anomalies regarding cosmic rays: the abundance of positrons and electrons at high altitudes despite their scarcity in primary radiation, and their observed creation near nuclear interactions in which charged π-mesons were also created. The mass of the π^0-meson was determined from the energy spectrum of photons in the upper atmosphere, although because π^0-mesons exist so briefly and are neutral, they could not be detected directly in nature but were discovered in 1950 by means of the cyclotron and synchrotron. The mean lifetime of the π^0-meson is about 10^{-16} seconds.

Neutrinos. The decay of $\pi^{\pm}$ mesons and muons in the atmosphere leads to a large flux of neutrinos. Muon neutrinos induce reactions in which further muons are produced. This reaction can, with great care, be used to detect cosmic ray neutrinos.

TRANSITIONS IN THE ATMOSPHERE

Because they are comparatively rare, unstable particles heavier than π-mesons will be neglected here and their brief existence, before decaying into the various components to be discussed, is unimportant to the major transitions in the atmosphere. Antinucleons (antiprotons and antineutrons) will be neglected too, because of their infrequent production.

Fate of heavy primaries. Heavy primary nuclei lose energy by ionization faster than do protons, in proportion to the square of their charge; hence, most of them are stopped near the top of the atmosphere. The more energetic interact with other nuclei and break up into protons and neutrons, which behave thereafter like the more numerous primary protons. The larger the nucleus, the shorter the mean path-length traversed before breakup; it varies from 45 g/cm² for alpha particles to 20 g/cm² for primary silicon nuclei and 14 g/cm² for iron. This explains why heavy primary nuclei are detected only at very high altitudes.

Interactions of primary protons. The typical first interaction of a primary proton in the atmosphere is its collision with a nucleus of nitrogen or oxygen. The immediate effects are: (1) $\pi^{\pm}$- and π^0-mesons are produced, in the approximate ratio of 2 to 1, the total number depending on the primary energy but being typically between 2 and 10; (2) the original proton continues with reduced energy, together with one or two nucleons of the struck nucleus that were given high velocities by the collision; and (3) the residual nucleus is left in an excited state, from which it recovers by emission of slower protons and neutrons in any direction, or by breaking up into deuterons, tritons, and alpha particles all positively charged. These fragments, called evaporation products, typically have energies on the order of 10 MeV.

Absorption of evaporation products. The charged evaporation products lose all their energy in less than 1 g/cm² of air. The neutrons lose no energy by ionization and travel through 5 percent to 10 percent (by weight) of the atmosphere before absorption. They are gradually slowed down by collisions with atoms of oxygen and nitrogen; and with decreasing velocity, their chance of being captured in a collision increases. Most reach energies of about ¼ eV before being captured by a nitrogen nucleus, which then emits a proton and becomes radioactive carbon, carbon-14.

Changes in cosmic-ray composition

Interactions of secondary nucleons. Secondary protons or neutrons of energy very much greater than a GeV behave like primary radiation, undergoing nuclear collisions from which further mesons, fast secondary nucleons, and evaporation products emerge. This step-like multiplication of mesons and nucleons through successive nuclear interactions is known as a nuclear cascade. The average distance between steps is about 1/13 of the vertical atmosphere by weight, or 80 g/cm². The number of steps depends on the primary energy. How low a cascade reaches depends also on chance variations of the distance between steps. The number of cascades still proceeding decreases rapidly as sea level is approached, changing by a factor of 2 in every 80 g/cm², or by a factor of 13 in the last 10,000 feet.

Secondary nucleons of energy less than a GeV usually undergo nuclear interactions that yield only charged nuclear fragments of rather short range, plus neutrons that in turn interact until only evaporation products are emitted.

Fate of positive and negative π-mesons. Primaries of extreme energy, more than 10^{13} eV, may produce $\pi^{\pm}$-mesons of more than 10^{11} eV. The slowing down of time for these particles, in accordance with the theory of relativity, is so great that they may travel far enough to experience nuclear collisions instead of decaying. Like the high-energy secondary nucleons, these mesons contribute to the nuclear cascade, producing further $\pi^{\pm}$- and π^0-mesons and fast secondary nucleons as well as evaporation fragments.

Much more common are $\pi^{\pm}$-mesons of energy less than 100 GeV. These usually decay without further interactions into muons and neutrinos.

Fate of muons. Muons retain about 80 percent of the energy of the $\pi^{\pm}$-mesons, and very seldom lose it by nuclear collisions. Instead, there is a gradual loss of energy by ionization, amounting on the average to about 2 GeV before a muon reaches sea level, since most of them are produced in the uppermost 200 g/cm² of the atmosphere.

Many muons, including a few having energy of tens of GeV, decay during flight through the atmosphere. About two-thirds of the energy and momentum of such a muon is taken away by neutrinos, and the rest by a high-energy electron or positron. Fast muons also generate a minor fraction of the high-energy electrons in the atmosphere by occasional close collisions with electrons belonging to atoms in the air. The particles so produced are called "knock-on electrons."

Most muons of many GeV energy survive, however, until they reach the ground, where they dissipate the remainder of their energy by ionizing atoms along their paths into the Earth. The rate of energy loss is about 2.2 MeV per g/cm², and the density of the Earth's surface is about 2.7 g/cm³; a muon of 20 GeV, for instance, will, thus, penetrate 110 feet before coming to rest.

Fate of neutral π-mesons; cascade showers. The π^0 mesons produced in primary and secondary collisions have such short lifetimes that they travel only negligible distances before decaying, each into two photons, which share the rest energy and kinetic energy of the mesons. Each photon travels on the average only 48 g/cm² in air before undergoing pair-production while passing near an atomic nucleus. The positron and electron thus produced generate new photons along their paths by radiation as they pass near nuclei; in travelling about 40 g/cm², each radiates a photon of about half its own energy, plus numerous photons of lesser energy. The new photons in turn make new pairs of electrons and positrons, and these radiate additional photons. Thus a cascade shower or electronic cascade develops, the average energy per particle decreasing as the number grows. This is the principal means by which secondary photons, electrons, and positrons are created. Ultimately the energies of the photons become too small to make pairs, while the energies of the electrons are too low to permit radiation of further high-energy photons; the remaining photons then lose energy (a reduction in frequency) through interaction with electrons by the process known as Compton scattering and are ultimately absorbed by the photoelectric process, while the energy of the electrons is absorbed by ionization.

Charged particles other than electrons and positrons directly initiate electronic cascades infrequently; the radiation process requires sudden acceleration in the electric field of a nucleus and only electrons and positrons are light enough to be accelerated sufficiently. Their ratio of charge to mass is 207 times greater than that of their nearest competitor, the muon. Cloud chambers and spark chambers are often equipped with lead plates, one of the

chief purposes of which is to make possible the detection of photons and the recognition of positrons and electrons by their unique ability to produce cascade showers.

In a heavy, condensed material such as lead, the mean distance between successive generations (called the "radiation length") is only 0.5 cm; hence, electronic cascades develop rapidly and are usually absorbed within 5 to 10 cm. In contrast, the distance per generation in nuclear cascades or penetrating showers is 14 cm in lead, and the total range is on the order of 1 m. Typical ranges of muons are even longer. Therefore electronic cascades are called "soft showers."

"Soft showers"

Transition curves. Graphs of the counting rates of the various kinds of particles as functions of depth in the atmosphere are called atmospheric transition curves. The transition curve of the total number of charged particles reflects the combined effect of all the processes discussed above.

The maximum in the atmospheric transition curve is due to the multiplication of particles through the mechanisms of multiple meson production, nuclear cascades, and cascade showers. At the top of the atmosphere the number of new particles produced exceeds the number of primaries absorbed, because the average energy is high. Lower in the atmosphere the number of particles capable of the multiplicative processes is much smaller, and absorption predominates over new production. Near the maximum in the transition curve, the most abundant charged particles are positrons and electrons, while photons and slow neutrons are present in comparable number. At the top of the atmosphere the particles are all primary protons, heavier nuclei, and some electrons. Near sea level, the primaries are gone, the soft component has been mostly absorbed, and the most abundant remaining particles are muons.

The relative counting rates (see above) at the top of the atmosphere, at the maximum of the transition curve, and at sea level reflect the total multiplying power of the primaries and the penetrating power of their secondaries. The ratios therefore are indicative of the average primary energy and vary with latitude.

Extensive air showers. An extensive air shower is the most complex of all cosmic-ray phenomena. It occurs when a primary of extremely high energy, 10^{14} eV or more, enters the atmosphere. The primary energy is so high that the cascade creation of particles can go on and on until huge numbers have been produced. All kinds of elementary particles have been found in extensive air showers, but by far the most numerous are photons, electrons, and positrons. In some instances, the number in a single shower exceeds 10^{10}.

Not all particles in air showers proceed in exactly the same direction. Although there is a core near which many are clustered, particles in single air showers have been detected more than 1,000 m apart. The majority, however, land in an area 70 m in radius (at sea level), and within about 10^{-7} sec of each other. Measurement of the energy contained in the biggest showers is a challenging experimental problem, and the acquisition of such energies by submicroscopic particles is difficult to explain.

EFFECTS OF COSMIC RAY BOMBARDMENT

Production of radioactive isotopes. Wherever cosmic-ray particles strike matter, nuclear collisions produce substantial amounts of radioactive isotopes.

Perhaps the most important isotope produced by cosmic rays is carbon-14, created by the capture of slow secondary neutrons in nitrogen-14 nuclei. Carbon-14, with a half-life of about 5,600 years, is employed in the carbon-dating technique that has played an important role in archeological and geological studies. When applied to objects of known age, it reveals that the average cosmic-ray intensity has been very nearly constant for the past 30,000 years. An isotope of hydrogen produced by cosmic radiation in the atmosphere is tritium, which has a half-life of about 12 years and can therefore serve for accurate dating of more recent events. A variety of other cosmic-ray produced radioactive isotopes of widely different half-lives are observed at the Earth. They add only insignificantly to the natural radioactivity of terrestrial origin.

Stable as well as radioactive isotopes of cosmic-ray origin are found in meteorites and on the lunar surface. Their abundance provides insight into the history of the solar system as well as evidence indicating that the average cosmic-ray intensity has been constant within 10 percent for the past 10^7 years and within a factor of 2 for the past 10^9 years.

Average cosmic-ray intensity

Cosmic ray exposure age. The effective duration of cosmic-ray bombardment of a meteorite, called the cosmic-ray exposure age, can be measured by determining the concentration and production rate of both radioactive and stable nuclides it has produced. Cosmic-ray bombardment is assumed to begin at a specific time, at the breakup of a mass of material sufficiently large to shield most of its volume from cosmic radiation. This assumption seems to be borne out since the exposure ages of many iron meteorites, for example, cluster around 630 and 900×10^6 years, indicating that the parent bodies of these meteorites broke up in catastrophic collisions at those specific times in the past. Other ages found range from 1.5×10^8 to 2×10^9 years. These investigations have provided information regarding the origin of meteorites and of the solar system. Study of lunar material promises to provide data on the history of the Moon.

BIBLIOGRAPHY. C.E. FICHTEL and F.B. MCDONALD, "Energetic Particles from the Sun," *A. Rev. Astron. Astrophys.*, 5: 351–398 (1967), a review article summarizing the present knowledge of particle acceleration on the sun and their propagation in interplanetary space; V.L. GINZBURG and S.I. SYROVATSKII, *The Origin of Cosmic Rays* (1964), a quantitative theoretical text covering the problems of galactic (supernova) and extragalactic cosmic-ray origin, and of cosmic-ray interactions with interstellar matter, interstellar magnetic fields, and photons; S. HAYAKAWA, *Cosmic Ray Physics* (1969), a comprehensive text with main emphasis on cosmic ray interactions in the atmosphere; P. MEYER, "Cosmic Rays in the Galaxy," *A. Rev. Astron. Astrophys.*, 7:1–38 (1969), a review article summarizing the present knowledge of intensity, energy spectra, and composition of galactic cosmic rays; B.B. ROSSI, *High Energy Particles* (1952), a classic text on the interactions of high-energy electromagnetic radiation and particles with matter; and *Cosmic Rays* (1964), a nonmathematical treatment of the story of cosmic rays, written for the intelligent layman.

(P.M.)

Costa Rica

A republic of Central America, Costa Rica is located on the great isthmus that joins North and South America, between Panama, situated on its eastern border, and Nicaragua, on its northern border. Costa Rica's greatest length, between the two neighbouring countries, is 288 miles, and its narrowest width, between the Pacific Ocean and the Caribbean Sea, is 74 miles, the shortest coast-to-coast distance on the isthmus outside of Panama. Its total area is approximately 19,650 square miles (50,900 square kilometres).

Costa Rica played a role in the federation of Central American states from 1823 to 1838 and is a member of the Organization of Central American States founded in 1951. Of the five states that have been partners in these two enterprises, Costa Rica is the most Spanish in character and has the highest percentage of literacy, the largest gross domestic product per capita, and the longest experience with democracy. Its well-populated and highly civilized heartland, dedicated to quality harvests of coffee, has long been admired by both its own citizens and visitors from outside. Its outlying reaches, good for stock raising and banana cultivation, have only recently, with the building of new roads, become readily accessible to visitors.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Physical geography.* Three mountain chains run almost the entire length of Costa Rica, from southeast to northwest; they are divided into the Cordillera Central, the Cordillera de Guanacaste to the northwest, and the Cordillera de Talamanca to the

Mountain ranges and drainage patterns

southeast. The highest peak is Chirripó Grande in the Talamanca system, at 12,529 feet (3,819 metres). Two of the four volcanoes in the Cordillera Central, Irazú (11,260 feet) and Poás (8,875 feet) have paved roads to the rims of their active craters. The Guanacaste range also contains four major volcanoes, the highest of which is Miravalles (6,627 feet).

The southern side of the Cordillera Central overlooks the Valle Central, often called the Meseta Central, the heartland of the country that includes San José, the capital, at 3,809 feet. The Valle Central is drained by the Río Reventazón to the Caribbean and by the Río Grande de Tárcoles to the Pacific. To the north and east of the mountain ranges lie the Caribbean lowlands, about one-fifth of the country and less than 400 feet in altitude.

Roughly half of the land is too rocky for farming. Other portions are too completely leached or have too poor drainage or (on the Caribbean shore) consist of sand and bogs. The best agricultural areas are the Valle Central, river bottoms, and areas where the older alluvium of the northern lowlands is well drained.

Climate. Thermal-heating and shore-breeze patterns bring abundant precipitation to the Pacific coast in the wet season, generally May to October in the north, April to December in the south. Northeasterly winds on the Caribbean coast provide plentiful rain all the year round. The Central and Talamanca ranges have warm temperate climates with the southern side of the mountains both wet and dry, the northern side always wet.

San José weather records show monthly averages of rainfall from well under one inch in February to nearly 14 inches in September, with over 75 inches the average for a year. Temperatures vary with the altitude, San José reporting a mean of 69° F (21° C), a nearby station at 7,665 feet, a mean of 59° F (15° C), and another at 682 feet, a mean of 80° F (27° C).

Vegetation and animal life. Dense broadleaf evergreen forest, which includes mahogany and tropical cedar trees, covers about half the landscape. On the Talamanca range, there are many evergreen oaks and, above the timberline, mountain scrub and grasses. The northwest, with the most pronounced dry season, contains open deciduous forest. Palm trees are common on the Caribbean coastline, and mangroves near the Pacific where peninsulas are attached to the mainland, both the palms and mangroves growing in swampy ground.

Mammalian life is both abundant and varied. It has major ties to both South and North American animal populations. The former tie includes monkeys, anteaters, sloths; the latter, deer, wildcats, weasels, otters, coyotes and foxes, and many other species and genera.

Character of human settlement. Almost everywhere in Costa Rica the family unit is tight knit, with sharply defined male, female, and child roles. In the Valle Central, rural men are employed in the coffee, sugarcane, or dairying industries, while some of those in urban areas work in manufacturing. The very Spanish atmosphere of the Valle Central, even though patterns of life are changing, is the milieu of which most outsiders speak when they refer to Costa Rica. The Pacific northwest, where many men work on large cattle ranches, or haciendas, and also maintain small agricultural plots of their own, is more typically Central American. From this district came the national dance, the *punto guanacasteco*. The Pacific south contains two banana zones, developed in the 1930s, and significant new settlement in the Valle del General that followed the construction of the Inter-American Highway. The Caribbean side contains many reminders of an older banana industry, abandoned in the 1930s, with railroads instead of roads (children even go to school on the trains) and with houses built on stilts and, under each one, large drawers for drying cacao. The San Carlos Plains, part of the northern lowlands, are attracting new settlers, who live like pioneers. The typical Costa Rican house has wooden walls and floors and roofs made of corrugated metal. Earthen floors predominate over wooden ones only in the Pacific northwest.

The built-up portion of the San José metropolitan area had an estimated population of almost 409,000 in the early 1970s. Deep gullies line the city proper on the north and south; some suburbs are built in irregular patterns, following the terrain. The city has an unhurried appearance: its life is congested only in a small area in which the larger stores and offices are located and, recently, the first modern edifices have been erected.

Character of San José

Outside the capital urban life is similar. Cartago offers a colonial flavour, being much older than any of the other cities. It and Heredia and Alajuela, each near 20,000 in population and very near San José, all cater to the outlying provinces they head. Limón and Puntarenas are the chief Caribbean and Pacific ports; Puntarenas (situated on a sandspit) serves also as a warm resort centre. Golfito, on the Pacific, has a life based on banana export.

Costa Rica, Area and Population

	area*		population†	
	sq mi	sq km	1963 census	1972 estimate
Provinces (*provincias*)				
Alajuela	3,668	9,500	241,000	326,000
Cartago	1,004	2,600	155,000	209,000
Guanacaste	4,015	10,400	143,000	200,000
Heredia	1,120	2,900	85,000	111,000
Limón	3,591	9,300	68,000	94,000
Puntarenas	4,358‡	11,288	157,000	224,000
San José	1,896	4,910	488,000	647,000
Total Costa Rica	19,652	50,898	1,336,000	1,811,000

*Approximate; subject to revision. †De jure. ‡Includes islands totaling 100 sq km or 38 sq mi. Figures do not add to total given because of rounding.
Source: Official government figures.

THE PEOPLE OF COSTA RICA

Ethnic and religious groups. The Valle Central, with over half of Costa Rica's population, is predominantly European, or, more specifically, Spanish, in both its manner of living and its ancestry. In this district, minority Indian and Negro strains have been absorbed into the general European population. Spanish is spoken, though with distinctive national accents and usages. In Central America, a Costa Rican is called a *tico;* Costa Ricans replace the diminutive ending *tito* with *tico*, a practice known elsewhere but uncommon in Central America.

Catholicism, the religious faith of most of the people in the Valle Central, generally is adhered to very seriously. Roman Catholicism is also the official religion of the country, receiving a small part of the national budget. There is a Catholic archdiocese centred in San José, with suffragan bishops in Alajuela, Tilarán (for the Pacific northwest), and San Isidro (for the Pacific south), as well as an apostolic vicar in less Catholic Limón. Though priests are forbidden to bring politics into the pulpit, the church maintains an interest in social conditions.

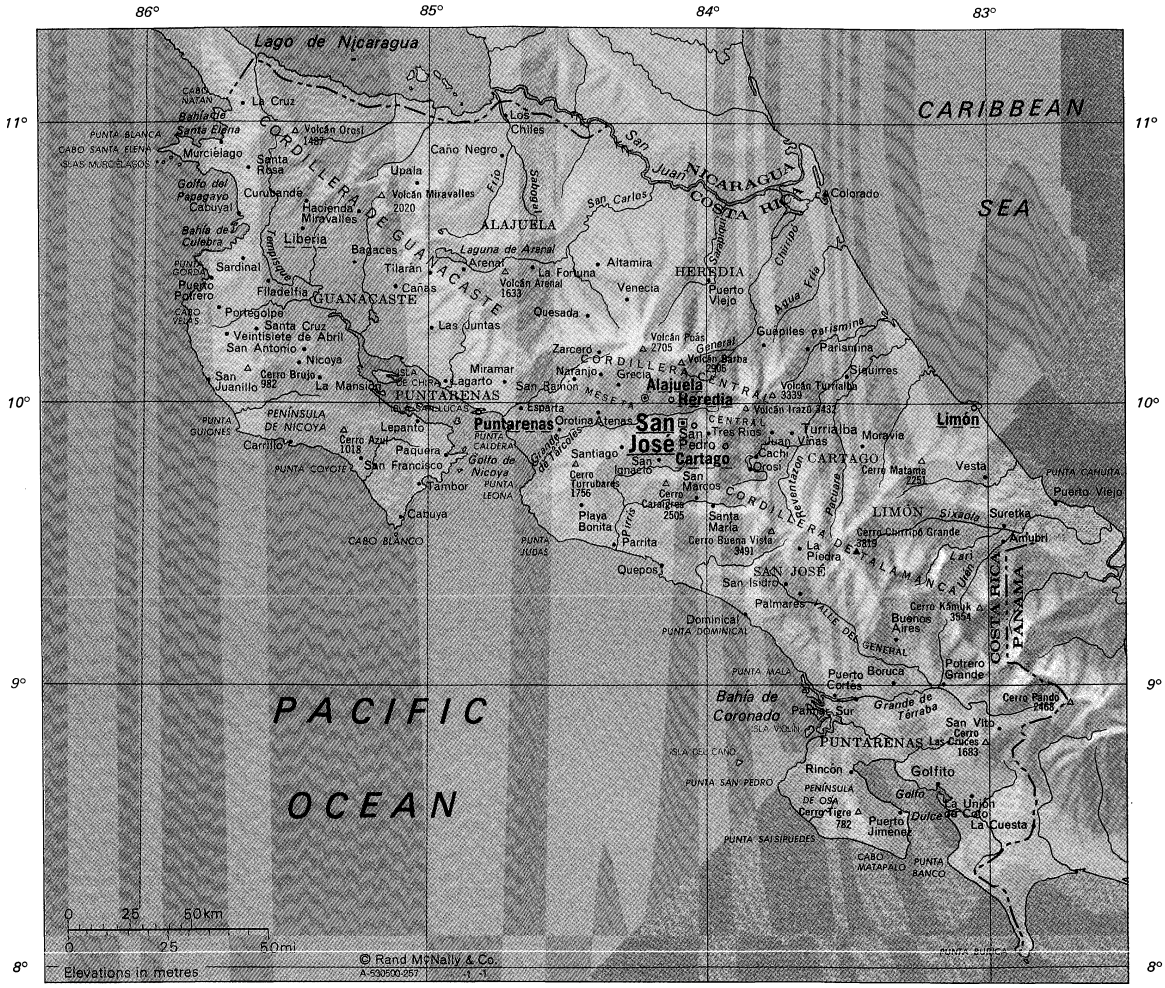
The racial and religious mix

Protestant evangelical activity is also strong in the Valle Central, although only a small minority of the Costa Rican people are Protestants. There is a language school, in which beginning missionaries study Spanish, and a seminary attended by Spanish American young people from many countries.

Nearly one-fifth of the population of the country is found in the Pacific northwest, often called Costa Rica's "brown" province. These people are a three-way blend of colonial Spanish, Indian, and Negro, whose Spanish is more like that of Nicaragua than that of the Valle Central. Their adherence to Catholicism is often nominal.

African ancestry is the strongest in the eastern Caribbean lowlands, which contain only about 6 percent of the population. The Negro people who live there are descendants of those brought from the West Indies to build railroads and raise bananas. Most of them speak both Spanish and a Jamaican style of English, a predominant number having come from that island. Protestantism is the most widespread faith in this district, some of the people of which are now moving to the Valle Central.

The Pacific south, with about 15 percent of the people, and the San Carlos Plains, with about 6 percent, are drawing upon the human resources of the other districts



COSTA RICA

MAP INDEX

Political subdivisions

Alajuela.....10·30n 84·30w
Cartago..... 9·50n 83·45w
Guanacaste.....10·30n 85·15w
Heredia.....10·30n 84·00w
Limón.....10·00n 83·15w
Puntarenas..... 9·00n 83·15w
San José..... 9·40n 84·00w

Cities and towns

Alajuela.....10·01n 84·13w
Altamira.....10·30n 84·23w
Amubri..... 9·31n 82·56w
Arenal.....10·29n 84·53w
Atenas..... 9·58n 84·23w
Bagaces.....10·31n 85·15w
Boruca..... 9·00n 83·20w
Buenos Aires.... 9·10n 83·20w
Cabuya..... 9·36n 85·06w
Cabuyal.....10·40n 85·40w
Cachí..... 9·50n 83·48w
Cañas.....10·25n 85·07w
Caño Negro.....10·54n 84·44w
Carrillo..... 9·52n 85·30w
Cartago..... 9·52n 83·55w
Colorado.....10·46n 83·35w
Covento..... 9·21n 83·30w
Curubandé.....10·43n 85·26w
Dominical..... 9·13n 83·51w
Esparta..... 9·59n 84·40w
Filadelfia.....10·26n 85·34w
Golfito..... 8·38n 83·11w
Grecia.....10·05n 84·18w
Guápiles.....10·13n 83·46w
Hacienda Miravalles.....10·41n 85·14w
Heredia.....10·00n 84·07w
Juan Viñas..... 9·54n 83·45w
La Cruz.....11·04n 85·39w
La Cuesta..... 8·30n 82·50w
La Fortuna.....10·30n 84·35w
Lagarto.....10·07n 84·56w
La Mansión.....10·06n 85·22w
La Piedra..... 9·29n 83·40w
Las Juntas.....10·16n 85·00w
La Unión de Coto..... 8·36n 83·03w
Lepanto..... 9·57n 85·02w
Liberia.....10·38n 85·27w
Limón.....10·00n 83·02w
Los Chiles.....11·02n 84·43w
Miramar.....10·06n 84·44w
Moravia..... 9·51n 83·26w
Murciélago.....10·55n 85·44w
Naranjo.....10·06n 84·22w
Nicoya.....10·09n 85·27w
Orosí..... 9·48n 83·51w
Orotina..... 9·54n 84·31w
Palmares.....10·03n 84·26w
Palmar Sur..... 8·58n 83·29w
Paquera..... 9·50n 84·56w
Parismina.....10·12n 83·38w
Parrita..... 9·30n 84·19w
Playa Bonita..... 9·39n 84·27w
Portegolpe.....10·20n 85·46w
Potrero Grande.. 9·00n 83·11w
Puerto Cortés.... 8·58n 83·32w
Puerto Jiménez.. 8·33n 83·19w
Puerto Potrero...10·28n 85·47w
Puerto Viejo.....10·26n 83·59w
Puerto Viejo..... 9·39n 82·45w
Puntarenas..... 9·58n 84·50w
Quepos..... 9·27n 84·09w
Quesada.....10·19n 84·26w
Rincón..... 8·42n 83·29w
San Antonio.....10·12n 85·26w
San Francisco... 9·49n 85·15w
San Ignacio..... 9·48n 84·09w
San Isidro..... 9·22n 83·42w
San José..... 9·56n 84·05w
San Juanillo.....10·02n 85·44w
San Marcos..... 9·40n 84·01w
San Pedro..... 9·56n 84·03w
San Ramón.....10·06n 84·28w
Santa Cruz.....10·16n 85·36w
Santa María..... 9·39n 83·57w
Santa Rosa.....10·51n 85·38w
Santiago..... 9·51n 84·18w
San Vito..... 8·50n 82·58w
Sardinal.....10·31n 85·39w
Siquirres.....10·06n 83·30w
Suretka..... 9·34n 82·56w
Tambor..... 9·43n 85·01w
Tilarán.....10·28n 84·59w
Tres Ríos..... 9·54n 83·58w
Turrialba..... 9·54n 83·41w
Upala.....10·47n 85·02w
Veintisiete de Abril.....10·15n 85·45w
Venecia.....10·22n 84·17w
Vesta..... 9·43n 83·03w
Zarcero.....10·11n 84·23w

Physical features and points of interest

Agua Fría, *river*..10·35n 83·30w
Arenal, Laguna de, *lake*.....10·32n 84·55w
Arenal, Volcán, *volcano*.....10·28n 84·44w
Azul, Cerro, *mountain*..... 9·55n 85·18w
Banco, Punta, *point*..... 8·22n 83·09w
Barba, Volcán, *volcano*.....10·09n 84·05w
Blanca, Punta, *point*.....10·58n 85·54w
Blanco, Cabo, *cape*..... 9·33n 85·06w
Brujo, Cerro, *mountain*.....10·08n 85·38w
Buena Vista, Cerro, *mountain*..... 9·33n 83·45w
Burica, Punta, *cape*..... 8·03n 82·51w
Cahuita, Punta, *point*..... 9·44n 82·48w
Caldera, Punta, *point*..... 9·54n 84·43w
Caño, Isla del, *island*..... 8·43n 83·53w
Caraigres, Cerro, *mountain*..... 9·43n 84·08w
Caribbean Sea...10·50n 83·00w
Central, Cordillera, *mountains*.....10·10n 84·05w
Central, Meseta (Valle Central), *upland*..... 9·55n 84·05w
Chira, Isla de, *island*.....10·06n 85·09w
Chirripó, *river*....10·41n 83·41w
Chirripó Grande, Cerro, *mountain*. 9·29n 83·29w
Coronado, Bahía de, *bay*..... 9·00n 83·50w
Coto Brus, *river*.. 8·56n 83·05w
Coyote, Punta, *point*..... 9·46n 85·30w
Culebra, Bahía de, *bay*.....10·37n 85·40w
Dominical, Punta, *point*..... 9·13n 83·50w
Dulce, Golfo, *gulf*..... 8·36n 83·15w
Frío, *river*.....11·07n 84·46w
General, *river*....10·12n 83·54w
General, Valle del, *valley*..... 9·11n 83·30w
Gorda, Punta, *point*.....10·31n 85·50w
Grande de Tárcoles, *river*.. 9·47n 84·38w
Grande de Térraba, *river*... 9·00n 83·40w
Guanacaste, Cordillera de, *mountains*.....10·45n 85·05w
Guiones, Punta, *point*..... 9·54n 85·41w
Irazú, Volcán, *volcano*..... 9·59n 83·51w
Judas, Punta, *point*..... 9·31n 84·32w
Kámuk, Cerro, *mountain*..... 9·16n 83·01w
Las Cruces, Cerro, *mountain*..... 8·48n 83·01w

MAP INDEX (continued)

Leona, Punta, *point*..... 9·42n 84·40w
Mala, Punta, *point*..... 9·02n 83·38w
Matama, Cerro, *mountain*..... 9·48n 83·14w
Matapalo, Cabo, *cape*..... 8·22n 83·18w
Miravalles, Volcán, *volcano*. 10·45n 85·10w
Mona, Punta, *point*..... 9·37n 82·36w
Murciélagos, Islas, *islands*.... 10·52n 85·57w
Nicoya, Golfo de, *gulf*..... 9·47n 84·48w
Nicoya, Península de, *peninsula*..... 10·00n 85·25w
Orosí, Volcán, *volcano*..... 10·59n 85·28w
Osa, Península de, *peninsula*... 8·35n 83·33w
Pacific Ocean... 8·45n 85·00w
Pacuare, *river*.... 10.14n 83.17w
Pando, Cerro, *mountain*..... 8·56n 82·43w
Papagayo, Golfo del, *gulf*..... 10·45n 85·45w
Parismina, *river*..... 10·18n 83·01w
Pirrís, *river*..... 9·28n 84·20w
Poás, Volcán, *volcano*..... 10·12n 84·13w
Reventazón, *river*..... 10·19n 83·20w
Sabogal, *river*.... 10·55n 84·45w
Salsipuedes, Punta, *point*.... 8·28n 83·36w
San Carlos, *river*..... 10·48n 84·11w
San Juan, *river*... 10·56n 83·42w
San Lucas, Isla, *island*..... 9·58n 84·54w
San Pedro, Punta, *point*.... 8·38n 83·45w
Santa Elena, Bahía de, *bay*... 10·59n 85·50w
Santa Elena, Cabo, *cape*..... 10·55n 85·57w
Sarapiquí, *river*.. 10·43n 83·55w
Sixaola, *river*..... 9·35n 82·32w
Talamanca, Cordillera de, *mountains*..... 9·30n 83·40w
Tempisque, *river*. 10·12n 85·14w
Tigre, Cerro, *mountain*..... 9·48n 83·14w
Turrialba, Volcán, *volcano*..... 10·02m 83·46w
Turrubares, Cerro, *mountain*. 9·48n 84·28w
Urén, *river*..... 9·33n 82·55w
Velas, Cabo, *cape*..... 10·21n 85·52w
Violín, Isla, *island*..... 8·50n 83·38w

in Costa Rica, and of Panama and Nicaragua as well. Language and religious preferences of these districts are thus a mélange, though basically Spanish and Catholic.

The Indian population

In the far south, a few thousand Indians are left. The Bribrí and Cabecar tribes inhabit valleys on both sides of the Cordillera de Talamanca, the Boruca group the Pacific side only. There are also a few hundred Guatuso Indian people on the northern plains. The languages of all four groups are related to the Chibcha speech of Colombia. Those on the Pacific side in the south and the Guatusos in the north are slowly losing their languages and becoming assimilated in every way. Those on the Caribbean side in the southern Talamanca district maintain their separate ways, including their animistic religions. The only pre-Columbian group oriented toward Mexico, the Orotiña of the Pacific northwest, contributed strongly to the district's population, but already they have lost their language, and only a few of their customs remain.

Demographic trends. The total population of Costa Rica was estimated at more than 1,811,000 in the early 1970s. Although Costa Rica is often cited as a nation whose population growth rate is high, it is not on the brink of becoming overpopulated. As general prosperity has increased, the population growth rate has gone down despite a drop in both the general and the infant mortality rates. In recent years the general mortality rate dropped from 8.6 to 6.6 deaths per 1,000 inhabitants; the infant mortality rate dropped in the same three years from 76 to about 60 deaths within the first year per 1,000 infants; and the annual growth rate dropped in four years (1965–69) from 3.5 percent to 2.7 percent. Life expectancy is 66.8 years, compared with an average of 51.1 years for the other four Central American nations.

Costa Rica's net loss of population by migration is of minor importance in explaining the drop in the growth rate. Of greater significance is a diminution of the birth rate, which, in 1965, stood at 42.3 births per 1,000 inhabitants but, in 1970, had dropped to 33.8. European immigration and life styles have molded Costa Rican history and influenced its character. The 1963 census counted 1,334 persons born in Spain and 2,597 in the remainder of Europe. Immigration from American countries, however, is much heavier. With 92 inhabitants per square mile at the beginning of 1972, Costa Rica had ample room to contain an expansion of the population.

THE NATIONAL ECONOMY

Costa Rica is neither rich, as its name ("rich coast") implies, nor as poor as many of its neighbours. Its gross domestic product per capita is the highest in Central America and comparatively is better distributed among all social classes—though this does not mean that the wealth is shared as much as it might be, or that even if it were the Costa Rican family would be living well by European or North American standards. Recent economic growth rates, however, suggest that Costa Rica may develop an economy characterized by a highly active and productive work force earning a decent livelihood, though continuing to depend heavily upon the nation's exports.

Natural resources. Costa Rica has some mineral resources, though because of small quantity or relative inaccessibility they may be disappointing commercially. The most important are bauxite deposits in the General and Coto Brus valleys. There is manganese also, in and near the Nicoya Peninsula; some gold in the Osa Peninsula; and magnetitic sand on scattered beaches, particularly on the southern Caribbean coastline. Timber reserves, already used domestically, have considerable commercial potential. Hydroelectric sources of power, now being developed, have capacity enough to support a great expansion of industry in the Valle Central. The greatest of the reservoirs is at Cachí, on the Río Reventazón at the eastern edge of the Valle Central. Elsewhere most electricity is generated by diesel plants, and many rural areas have no electricity at all.

Agriculture. Agriculture occupies almost half of Costa Rica's economically active people and contributes nearly one-fourth of the national income. Maize (corn) is grown less widely than in other Central American lands, but corn and beans remain the most significant crops for home consumption. Both of them, and rice as well, are raised chiefly on the Pacific side of the mountains. Coffee and bananas vie for first place among the nation's valuable exports. Coffee of high quality comes from the Valle Central and a few outlying areas. Bananas are grown commercially on both shores, though in larger quantity along the southern coast of the Pacific. Three lesser but important exports of food are beef, chiefly from the Pacific northwest; sugar, from cane grown mainly in the lower altitudes in the vicinity of the Valle Central; and cacao, from the eastern Caribbean lowlands.

Principal crops

Manufacturing and services. Manufacturing, contributing nearly one-fifth of the national income in recent years, involves little more than one-tenth of the economically active population. Most of it is concentrated in the Valle Central, with a few new plants in Puntarenas and Limón. For domestic consumption, the processing of food and beverages and the making of textiles, shoes, and furniture are important. The main items for export, chiefly to other Central American nations, are fertilizers, rubber tires, cotton fabrics, sheets of galvanized iron, plywood, medicines, electric batteries, and metal containers (in rough order of importance).

The number of Costa Ricans employed in providing services is greater than the number in manufacturing. A third of those contributing services are domestic help, and another third are teachers and government employees. Commerce, construction, transport, utilities, and mining follow manufacturing as branches of activity.

Foreign trade. To round out its food supply, Costa Rica imports wheat from the United States and both maize and beans from isthmian neighbours. The larger values in nonfood imports, by country in 1970, were paper from the U.S., partially refined petroleum from Venezuela, excavation machinery from the U.S., iron sheets from Japan, cottonseed oil from Nicaragua, and insecticides from the United States, in decreasing order.

Coffee is shipped extensively but in greatest quantity to the U.S., seven countries of Western and Central Europe, the Soviet Union, and Saudi Arabia. The bananas go chiefly to the U.S., Italy, West Germany, and Belgium. The U.S. buys most of the beef, sugar, and cacao.

Economic organization. There are three trade union associations. The Costa Rican Confederation of Democratic Workers is affiliated with the International Confederation of Free Trade Unions, the Costa Rican General Confederation of Workers with the Communist-led

World Federation of Trade Unions, and the Costa Rican Christian Confederation of Workers and Farmers with the World Confederation of Labour. There are also a number of independent unions and a few nationwide employer associations.

Private and public sectors of the economy

Private industry has retained its place in Costa Rica even though there is much sentiment for a larger public investment role. Individual autonomous agencies created by the government cover the entire insurance field, operate one railroad, provide low-cost public housing, develop sources of electricity, hasten economic growth on the Caribbean coast, and operate many of the country's banks. Small national budgets, however, have made it difficult for the government to go much further, so that stress has been placed on new private investment instead.

Costa Rica is one country that made effective use of assistance granted under the United States aid program, which formed part of the Alliance for Progress. Despite temporary balance of payment difficulties, it has benefitted also by its membership in the Central American Common Market.

Transportation. The hub of almost all Costa Rican transportation is the Valle Central. From San José, Cartago, Heredia, and Alajuela, all very close together, fan out narrow, often tortuous, paved routes with little interconnection, that reach the many valley and mountain communities in the immediate area. From this centre of human activity there extend two narrow-gauge railways—an electrified, government-operated line to Puntarenas and a British-operated, diesel-engine line to Limón. Three lesser lines are maintained by a fruit company.

The Inter-American Highway

The Inter-American Highway has since 1955 made Costa Rica far more accessible for trade. The provincial capitals of Puntarenas and Liberia are now reached with ease from the Valle Central. So are Nicaragua and the Central American countries beyond, with freight vans hauled by land even as far as Guatemala's ports on the Caribbean. There is also steady traffic to Panama, the obstacles being greater distance, a difficult crossing over the Talamanca range, and a lack of pavement on the road until very recently. Of the seven provincial capitals, only Limón remains without a paved highway from San José, and that connection entered the construction stage in 1971.

Limón has two piers, and Puntarenas and Golfito one each, to handle freight transported by railway. Limón's traffic is heavier than that of the other two combined, despite Golfito's status in the banana industry. El Coco, near San José, is the only international airport used by jet planes. Paved runways exist at Limón and Golfito and gravel strips at a few other airports offering local service by Líneas Aéreas Costarricenses, known as LACSA.

ADMINISTRATION AND SOCIAL CONDITIONS

Costa Rica is governed by its constitution of November 1949, the tenth in its history. A president, two vice presidents, and a unicameral legislative assembly are elected at one time for a term of four years, the assembly by proportional representation. Magistrates of the Supreme Court are chosen by the assembly for eight-year terms, being then automatically continued in office unless removed by a two-thirds vote.

Local government. The nation's seven provinces are ruled by governors appointed by the president. Guanacaste, the "brown" province, lies in the Pacific northwest. Puntarenas contains two areas, that in the vicinity of its capital and another in the Pacific south, connected by a narrow strip of territory along the sea. Limón contains all the Caribbean coastland. The other four, though they include much of the untamed territory, are governed at, and have the same names as, the four cities of the Valle Central. Each province is divided into cantons, and each canton into districts. Councilmen for the cantons are elected locally, but budgets for all are approved by the national government.

Government finance. Import duties, until the 1960s the biggest source of government revenue, have been reduced to about one-half their former level. Other government revenue comes from a graduated income tax that runs from 1 to 30 percent of net income, a similarly progressive real estate tax, rather heavy consumption taxes on a variety of nonfood items, a new general sales tax of 5 percent, and a number of other levies.

Political parties. The fairness of national elections in recent decades has been indicated by the fact that every four-year period has seen a change in the party winning the presidency. In effect, the political situation since 1949 has been a struggle between the Party of National Liberation (Partido de Liberación Nacional) and its enemies. This organization, founded by the moderate Socialist José Figueres and friends, has impelled the nation into cooperative activity of various kinds, in the belief that good fortune should be shared by all Costa Rica's inhabitants. Opposition groups, highly personalist in nature and oriented toward the liberal left as well as the conservative right, generally unite behind one presidential candidate. Small parties, devoted to Marxism or the Christian Democratic movement, have not taken firm hold in the minds of the people, their role as left-wing reform groups already having been appropriated by the Party of National Liberation.

The Party of National Liberation

Police and judiciary. Costa Rican people were once proud of the fact that their country had more schoolteachers than army personnel. They now cite the fact that there is no army at all but only a nonconscripted civil guard that has police duties. There are district police also. A person accused of a crime will have his case decided by a single judge or, on appeal, by a panel of judges, without the use of a jury. Capital punishment is banned, and sentences to the penitentiary must be for a stated number of years.

Education. For two decades, the government has been taking vigorous action to see that all children attend school, that health standards are raised, and that adequate housing is available in urban zones. Education occupied 27.8 percent of the nation's budget in 1969. The census of 1963 showed a literacy count of 85.7 percent of the population ten or more years old, reaching to 94.8 percent in urban zones. Less than half of the people ten or more years old had, however, completed fourth grade. School attendance figures since then indicate a marked prolongation of the years spent in learning, especially in the cities. The University of Costa Rica has a well-planned, functional campus in San Pedro, very near San José, with an enrollment of more than 11,000 in 1970. More than one-third of the students were women.

Health and welfare. The greatest health problem for the Costa Rican people is protein-calorie malnutrition. Vitamin A deficiencies and goitre are common, as are infectious and parasitic diseases.

Though the break between the wealthy and the manual worker is less sharp in Costa Rica than in other Central American lands, there remains a large number of employees who are paid a very small wage. In the entire country, at the beginning of 1969, a sampling study showed that, of the urban working force receiving its compensation only in money (over three-fourths of whom worked more than 40 hours a week), 76 percent received less than 175 colones (about $26) a week, while only 6 percent received 400 colones (about $60) or more per week. At the same time, in the San José metropolitan area, black beans cost 12.1 cents per pound in United States currency, first-class rice 13.6 cents per pound, and workingman's shoes $4.63 per pair.

CULTURAL LIFE AND INSTITUTIONS

Tourist impressions of Costa Rica may include the national dance, called the *punto guanacasteco*, the gaily painted oxcarts now made in miniature for the tourist market, or the baiting of the bulls, which is done by amateurs during the week's holiday from Christmas to the New Year. Altogether, however, most Costa Rican diversions are cosmopolitan rather than nationalistic in nature. The people attend film shows with great frequency, enjoying international cinema. They listen to an extraordinary variety of music, especially from the more than 50 radio transmitters in the country. Residents of the Valle Central attend the national theatre, where the

music played and the drama performed may come from any part of the world.

Costa Ricans take an interest in their pre-Columbian art, which includes large statues from the Pacific northwest, smaller examples of carved relief in stone from other districts, and some very fine work done in the form of small objects of gold. Samples of all of these may be seen in the national museum. Genuine colonial architecture is little found, the most famed example being a 17th-century mission in Orosí. Cartago's older buildings, destroyed by earthquake, have in some cases been restored; new ones have also been built like them. Painting and sculpting have re-entered the Costa Rican scene only in recent decades. In the early 1970s a new art centre was being constructed in San José near the national theatre.

Costa Ricans are not inactive in the field of literature. *El Repertorio Americano*, edited in San José from 1919 to 1958 by Joaquín García Monge, was heralded throughout the Western Hemisphere as a magazine of high intellectual and literary value. Roberto Brenes Mesén and Ricardo Fernández Guardia were widely known as independent thinkers in the fields of education and history, respectively. Fabián Dobles has attracted international attention as a writer of novels on social-protest themes.

La Nación, independent but conservative, is the most widely read of Costa Rica's newspapers. A combination of *La Nación* with either *La República* or *La Prensa Libre*, which lean more toward reform ideas, would provide the subscriber with a fairly full account of Costa Rican news. Television programming, which relies heavily upon productions from the United States and Mexico, is only slowly becoming part of the life of the majority of the people. Residents of the Valle Central have a choice of stations, while other areas receive only the relayed transmissions of the one national network.

PROSPECTS

The character of present-day Costa Rica was set in 1943 in a letter written by José Figueres, who three times since then has been the country's leader (as chairman of a ruling junta, 1948–49; and president 1953–58, 1970–74). In setting goals for the nation, Figueres listed (1) honesty in government; (2) liberty for the people; (3) professionalism in public administration; and (4) a distinct social orientation, the state gradually assuming the direction of all economic activity, so that there might be a greater production of wealth and more equity in its distribution.

All four of these goals have been pursued vigorously by the Party of National Liberation while in power (1962–66 and the three periods with Figueres), without their being negated by National Liberation's opponents while they held the presidency (1949–53, 1958–62, 1966–70). This was true even though some of the opponents denounced the prospect of complete governmental control of economic activity.

BIBLIOGRAPHY. JAMES L. BUSEY, *Notes on Costa Rican Democracy* (1962), provides a brief but in-depth study of the social and economic conditions that provide the bases for democracy in Costa Rica. DONALD E. LUNDBERG, *Costa Rica* (1968), is an informal guide for tourists and prospective immigrants. HOWARD I. BLUTSTEIN *et al.*, *Area Handbook for Costa Rica* (1970), is a comprehensive fact book for government personnel. FRANKLIN D. PARKER, *The Central American Republics* (1964), gives historical background and a Spanish-language bibliography.

(F.D.P.)

Counterpoint

A peculiarly Western phenomenon, counterpoint is the art of combining different melodic lines in a musical composition. It is the most characteristic element in Western music and a major distinguishing feature between the music of the West and that of the Orient and of primitive peoples.

Counterpoint and polyphony distinguished

The term counterpoint is frequently used interchangeably with the term polyphony. This is not properly correct since polyphony refers generally to music consisting of two or more distinct melodic lines, while counterpoint refers to the compositional technique involved in the handling of these melodic lines. Good counterpoint requires two qualities: (1) a meaningful or harmonious relationship between the lines (a "vertical" consideration—*i.e.*, dealing with harmony), and (2) some degree of independence or individuality within the lines themselves (a "horizontal" consideration, dealing with melody). Musical theorists have tended to emphasize the vertical aspects of counterpoint, defining the combinations of notes that are consonances (euphonious combinations, implying musical repose) and dissonances (clashing combinations, implying musical tension) and prescribing where consonances and dissonances should occur in the strong and weak beats of musical metre. In contrast, composers, especially the great ones, have shown more interest in the horizontal aspects: the movement of the individual melodic lines and long-range relationships of musical design and texture, the balance between vertical and horizontal forces, existing between these lines. The freedoms taken by composers have in turn influenced theorists to revise their laws.

The term counterpoint is occasionally used by ethnomusicologists to describe aspects of heterophony—duplication of a basic melodic line, with certain differences of detail or of decoration, by the various performers. This usage is not entirely appropriate, for such instances as the singing of a single melody at parallel intervals (*e.g.*, one performer beginning on C, the other on G) lack the truly distinct or separate voice parts found in true polyphony and in counterpoint.

Finally, contemporary theorists generally use the term counterpoint in a narrow sense for musical styles resembling those of Palestrina or Bach and emphasizing clear melodic relationships (*e.g.*, melodic imitation) between the voice parts.

In the present article, counterpoint will be considered more broadly, as an essential element in many styles within Western music. Composers in different periods have used counterpoint differently: in the Middle Ages they used it for the superimposing of different rhythmic groupings, in the Renaissance for melodic imitation, in the Baroque for contrasts between groups of instruments or voices, in the Classical period in conjunction with tonality, the organization of music in terms of key (a specific group of interrelated notes and chords), in the Romantic in the combining of leitmotifs, or short melodic fragments, and in contemporary music in the arrangement of isolated components of sound.

Historical survey. *The Middle Ages.* The earliest examples of actual written counterpoint appear in the late-9th-century treatise *Musica enchiriadis.* Here a Gregorian

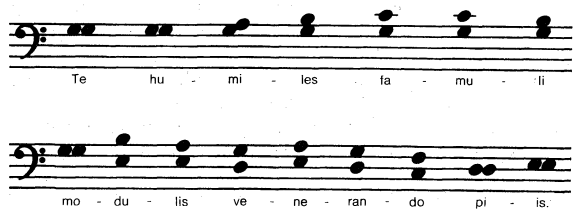


Figure 1: Early organum; from *Musica enchiriadis* (*c.* 859).

chant melody, or "principal voice" (*vox principalis*), is combined with another part, "organal voice" (*vox organalis*), singing the same melody in parallel motion a perfect fourth or fifth below (*e.g.*, G or F below C). Such music was called organum, probably because it resembled the sound of contemporary organs. In the early 11th century the teacher and theorist Guido of Arezzo in his *Micrologus* described a variety of organum in which the accompanying or organal voice had become more individualized. In addition to moving parallel to the main voice, it included oblique (diverging or converging) motion and contrary (opposite) motion. In this period the organal voice remains melodically awkward and subservient to the chant voice, as though it were composed one note at a time simply to colour or ornament each note of the chant. Early organum is thus not far removed from heterophony. Until the end of the 11th century organum was written entirely in note-against-note style, described, in

Organum: the earliest counterpoint

1336, as *punctus contra punctum* (point against point—*i.e.*, note against note), hence the name counterpoint (see Figure 1).

In the 12th century true polyphony comes into being; the melodic lines become individualized mostly by being given different rhythms. There emerges a hierarchy between the voice parts. The emphasis is upon the chant voice, which now becomes the lower part. Its notes are prolonged, or "held," and this part is now called the tenor, from the Latin *tenere*, to hold. The contrapuntal genius of the Middle Ages realizes itself mostly through the use of rhythmic contrasts between the different voice parts, and such contrasts gradually increase in complexity from *c.* 1100 to *c.* 1400. Around 1200 Pérotin, composer at Notre Dame in Paris who wrote some of the earliest music in three and four parts, superimposed different rhythmic modes (short fixed rhythmic patterns) in the voice parts. In his three-part *Alleluia Nativitas* (see Figure 2), the voices are in different rhythmic modes, and they are also distinguished by different phrase lengths, consisting of more or fewer repetitions of the rhythmic pattern. During the 13th century such contrasts were carried still further in the motet, a musical form usually in three voice parts, each in a different rhythmic mode. The theorist Franco of Cologne advocated the use of consonance at the beginning of each measure; such consonances (usually a chord made up of the unison, fifth, and octave, such as C–G–C′) served as fixed pillars in terms of which the horizontal extensions of different rhythmic lengths were like soaring arches of sound. The tenor voice part in the motets of the 14th and early 15th centuries was organized by huge rhythmic recurrences known as isorhythm (*i.e.*, the return throughout the piece of a complex rhythmic pattern, not necessarily in conjunction with the same pitches of the melody). During the 14th century, particularly in the works of Guillaume de Machaut, the upper voice part was sometimes displaced by a beat or more in respect to the other parts, giving it further rhythmic independence. In the late 14th century complicated syncopations (displaced accents) and the simultaneous use of different metres characterized some of the most complex counterpoint in history.

Medieval rhythmic complexity
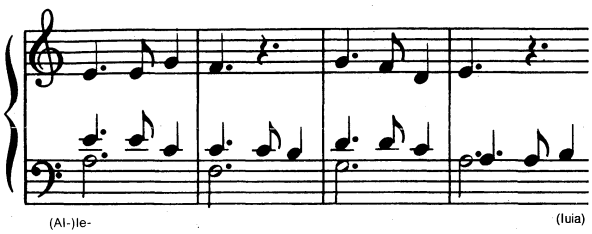


Figure 2: From *Christmas Alleluia* by Pérotin.

The Renaissance. If the medieval composer explored mostly the possibilities of rhythmic counterpoint, the Renaissance composer was concerned primarily with melodic relationships between the voice parts. The predominant technique used was that of imitation; *i.e.*, the successive statement of the same or similar melody in each of the voice parts, so that one voice imitates another.

Imitation had appeared earlier in the Italian caccia and French *chace*, roundlike vocal forms of the 14th century, and in England in the 13th-century round, *Sumer is icumen in*. These compositions anticipate the Renaissance and also emphasize the rhythmic relationships typical of medieval counterpoint.

During the Renaissance the technique of imitation contributed to a new unity between the voices, as opposed to the hierarchy found in medieval counterpoint. Renaissance composers strove also for clear melodic relationships between voices; consequently imitations usually began on the same beat of a measure and were separated in pitch by simple intervals such as the fifth (as, c–g) or octave (as, c–c′). The Renaissance theorists, among them Johannes Tinctoris and Gioseffo Zarlino, categorized dissonances according to type and governed each type by definite rhythmic and melodic restrictions.

What is often proclaimed as the "golden age" of counterpoint—meaning melodic counterpoint—stretches from the late 15th to the late 16th centuries, from the Flemish–French master Jean d'Okeghem to the Spanish Tomás Luis de Victoria and the Elizabethan William Byrd. Its leading masters were Josquin des Prez, Giovanni Pierluigi da Palestrina, and Orlando di Lasso. The northern composers in particular showed a penchant for complex melodic relationships. Okeghem's *Missa prolationum* (*Prolation Mass*), for example, involves simultaneous canons in two pairs of voices. (In a canon, one melody is derived from another. It may be identical, as in a round, or it may be given various alterations, as of speed, or metre or omission of certain notes.) The most versatile craftsman of the Renaissance was Josquin, whose music displays a continual variety of contrapuntal ingenuities, including melodic imitation. His use of successive imitation in several voices, as in his *Missa da pacem* (see Figure 3) based on the chant melody "Da pacem" ("give peace"), is coupled with melodic smoothness and rhythmic vitality. The imitative style came to its fullest flowering in the late 16th century, not only in the masses and motets of Lasso and Palestrina but also in secular songs such as the French chanson and Italian madrigal. It also flourished in instrumental music in such contrapuntal forms as fantasias, canzonas, and ricercares.

Golden age of melodic counterpoint
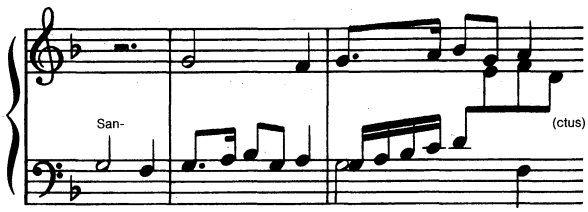


Figure 3: From *Missa da Pacem* by Josquin des Prez.

The Baroque period. During the 17th and early 18th centuries the pure linear—*i.e.*, melodic—counterpoint of the Renaissance, now called the first practice, was retained alongside the newer type of counterpoint known as the second practice. This latter type was characterized by a freer treatment of dissonances and a richer employment of tone colour. The new liberties with dissonance disturbed the conservative theorists of the time; but they were justified by their proponents on the ground that they allowed a more expressive treatment of the text. Still more distinct was a new use of tone colour. Although the individual melodic lines often resembled those of the Renaissance, they were intensified and made to stand out through differences of scoring or instrumentation. In figured bass compositions (in which a keyboard instrument improvised the harmonies over a given bass melody) the counterpoint was between the upper melody and the bass line. These stood out clearly from one another because of their differences of instrumental or vocal tone colour. Also significant at this time was the development of concerto-like scoring. In a concerto a soloist or group of instruments is contrasted with the entire orchestra. Hence concerto style emphasized contrasts between the numbers of performers, the high and low registers, and the tone colours of two or more performing groups. This was anticipated in some of the madrigals (Italian part-songs) of the late Renaissance, especially those of Luca Marenzio (1553–99) and Don Carlo Gesualdo, in which two or three voice parts in a high or low register were immediately answered by parts in a contrasting register. Giovanni Gabrieli of Venice expanded this principle in his *Symphoniae Sacrae* (*Sacred Symphonies*) by setting off choirs of voices or instruments, thus achieving a counterpoint of contrasting sonorities. Such concerto-like effects became an essential part of the later madrigals and operas of Claudio Monteverdi. In his madrigal *Lament of the Nymph* (see Figure 4), a single soprano voice is pitted against three male voices, and both in turn against an instrumental continuo (figured bass played, for example, by cello and harpsichord) in the background. This type of counterpoint was ideal for emphasizing dramatic con-

Influence of the concerto
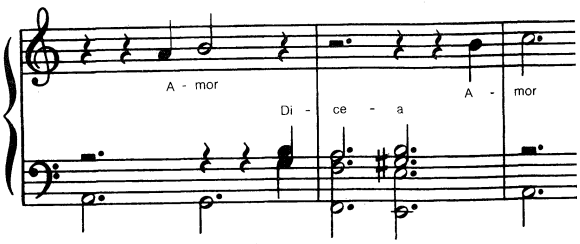


Figure 4: From *Lament of the Nymph* by Claudio Monteverdi.

trasts in the new forms of the opera and the oratorio. In these forms soloists, ensembles, and instrumental parts were opposed and combined in a great variety of ways by composers like Heinrich Schütz, Giacomo Carissimi, and Henry Purcell. In the late Baroque Arcangelo Corelli and Antonio Vivaldi added this style of dramatic contrasts to the purely instrumental contrasts of the concerto. The Baroque concerto culminated in the *Brandenburg Concertos* of J.S. Bach, which are characterized by a remarkable fusion of contrapuntal lines and instrumental colours (see Figure 5).

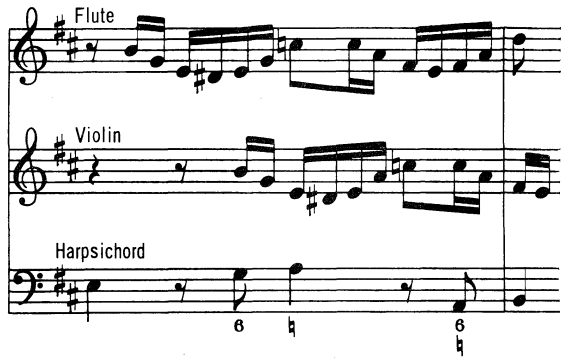


Figure 5: From *Brandenburg Concerto No. 5* by J.S. Bach.

Bach's counterpoint has a retrospective side, which uses a mainly melodic approach. The fugue, a composition using the technique of melodic imitation, became highly developed in Bach's hands—*e.g.*, the fugues of the *Well-Tempered Clavier*, and his final compendium of contrapuntal devices, *The Art of the Fugue*. A similar melodic, rather than tone-colour, approach occurs in works such as the *Inventions* and in the canons of the *Musical Offering*. These works are akin to "the first practice," the melodic counterpoint of the Renaissance, although in their use of dissonance and harmony they go considerably beyond Renaissance convention.

The Classic period. The turn from the Baroque to the Classic period in music was marked by the change from a luxuriant polyphonic to a relatively simple homophonic texture—*i.e.*, a texture of a single melodic line plus chordal accompaniment. Composers of the early Classic period (*c.* 1730–70) largely eschewed counterpoint altogether, drawing on it only when preparing church music in the "learned style," as the Renaissance style was then called. Many of the keyboard sonatas of Domenico Scarlatti and Carl Philipp Emanuel Bach, despite a basically homophonic approach, reveal a skillful interplay between the main melody and accompaniment. In the late Classic period (*c.* 1770–1820), especially in the music of the Viennese school of Haydn, Mozart, and Beethoven, there was an ever-increasing penetration of counterpoint into musical forms based on this homophonic style and its contrasts of tonality, or key (the particular notes and chords related to a given keynote). This counterpoint in turn was tempered by the Classical style and musical forms. For example, although combined melodic lines are heard as counterpoint, together they can also be heard as a series of harmonies. In this way they form unified phrases in the homophonic style. This satisfied demands for symmetrical phrase lengths and clear-cut cadences, or stopping points, necessary to mark the sections of Classical forms such as the sonata.

Use of counterpoint by Haydn, Mozart, and Beethoven

Haydn underwent his contrapuntal "crisis," or movement toward counterpoint, during the 1770s, the period of "Storm and Stress" in German literature, which had a deepening effect on other arts as well. Three of his *Sun Quartets* (1772) had fugues as final movements, and in the *Russian Quartets* (1781) Haydn proclaimed "an entirely new manner," in which the thematic material was to be more equally shared by all of the string instruments instead of being given to a single principal melody instrument.

Haydn heard Handel's oratorios in London, which inspired him to write his own richly contrapuntal late oratorios, *The Creation* and *The Seasons*.

Mozart's discovery of the contrapuntal art of Bach and Handel impressed him so deeply that almost all of his later works were affected. The ensembles of the operas—*e.g.*, *Don Giovanni* and *Così fan tutte*—with their clear delineation of several characters through their vocal lines, only became possible because of his new feeling for counterpoint. And at one point in his *Jupiter Symphony* five different themes are stated simultaneously, singly, or in combination. Nevertheless the counterpoint is kept entirely subservient to the harmonies of the symphony's tonal design, or its use of keys. Each voice is also governed by an underlying phrase structure applied to all of them, so that the combined parts form unified musical phrases (see Figure 6).
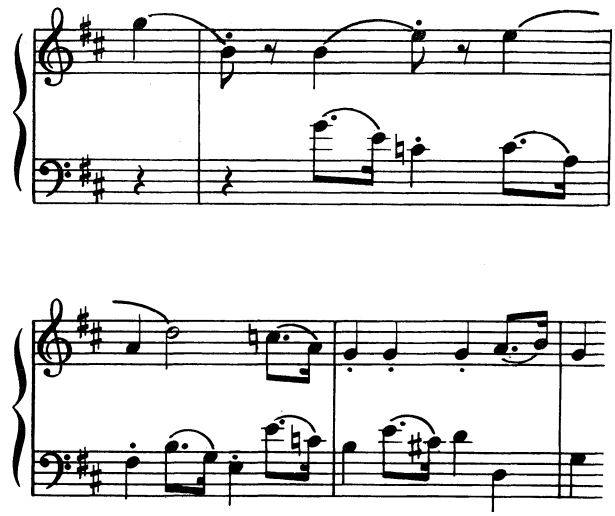


Figure 6: From *String Quartet* (K. 490) by W.A. Mozart.

Beethoven began his career in Vienna under the tutelage of the noted contrapuntal theorist Johann Albrechtsberger, and this, coupled with his admiration for Handel, probably accounts for his lifetime interest in counterpoint. He drew upon counterpoint to create musical intensity, especially in the development section of sonata (*q.v.*) form (the form prominent in Classical symphonies and chamber music), as in the first movement of the *Razumovsky Quartet*, Opus 59, No. 1, for example. In his late sonatas and quartets, except for obvious fugal works such as the first movement of Opus 131, or the *Great Fugue*, Opus 133, almost every movement shows the interpenetration of the principles of counterpoint, which deals with melodic lines, and tonality, which deals with harmonies.

The Romantic period. Counterpoint in the 19th century had a retrospective side in addition to a characteristically Romantic style. Richard Wagner admired the counterpoint of Palestrina, and Johannes Brahms revered the Baroque masters. Felix Mendelssohn revived Bach's *St. Matthew Passion* in 1829, and this led to numerous Bach-like works, such as the organ sonatas of Mendelssohn and numerous organ works by Max Reger, as well as arrangements of Bach's works by Franz Liszt. Yet the true bent of Romantic composers was toward combinations of motives (small melodic fragments), use of mo-

Combinations of motives and leitmotifs

tivic accompaniments against themes, and later, of the combination of leitmotifs, or motives with significance beyond the music itself. The lieder (songs) of Franz Schubert were highly innovative because of their motivic accompaniments, which balance in interest the vocal part itself and contrapuntally interact with it. This technique is still more pronounced in the songs of Robert Schumann and Hugo Wolf. It is also the tendency in 19th-century opera. In the later operas of Giuseppe Verdi the voices often have a parlante character (imitating speech through music) while the orchestra defines the dramatic substance. This, too, is the principle of the Wagner music dramas, with their "speech-song" (*Sprechgesang*) in the voice balanced contrapuntally by the leitmotifs of the accompaniment. In *Tristan und Isolde* Wagner set the leitmotifs in counterpoint against one another. Similarly, in the Prelude to Act III of *Siegfried*, a motif known as the

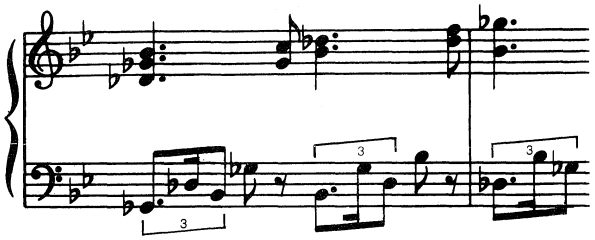


Figure 7: From *Siegfried* (Act III) by Richard Wagner.

"Need of the Gods" is cast against one associated with the "Valkyries." This results in a "counterpoint" of connotations and of emotions as well as in a musical counterpoint (see Figure 7). In purely instrumental music a similar joining of motives previously heard separately is encountered in the finale of Hector Berlioz's *Symphonie fantastique* when the plainchant melody "Dies Irae" ("Day of Wrath") is heard together with the theme called "Round of Sabbath." Richard Strauss, in his tone poem *Ein Heldenleben* (*A Hero's Life*), skillfully combines several themes taken from his earlier tone poems. And in the late symphonies of Gustav Mahler there is sometimes a complex of interwoven motives, each of which stands out contrapuntally through its presentation by a solo instrument.

In the 20th century Arnold Schoenberg carried this technique further, especially in his 12-note works, which are based on a 12-tone row, or specific ordering of the 12 notes of the chromatic scale, arranged in such a way as to avoid a sense of tonality. In some 12-note operas—*e.g., Moses und Aron* by Schoenberg and *Lulu* by Alban Berg—there is but one tone row used in the entire work; nonetheless, several hours of music are spun out of it through a continual variety of thematic shapes and contrapuntal combinations.

The contemporary period. The 20th century, like the 19th, has had its counterpoint inspired by earlier music. Anton Webern, for example, advocated a return to the forms of counterpoint used by Renaissance composers such as Heinrich Isaac, and in numerous of his own works (*e.g., Symphonie*) he makes use of Renaissance contrapuntal devices such as simultaneous canons and retrograde movement between the voice parts—*i.e.*, one voice using the other's melody but with the notes in reverse order. Out of a similar return to Baroque forms came musical works such as the double fugue (a fugue based on two themes) that forms the second movement of the *Symphony of Psalms* by Igor Stravinsky.

Counterpoint of polytonality and tone colours

But the use of older musical forms is no more of the essence of 20th-century counterpoint than it was of the 19th. A basic characteristic of 20th-century counterpoint is the separation of the voice parts into isolated entities of sound that are of themselves rather static. This may take the form of polytonality (the simultaneous use of two or more keys), using as static entities the notes of each key. It may also take the form of contrast of individual tone colour effects rather than of melodies, found in much electronic music. (This use extends beyond the original definition of counterpoint simply as the combination of melodies.)

Richard Strauss's *Elektra* (1909) was one of the earliest works to make use of polytonality; in certain passages the instruments and voice parts are grouped into layers, each of which defines a different tonality, or key, although in this case all of the keys can also be interpreted as complicated aspects of the basic key. Stravinsky's *Three Pieces for String Quartet* suggests four keys at the same time: G, B, D, and A♭. In this particular work each instrument is limited throughout the piece to a few notes assigned to it. Thus each part is absolutely individual and, except for the viola, consists of an ostinato melodic and rhythmic pattern. The coming together of these ostinato patterns at different times and in continually shifting arrangements suggests the effect of a mobile (see Figure 8).

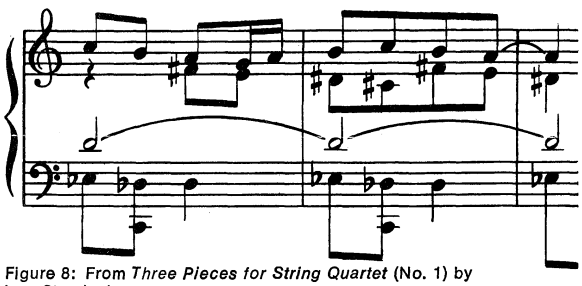


Figure 8: From *Three Pieces for String Quartet* (No. 1) by Igor Stravinsky.

Béla Bartók carried out a similar procedure in many of the short piano pieces of his *Mikrokosmos*, and in his *Fourth Quartet* (1928) he set apart tone clusters (chords built up in seconds, as C–D–E–F–G) in this way.

Turning now to a counterpoint purely of tone colours, *Intégrales* (1925) by Edgard Varèse presents 11-note "sound-clouds" in the wind instruments in opposition to the sounds of a large battery of percussion instruments. This approach probably grew directly out of earlier experiments with polytonality, but here tone colours, rather than keys or tones, are differentiated. Elliott Carter in his *Double Concerto* (1961) set apart two groups of instruments, one around a piano, another around a harpsichord, each with its distinctive tone colours and its own distinctive harmonic intervals or note combinations. In György Ligeti's *Atmospheres* every instrument in a symphony orchestra, including every string part, plays its own unique, melodic pattern; all of these parts coalesce into gigantic bands or spectra of tone colour that contrast with one another. In later experiments, the sound-producing groups are further set off by visual or spatial contrasts in the physical placement of performers; *e.g.*, Ramon Zupko's *Third Planet from the Sun*, 1970.

Evaluation and discussion of writings on counterpoint. Most of the writings on counterpoint have sought to increase the student's skill in musical composition. From the 18th century onward, textbooks of counterpoint have recommended as a model usually Palestrina or Bach, and in some recent cases 20th-century composers. Medieval and Renaissance treatises also were originally intended for student guidance and reflect the taste and attitudes of their own time. Several 20th-century studies deal with the contrapuntal technique of a particular composer or group of composers. At present there is no comprehensive study analyzing both the actual practices of composers and the theoretical writings on counterpoint throughout history.

BIBLIOGRAPHY

General studies: HUGO RIEMANN, *History of Music Theory*, trans. by RAYMOND HAGGH (1962), rather outdated, but still the most thorough summary of medieval and Renaissance theoretical studies on counterpoint; KNUD JEPPESEN, "Outline History of Contrapuntal Theory," in *Counterpoint*, pp. 3–53 (1939), the main theoretical views on the subject; GUSTAVE REESE, *Fourscore Classics of Music Literature* (1957), a synopsis of 80 theoretical sources, many of which deal with counterpoint; OLIVER STRUNK, *Source Readings in Music His-*

tory (1950), numerous excerpts from musical theorists on the subject of counterpoint.

Historical treatises: JOHANNES TINCTORIS, *Liber de arte contrapuncti* (1477; trans. by ALBERT SEAY, *The Art of Counterpoint*, 1961), a famous landmark, the first extensive outline of contrapuntal principles; GIOSEFFO ZARLINO, *Le istitutioni harmoniche* (1558; pt. 3 trans. by GUY MARCO and CLAUDE PALISCA as *The Art of Counterpoint*, 1968); THOMAS MORLEY, "Treating of Descant," *A Plain and Easy Introduction to Practical Music*, pt. 2 (1597; new ed. by R. ALEC HARMAN, 1952), a pupil–master discussion that offers firsthand information concerning the 16th-century approach to counterpoint; LODOVICO ZACCONI, *Prattica di musica*, pt. 2 (1622), one of the first presentations of the five species as a means of teaching counterpoint; JOHANN FUX, *Gradus ad Parnassum* (1725; trans. by ALFRED MANN and JOHN EDMUNDS, *Steps to Parnassus*, 1943), probably the most celebrated of all books on this subject; mainly concerned with the problems encountered in writing counterpoint.

Textbooks: Information on the 16th-century (Palestrina) style may be found in EBENEZER PROUT, *Counterpoint Strict and Free* (1890), still among the most thorough presentations; C.H. KITSON, *The Art of Counterpoint* (1907), species approach; R.O. MORRIS, *Contrapuntal Technique in the Sixteenth Century* (1922); A.T. MERRITT, *Sixteenth-Century Polyphony* (1939); G.F. SODERLUND, *Direct Approach to Counterpoint in 16th Century Style* (1947); S.I. TANEEV, *Convertible Counterpoint in the Strict Style*, trans. by G.A. BROWER (1962), exercises in invertible counterpoint; and OWEN SWINDALE, *Polyphonic Composition* (1962). For 18th-century (Bach) style, see A.I. MCHOSE, *The Contrapuntal Harmonic Technique of the 18th Century* (1947); WALTER PISTON, *Counterpoint* (1947), extends also to 19th-century music; and K.W. KENNAN, *Counterpoint Based on Eighteenth Century Practice* (1959). Works on the 20th-century style include: ERNST KRENEK, *Studies in Counterpoint Based on the 12-Tone Technique* (1940), with emphasis on the melodic aspects of twelve-note composition; HUMPHREY SEARLE, *Twentieth Century Counterpoint* (1954), contain chapters outlining the contrapuntal theories of Hindemith and the 12-note technique of Schoenberg; and R.E. MIDDLETON, *Harmony in Modern Counterpoint* (1967).

Specialized studies: GILBERT REANEY, "Fourteenth Century Harmony," in *Musica Disciplina*, 7:129–146 (1953), an examination of the medieval approach to counterpoint and its application in the music of Guillaume de Machaut; KNUD JEPPESEN, *The Style of Palestrina and the Dissonance*, 2nd rev. ed. (1946); H.K. ANDREWS, *An Introduction to the Technique of Palestrina* (1958); ALFRED EINSTEIN, "Mozart and Counterpoint," in *Mozart*, trans. by ARTHUR MENDEL and NATHAN BRODER (1965); DOROTHY SLEPIAN, "Polyphonic Forms and Devices in Modern American Music," *The Musical Quarterly*, 33:311–326 (1947); C.W. FOX, "Modern Counterpoint: A Phenomenological Approach," in *Music Library Association Notes*, 6:46–57 (1948).

(R.J.J.)

Couperin Family

For over 200 years, beginning in the early 17th century members of the Couperin family, one of the outstanding dynasties of music, worked in and around Paris as composers, performers, and teachers. Their name is linked inseparably with the church of Saint-Gervais, where for 173 years the post of organist was held by a Couperin.

The family is first heard of in the district of Brie, about 30 miles southeast of Paris. There, in the little walled town of Chaumes-en-Brie, between 1626 and 1638, three brothers were born. Their father, a merchant and small landowner, was also the organist of the local abbey church, and all three learned to play respectably on the violin, viol, harpsichord, and organ. Still, they might have remained provincial musicians but for a gesture of friendly respect. One summer, around 1650, Jacques Champion de Chambonnières, the best harpsichordist in France, was celebrating his name day at his house near Chaumes-en-Brie. The Couperin brothers and some friends honoured the great man with a serenade. The delighted Chambonnières, learning that it was Louis (*c.* 1626–61), the eldest, who had composed the music, insisted that he go to Paris.

In 1653 Louis had become the first Couperin to occupy the post of organist at Saint-Gervais, on the right bank of the Seine, across from Notre-Dame Cathedral. He also held a court appointment as a treble viol player, but it was for his performing ability as a harpsichordist that he was best known. Until about 1960 when a collection of 70 organ pieces was discovered, his known compositions had consisted of 123 pieces for harpsichord and a handful for viol and organ. This small surviving sample of his life's work suggests that when he died in 1661, at only 35, the 17th century lost one of its greatest musical talents. Into conventional forms he crowded invention, learning, and passion in a measure surpassing even Chambonnières.

Giraudon



François Couperin le Grand, portrait by an unknown French artist, 18th century. In the Musée de Versailles.

The two younger brothers followed Louis to Paris and, although less illustrious, became successful musicians. François was described as a "great musician and great drunk"; no compositions are known, but his line of the family carried the name of Couperin into the 19th century. Charles succeeded Louis at Saint-Gervais, styled himself Sieur de Crouilly after some land inherited from his father, and, in 1668, produced an only child, François Couperin le Grand (1668–1733), who stands far above all the other Couperins with the exception of Louis, his equal in genius.

François Couperin

Although François was only 11 when his father died, the wardens of Saint-Gervais reserved the office of organist for him until he was 18. The organist Jacques Thomelin acted as a "second father" to François, and his mother went to considerable expense to provide teachers of "music, harpsichord, and organ" so that François would be ready at 18 to assume his father's position. The boy took over before his 18th birthday and in 1693 became one of the four organists of the royal chapel. One honour followed another: harpsichord teacher to the royal children (1694), *chevalier de l'ordre du Lateran* (1702), and the *survivance* (right to succeed) of Jean-Henri d'Anglebert as court harpsichordist (1717).

By 1723 Couperin's health obliged him to bestow the *survivance* at Saint-Gervais upon his cousin Nicolas, and in 1730 the d'Anglebert *survivance* went to his daughter Marguerite-Antoinette. François had married in 1689, and of his four children none gave him a grandchild. Only Marguerite-Antoinette remained to carry on his work, until ill health forced her, too, to desist. Despite his poor health, Couperin composed, collected, and edited at a furious rate, bringing out major publications in 1724–26, 1728, and 1730—half of all his surviving works. Like Louis, François is known above all for his harpsichord music. It carried his name across Europe and passed for the quintessence of civilized elegance. J.S. Bach knew it and copied it. Most of his 254 harpsichord pieces have titles and were written with some person or thing in mind, now often impossible to identify; all are alive with colours of the age captured in the cool, luminous precision of harpsichord sound.

Nicolas Couperin

Nicolas, who had been acting organist at Saint-Gervais, took over the full title at the death of François in 1733. No music can be definitely attributed to Nicolas, and he seems not to have been much talked about as a performer. He died in 1748, and the post went to his only son, Armand-Louis, who kept it until his death in a traffic accident in 1789. The surviving music of Armand-Louis, the last of the Couperins to enjoy an international reputation, reflects the cosmopolitan ferment of mid-18th century Paris and shows a strong experimental vein. His son Pierre-Louis took over at Saint-Gervais but died the same year and was succeeded by his brother, Gervais-François, last male of the line.

In Revolutionary Paris in 1793, all the churches were closed; but Gervais-François bent with the political winds of the next 30 years, celebrating in music the Republic, the Empire, and the restored Monarchy, all with impartial loyalty. He ended his days in 1826 in the same organ loft at Saint-Gervais that his great-great-uncle had entered 173 years before. The surviving music by both brothers is mostly for piano.

In 1830 Gervais-François's daughter Céleste and her mother retired to Beauvais, where Céleste gave lessons until 1843, when they returned to Belleville, near Paris. In 1848 the mother sold the family portraits to the state for a pittance. Céleste died in 1860 survived by her mother, with whom the dynasty perished in 1862.

MAJOR WORKS

François Couperin ("le Grand")

THEORETICAL WORK: *L'Art de toucher le clavecin* (1716 and 1717).

CHURCH MUSIC: 3 *Leçons de ténèbres* (*c.* 1714–15); *Élévations;* various motets.

CHAMBER MUSIC: 4 *Concerts royaux* (published 1722); *Les Goûts réunis* (1724); *Le Parnasse ou l'Apothéose de Corelli* (1724); *Apothéose de Lully* (1725); *Les Nations* (1726).

HARPSICHORD: *Livres de clavecin:* these works were published in four books in 1713, 1717, 1722, and 1730, arranged in 27 *Ordres* each comprising pieces in the same key, major or minor.

ORGAN: *Pièces d'orgues consistantes en deux messes: Messe pour les paroisses et Messe pour les couvents* (1690).

Louis Couperin

INSTRUMENTAL MUSIC: 123 pieces for harpsichord, including unmeasured preludes and dances: allemandes, courantes, sarabands, etc.; 70-plus organ pieces including fantasies, fugues, and versets.

Armand-Louis Couperin

A book of harpsichord pieces; 9 sonatas for harpsichord with the accompaniment of violin and violin and cello; 2 sonatas for two harpsichords with expressive devices.

BIBLIOGRAPHY. Owing to the enormous richness of French archives and the continuing researches of musical archivists, each year brings new discoveries about the Couperins, so that opinions about even such basic matters as how many Couperins there were are being continually revised. This is why there is so little agreement among the most recent articles in reference works. The following studies, however, are particularly recommended: MAURICE CAUCHIE, *Thematic Index of the Works of François Couperin* (1949); "Bibliographie chronologique des Couperin," *Mélanges François Couperin*, pp. 132–136 (1968); WILFRED MELLERS, *François Couperin and the French Classical Tradition* (1950), the most comprehensive study to date; CHARLES BOUVET, *Les Couperin* (1919), the first and most detailed work (copiously documented) on the entire family; *Nouveaux documents sur les Couperin* (1933); JULIEN TIERSOT, *Les Couperin* (1926); GUY OLDHAM, "A New Source of French Keyboard Music of the Mid-17th Century," *"Recherches" sur la musique française classique*, vol. 1, pp. 51–59 (1960).

(D.Fu.)

Courbet, Gustave

Gustave Courbet founded and dominated the school of Realism in mid-19th century France in opposition to the academic art that had dominated French painting since 1800. His paintings, demonstrations of his belief that art should render reality unadorned and unaltered, shocked a society that had long been accustomed to an art that prettified life.

Early life and work. Courbet was born at Ornans, in eastern France, on June 10, 1819, the son of Eléonor-Régis, a prosperous farmer, and Sylvie Courbet. After attending both the Collège Royal and the college of fine arts at Besançon, he went to Paris in 1841, ostensibly to attend law school. He devoted himself more seriously, however, to studying the paintings of the masters in the Louvre. Father and son had great respect for each other, and, when Courbet told his father of his intention to become a painter rather than a provincial lawyer, his father consented, saying, "If anyone gives up, it will be you, not me," and adding that, if necessary, he would sell his land and vineyards and even his houses.

Freed from all financial worry, young Courbet was able to devote himself entirely to his art. He gained technical proficiency by copying the pictures of Diego Velázquez, Ribera, and other 17th-century Spanish painters. In 1844, when he was 25, after several unsuccessful attempts, his self-portrait "Courbet with a Black Dog," painted in 1842, was accepted by the Salon—the only annual public exhibition of art in France, sponsored by the Royal Academy. When in the following years the jury for the Salon thrice rejected his work because of its unconventional style and bold subject matter, he remained undaunted and continued to submit it.

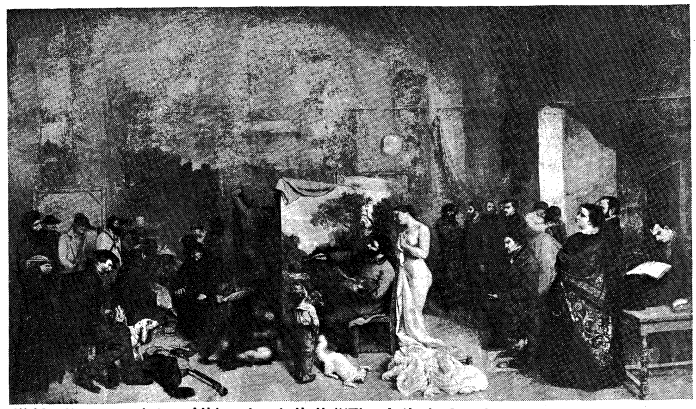
The development of Realism. The revolution of 1848 ushered in the Second Republic and a new liberal spirit that greatly affected the arts. The Salon held its exhibition not in the Louvre itself but in the adjoining galleries of the Tuileries. Courbet exhibited there in 1849, and his early work was greeted with considerable critical and public acclaim.

In 1849 he visited his family at Ornans to recover from the hectic life in Paris and, inspired again by his native countryside, produced two of his greatest paintings: "The Stone-Breakers" and "Burial at Ornans." Painted in 1849, "The Stone-Breakers" is a realistic rendering of two figures doing menial labour in a barren, rural setting. The "Burial at Ornans," from the following year, is a huge representation of a peasant funeral, containing more than 40 life-size figures. Both works depart radically from the more controlled, idealized pictures of either the Neoclassic or Romantic schools; they portray the life and emotions not of aristocratic personages but of humble peasants, and they do so with a realistic urgency. The fact that Courbet did not glorify or ennoble his peasants but presented them boldly and starkly created a violent reaction in the art world.

Early realistic works

Leader of the new school of Realism. Courbet, an intimate of many writers and philosophers of his day, including the poet Charles Baudelaire and the social philosopher Pierre-Joseph Proudhon, became the leader of the new school of Realism, which in time prevailed over other contemporary movements. One of the decisive elements in his development of Realism was his lifelong attachment to the traditions and customs of his native province, the Franche-Comté, and of his birthplace, Ornans, one of the most beautiful towns in the province. After a brief visit to Switzerland, he returned to Ornans and in late 1854 began an immense canvas, which he completed in six weeks, "The Artist's Studio," an allegory of all the

Giraudon



"L'Atelier du peintre, Allégorie réelle" ("The Artist's Studio"), with Courbet at the easel; oil on canvas, 1855. In the Louvre.

influences on Courbet's artistic life, portrayed as human figures from all levels of society. Courbet himself presides over all the figures with ingenuous conceit, working on a landscape and turning his back to a nude model. When the painting was refused by the jury for the 1855 Exposition Universelle, Courbet, with the financial support of a friend, opened his own pavilion of Realism to exhibit his works close to the exposition. His show provided the first group exhibition of Realistic paintings. The enterprise failed; the painter Eugène Delacroix alone, in his journal, praised the audacity and talent of Courbet, who made another attempt in 1867 but was again unsuccessful.

In 1856 Courbet visited Germany, where he was warmly welcomed by his fellow artists. Three years later, at the age of 40 and still working in defiance of severe criticism in his own country, he was the undisputed master and leader of a new generation of painters that had turned away from the traditional schools of painting, which they considered only barriers to artistic inspiration. Courbet worked in all genres. A lover of women, he glorified the female nude in paintings of stunning warmth and sensuality. He executed admirable portraits, but above all he celebrated the Franche-Comté, the forests, springs, rocks, and cliffs of which were immortalized by his vision. In 1865 he set up his easel before the cliffs of Étretat, Deauville, Trouville, and other resorts fashionable during the Second Empire. Carefully observing air currents and storm skies, he successfully depicted the architecture of a tempest in a series of seascapes. These pictures were an extraordinary achievement that amazed the world of art and opened the way for Impressionism, which was to achieve an even greater sensualism by reproducing the colour and light reflected by an object rather than its strict linear shape.

Portraits, landscapes, and seascapes

Political activist. The Franco-Prussian War broke out in 1870, the Second Empire collapsed, and the Third Republic was proclaimed. On March 18, 1871, the republican Paris Commune was established to fight the Germans in France as well as to fight the Army of Versailles, which had remained loyal to Napoleon III and had concluded an armistice with the Germans that was judged to be dishonourable by the members of the Commune. Courbet, who had been recently elected president of the artists' federation and was charged with reopening the museums and organizing the annual Salon, took part in the revolutionary activities of the Commune. Instead of opening the museums, he decided to protect the major public monuments, especially the Sèvres china factory and the palace at Fontainebleau, for Paris had been under constant bombardment by the Germans. Alarmed by the excesses of the Commune, he resigned May 2.

The Commune had voted to destroy the column in the Place Vendôme commemorating the Grand Army of Napoleon Bonaparte, and it carried out the decision on May 16. But on May 28 the Commune was crushed by the Army of Versailles, and on June 7 Courbet was arrested at the home of a friend. Because he was thought to be responsible for the demolition of the column, he was brought before a military court. As he had often made known his disgust of the militarism represented by the monument, he was charged with having been the instigator, although he had in no way participated in its destruction. A scapegoat was needed, and Courbet was arbitrarily chosen, despite his protests and those of the persons actually responsible for the demolition, who had fled to England. He was sentenced to six months in prison, and, thanks to the intervention of Adolphe Thiers, head of the provisional government of the French Republic, he was given a minimum fine of 500 francs. He served his sentence first at the Sainte-Pélagie prison, and, when he became seriously ill, he was moved to a clinic near Paris. Once freed, he hastened to Ornans in the hope of regaining his strength.

Arrest and imprisonment

When Thiers resigned in 1872, the Bonapartist deputies reopened Courbet's case and sued him for the cost of rebuilding the column. His entire personal property and all his paintings were seized, and he was fined 500,000 gold francs. Having no other alternative but to leave France because he could not pay the fine, he crossed the border into Switzerland on July 23, 1873, and settled in the small town of Fleurier. He set to work again, but, feeling unsafe so close to France, he first went to Vevey and then to La Tour-de-Peilz, where he bought an old inn, appropriately named the Bon-Port (Safe Arrival). There he died on December 31, 1877, at the age of 58, physically and morally exhausted by illness and disappointment.

Flight to Switzerland

Courbet's reputation has continued to grow since his death. His detractors often judge his art only on the basis of his Socialism, ignoring the fact that his political beliefs grew out of his generosity and compassion. His work, however, exerted much influence on the modern movements that followed him. He offered succeeding generations of painters not so much a new technique as a whole new philosophy. The aim of painting was not, as previous schools had maintained, to embellish or idealize reality but to reproduce it accurately. Courbet succeeded in ridding painting of its artistic clichés, contrived idealism, and timeworn models.

MAJOR WORKS

"Courbet au chien noir" ("Courbet with a Black Dog," 1842; Musée du Petit Palais, Paris); "Juliette Courbet" (1844; Musée du Petit Palais, Paris); "L'Homme à la pipe" ("Man with a Pipe," *c.* 1846; Musée Fabre, Montpellier France); "St. Nicolas ressuscitant les petits enfants" (1847; Église de Saules, near Ornans, France); "L'Après-dinée à Ornans" ("After Dinner at Ornans," 1849; Musée des Beaux-Arts, Lille, France); "Un Enterrement à Ornans" ("Burial at Ornans," 1849; Louvre, Paris); "Les Paysans de Flagey revenant de la foire" ("The Peasants of Flagey Returning from the Fair," 1850; Musée des Beaux-Arts et d'Archéologie, Besançon, France); "Young Ladies from the Village" ("Les Demoiselles de village," 1851–52; Metropolitan Museum of Art, New York); "Les Baigneuses" (1853; Musée Fabre, Montpellier, France); "Portrait de Bruyas" (1854; Musée Fabre, Montpellier, France); "La Rencontre ou bonjour, Monsieur Courbet" (1854; Musée Fabre, Montpellier, France); "La Roche de dix-heures" (*c.* 1854; Louvre); "Les Cribleuses de blé" ("The Winnowers," 1854; Musée des Beaux-Arts, Nantes); "L'Atelier du peintre, Allégorie réelle" ("The Artist's Studio, a Real Allegory of a Seven-Year Long Phase of My Artistic Life," 1855; Louvre); "Mère Grégoire" (1855; Art Institute, Chicago); "Les Demoiselles desbords de la Seine" ("Young Women on the Banks of the Seine," 1856; Musée du Petit Palais, Paris); "La Toilette de la mariée" ("The Bride at Her Toilet," 1859; Smith College Museum of Art, Northampton, Massachusetts); "La Diligence dans la neige" (1860; National Gallery, London); "Combat de cerfs. Le Rut du printemps" ("Battle Between Two Stags," 1861; Louvre); "The Trellis" (*c.* 1863; Toledo Museum of Art, Toledo, Ohio); "Proudhon et ses enfants" ("Proudhon and His Family," 1865; Musée du Petit Palais, Paris); "La Remise de chevreuils" ("Roe-Deer in Cover by the Plaisir-Fontaine Stream," 1866; Louvre); "The Woman with a Parrot" ("La Femme au Perroquet," 1866; Metropolitan Museum of Art, New York); "Les Dormeuses" ("Sleeping Women," 1866; Musée du Petit Palais, Paris); "La Source ou baigneuse à la source" (1868; Louvre); "Falaise d'Étretat après l'orage" ("The Cliffs at Étretat," 1870; Louvre); "Mer orageuse" ("La Vague," 1870; Louvre); "Portrait de Regis Courbet" (1874; Musée du Petit Palais, Paris).

BIBLIOGRAPHY. G. CASTAGNARY, *Gustave Courbet et la colonne Vendôme* (1883), was a courageous attempt of the author to absolve Courbet of the unjust charge that he took part in the demolition of the Vendôme column. GEORGES RIAT, *Gustave Courbet, peintre* (1906), written in collaboration with JULIETTE COURBET, who made important documents available to the author, is the definitive work upon which most later studies were based. BELA LAZARE, *Courbet et son influence à l'étranger* (1911), is a detailed analysis of Courbet's influence on the painters of central Europe. CHARLES LEGER, *Courbet* (1929), is an abundantly illustrated volume published on the 50th anniversary of Courbet's death (in French) that often sheds new light on the art of the painter. MARCEL ZAHAR, *Gustave Courbet* (1950), analyzes with rare perception the character of the man and the genius of the painter (in French). GERSTLE MACK, *Gustave Courbet* (1951), is a meticulously and heavily documented study of the artist (in English). ROBERT FERNIER, *Gustave Courbet* (Eng. trans. 1969), analyzes Courbet the craftsman; it is written as an homage to the genius of the painter whom the author greatly admired. The Société des "Amis de Gustave Courbet" *Bulletin* (issued twice a year since 1947) describes the society's

activities and records all new information on Courbet's life and work.

(R.J.F.)

Courts and the Judiciary

This article deals with the operations of the judicial branch of government. It explores some of the fundamental relationships of this branch with legislative and executive branches and analyzes the functions, the structure and organization, and finally, the key personnel of courts: the judges.

The approach is comparative, contrasting and comparing the systems of the two predominant legal traditions of the contemporary world: first, that of the common law, represented by England, the United States, Canada, Australia, and other nations deriving their legal systems from the English model; and second, that of the civil law, as represented by nations of western Europe and Latin America and certain Asian and African nations that have modelled their legal systems on western European patterns. Reference is made to the legal institutions in the Soviet Union and other Communist nations that display distinctive characteristics different from those of the civil-law tradition, from which, basically, they developed. Furthermore, at the end of the article there has been appended a section dealing more specifically with systems in Communist countries.

FUNCTIONS OF COURTS

Keeping peace. The primary function of any court system in any nation—to help keep domestic peace—is so obvious that it is rarely considered or mentioned. If there were no agency to decide impartially and authoritatively whether a man had committed a crime and, if so, what should be done with him, other persons offended by his conduct would take the law into their own hands and proceed to punish him according to their uncontrolled discretion. If there were no agency empowered to decide private disputes impartially and authoritatively, self-help, quickly degenerating into physical violence, would prevail and anarchy would result. Not even a primitive society could survive under such conditions. All social order would be destroyed. In this most basic sense, courts constitute an essential element in society's machinery for keeping peace.

Deciding controversies. In the course of helping to keep the peace, courts are called upon to decide controversies. If, in a criminal case, the defendant denies committing the acts charged against him, the court must choose between his version of the facts and the prosecution's; and if he asserts that his conduct did not constitute a crime, the court must decide whether his view of the law or the prosecution's is correct. In a civil case, if the defendant disputes the plaintiff's account of what happened between them—for example, whether they entered into a certain agreement—or if he disputes the plaintiff's view of the legal significance of whatever occurred—for example, whether the agreement was legally binding—the court again must choose between the contentions of the parties. The issues presented to, and decided by, the court may be either factual, legal, or both.

The court's administrative role

It would be a mistake, however, to assume that courts spend all of their time deciding controversies. Many cases brought before them are not contested. They represent potential, rather than actual, controversies in which the court's role is more administrative than adjudicatory. The mere existence of a court renders unnecessary any very frequent exercise of its powers. The fact that it operates by known rules and with reasonably predictable results leads those who might otherwise engage in controversy to compose their differences.

Most people arrested and charged with crime in the common-law world plead guilty. If they do so understandingly and without coercion of any sort, there is no need to determine guilt, for the sole question is whether the defendant should go to jail, pay a fine, or be subjected to other corrective treatment. In civil-law countries some judicial inquiry into the question of guilt or innocence is required even after a confession. But the inquiry is brief and tends to be perfunctory. The main problem to be resolved, usually without contest, is what sentence should be imposed.

The vast majority of civil cases are also uncontested or, at least, are settled before trial. The court keeps the calendar moving, sometimes encouraging settlement, and decides such questions of law or fact as are presented by the parties; but the number of cases actually tried is small compared to the number settled.

Most divorce cases are uncontested, both parties usually being anxious to terminate the marriage and often agreeing on related questions concerning support and the custody of children. All the court does in such cases is to review what the parties have agreed upon and give its official approval.

Many other uncontested matters come before courts, such as the adoption of children, the distribution of assets in trusts and estates, and the setting up of corporations. Occasionally questions of law or fact arise that have to be decided by the court, but normally all that is required is judicial supervision and approval.

Judicial lawmaking. As courts decide controversies they create an important by-product beyond the peaceful settlement of disputes; that is, the development of rules for future cases. Law is thus made not only by legislatures but also by the courts.

To an extent that varies greatly between common-law and civil-law nations, all courts apply pre-existing rules formulated by legislative bodies. In the course of doing so, they interpret those rules, sometimes distorting them, sometimes transforming them from generalities to specifics, sometimes filling gaps to cover situations never considered by the original lawmakers. The judicial decisions embodying these interpretations then become controlling for future cases, sometimes to the extent of virtually supplanting the legislative enactments themselves.

The uses of experience in the law: stare decisis

This is one aspect of the doctrine of precedent, or, as it is sometimes called, stare decisis (literally, "to stand by decided matters"). Judges follow earlier decisions, not only to save themselves the effort of working out fresh solutions for the same problems each time they recur but also, and primarily, because their goal is to render uniform and stable justice. If one individual is dealt with in a certain way today, the theory is that another individual engaging in substantially identical conduct under substantially identical conditions tomorrow or a month or year hence should be dealt with in the same way. This, reduced to its essentials, is all that precedent means.

In civil-law nations all judicial decisions are, in theory, based upon legislative enactments, and the doctrine of judicial precedent does not apply. Practice, however, departs from theory. While there are comprehensive legislative codes in these countries, supposedly covering almost every aspect of human conduct and supplying ready-made answers for all problems that can arise, in fact many of the provisions are exceedingly vague and are sometimes almost meaningless until applied to concrete situations, when judicial interpretation gives them specific meaning. Furthermore, the legislative codes cannot anticipate all situations that may arise and come before the courts. The gaps in legislative schemes must be and are filled by judicial decisions, for no court in any nation is likely to refuse to decide a case on the ground that it has not been told in advance the answers to the questions presented to it. Decisions dealing with circumstances unforeseen by the codes and giving specific meaning to vague legislative provisions are published in most civil-law countries and are frequently referred to by lawyers and relied upon by judges. They are not considered "binding," but neither are they forgotten or disregarded. In actual practice, they have almost as much influence as statutory interpretations in nations that formally adhere to the doctrine of stare decisis.

It remains true that in common-law countries judicial lawmaking is more pervasive and more frankly acknowledged than in civil-law countries. In addition to rendering decisions that authoritatively interpret statutes, the courts

of these nations have created a vast body of law without any statutory foundation whatever. Centuries ago, when there was no legislation to guide them, judges began to decide cases in accordance with their own conceptions of justice. Later judges followed them, deciding like cases in the same manner but distinguishing earlier cases when dissimilar factors were discovered in the cases before them. The later cases also became precedents to be followed in still later cases presenting substantially similar fact patterns. So the process has continued over centuries and is still continuing. The total accumulation of all these judicial decisions is what constitutes "the common law" —the by-product of judges deciding cases and setting forth their reasons. In the common-law nations, legislation is, as a result, more limited in scope than in the civil-law countries. It does not purport to provide for all possibilities but leaves large areas of conduct to be governed solely by judge-made law.

To speak of precedent as "binding" even in common-law systems is misleading. As already noted, earlier decisions can be and are distinguished when judges conclude that they are based upon situations different from those before the court in later cases. Even more significant, earlier decisions can be overruled by the courts that rendered them (not by courts lower in the judicial hierarchy) when the judges conclude that they have proved to be so erroneous or unwise as to be unsuited for current or future application. The Supreme Court of the United States has overruled many of its own earlier decisions, to the consternation of those who yearn for a rigid separation of powers and who are unable to accept the inevitability of judicial lawmaking. Many of these overrulings are in the field of constitutional law, in which legislative correction of an erroneous judicial interpretation of the Constitution is impossible and in which the only alternative is the exceedingly slow, cumbersome, costly, and difficult process of constitutional amendment. Nevertheless, the power to overrule decisions is not restricted to constitutional interpretations. It extends to areas of purely statutory and purely judge-made law as well, areas in which legislative action would be equally capable of accomplishing needed changes. Even in England, which has no written constitution and which has traditionally followed a far more rigid doctrine of stare decisis than the United States, the House of Lords, in its role as the highest court, has announced its intention of departing from precedent "in appropriate cases."

Conflicting views of the court's role

The desirability of judicial lawmaking has long been the subject of lively debate in both civil- and common-law countries. That courts should not arrogate to themselves unrestricted legislative power is universally accepted. But when existing statutes and precedents are outmoded or barbarous as applied to specific cases before the courts, should not judges be able to change the law in order to achieve what they conceive to be just results or, stated differently, to avoid what they consider unjust results?

The extent to which the judges should be bound by statutes and case precedents as against their own ethical ideas and concepts of social, political, and economic policy is an important question as is the matter of which should prevail when justice and law appear to the judges to be out of alignment with each other. These are questions upon which reasonable men disagree vigorously even when they are in basic agreement on the proposition that some degree of judicial lawmaking is inevitable. What is mainly at issue is the proper tempo and scope of judicial change. How quickly should judges act to remedy injustice and when should they consider an existing rule to be so established that its alteration calls for constitutional amendment or legislative enactment rather than judicial decision? As many dissenting opinions attest, judges themselves disagree on the answers to these questions, even when they are sitting on the same bench hearing the same case.

Constitutional decisions. In some nations courts not only interpret legislation but determine its validity and in so doing sometimes render statutes inoperative. This happens only in nations that have written constitutions and have developed a doctrine of "judicial supremacy." The prime example is the United States and the classic statement of the doctrine is the Supreme Court's decision in *Marbury* v. *Madison* (1803), in which Chief Justice Marshall said:

> The powers of the legislature are defined and limited; and that those limits may not be mistaken, or forgotten, the Constitution is written. To what purpose are powers limited, and to what purpose is that limitation committed to writing, if these limits may, at any time, be passed by those intended to be restrained? The distinction between a government with limited and unlimited powers, is abolished, if those limits do not confine the persons on whom they are imposed, and if acts prohibited and acts allowed, are of equal obligation. It is a proposition too plain to be contested, that the Constitution controls any legislative act repugnant to it. . . . It is emphatically the province and duty of the judicial department to say what the law is. Those who apply the rule to particular cases, must of necessity expound and interpret that rule. If two laws conflict with each other, the courts must decide on the operation of each.

Armed with the authority asserted at this early date, the Supreme Court of the United States has held many statutes, federal as well as state, unconstitutional and has also invalidated executive actions that violated the Constitution. Even more surprising is the fact that lower courts also possess and exercise the same powers. Whenever a question arises in any U.S. court at any level as to the constitutionality of a statute or executive action, that court is obligated to determine its validity in the course of deciding the case before it. The case may have been brought for the sole and express purpose of testing the constitutionality of the statute or it may be an ordinary civil or criminal case, in which a constitutional question incidental to the main purpose of the proceeding is raised. Of course, when a lower court decides a constitutional question, its decision is subject to appellate review, sometimes at more than one level. When a state statute is challenged as violating the state constitution, the final authority is the supreme court of that state; when a federal or state statute or a state constitutional provision is challenged as violating the Constitution of the United States, the ultimate arbiter is the Supreme Court of the United States.

In a few American states, questions as to the constitutional validity of a statute may be referred in abstract form to the state's highest court by the chief executive or the legislature for an advisory opinion. This, however, is unusual and, in any event, supplementary to the normal procedure of raising and deciding constitutional questions. The normal pattern is for a constitutional question to be raised at the trial-court level in the context of a genuine controversy and to be decided finally on appellate review of the trial-court decision.

Other methods of constitutional adjudication

The U.S. pattern of constitutional adjudication is not followed in all nations that have written constitutions. In some, such as Germany, there is a special court at the highest level of government that handles only constitutional questions and to which all such questions are referred as soon as they arise. A constitutional question may be referred to the special court in abstract form for a declaratory opinion by a procedure similar to that prevailing in the minority of U.S. states that allow advisory opinions.

In other nations, written constitutions may be in effect but not accompanied by any conception that their authoritative interpretation is a judicial function. Legislative bodies, rather than courts, act as the guardians and interpreters of the constitution, being guided by their provisions but not bound by them in any realistic sense.

Finally, there are some nations, such as England, that have no written constitutions. Here parliamentary supremacy clearly prevails. The courts have no power to invalidate statutes, although they can and do interpret them.

Procedural rule making. Distinct from the type of lawmaking just described is a more conscious and explicit type of judicial legislation and one that is less controversial. It is directed toward the rules of procedure by which

the courts operate. This is a technical area in which expert knowledge of the type possessed by judges and lawyers is needed; in which constant attention to detail is required; and in which major problems of social, economic, or political policy are seldom encountered. Some legislative bodies, able or willing to devote only sporadic attention to the day-to-day problems of the management of litigation, have delegated the power to regulate procedure to the courts themselves. This is not ad hoc judicial lawmaking as a by-product of deciding cases but openly acknowledged promulgation of general rules for the future, in legislative form, by courts rather than legislatures.

An outstanding example of judicial rule making is found in the United States where Congress has delegated to the Supreme Court broad power to formulate rules of civil, criminal, and appellate procedure for the federal courts. The Supreme Court also has and exercises the power to amend the rules from time to time as experience indicates that changes are desirable. Congress reserves the power to veto the rules so promulgated but has felt no need to exercise it.

Other legislative bodies, including those of some American states and most of the nations of continental Europe, have been unwilling to repose equal trust in the courts and have retained for themselves the power to regulate procedure. The results have been varied. Courts sometimes become so immersed in day-to-day decision making that they fail to pay adequate attention to the proper functioning of the judicial machinery and so perpetuate rules that are unduly rigid, unrealistic, and unsuited to the needs of litigants, which was the case in England and the American colonies during the 18th and first part of the 19th century. When such a condition occurs, reform through legislative action is indicated. Apart from the occasional necessity of major sweeping changes, however, experience in the common-law countries, at least, indicates that procedural rule making is better vested in the courts than in legislative bodies.

Review of administrative decisions. Existing alongside the courts in any nation are administrative agencies of various kinds. Some do substantially the same kind of work as is done by courts and in substantially the same manner; some have quite different functions such as the issuing of licenses and the payment of welfare benefits.

The relationship between such agencies and regular courts differs markedly between common-law countries and civil-law countries. In common-law countries the actions of administrative agencies are subject to review in the ordinary courts. If the agency is one that decides controversies in substantially the same manner as a court, but in a different and more limited area, judicial control takes much the same form of appellate review as is provided for the decisions of lower courts. The objective of reviewing the record of proceedings is to determine whether the administrative agency acted within the scope of its jurisdiction, whether there was any evidence to support its conclusion, and whether the governing law was correctly interpreted and applied. Administrative decisions are seldom upset by the courts because of a belief on the part of most judges that administrative agencies have special expertise in the area of their specialty. However, they can be and occasionally are upset, thus underscoring the large degree of judicial control over other agencies of government that characterizes common-law systems. If the administrative agency does not engage in formal adjudication, it produces no record of proceedings for judicial review. Nevertheless its action can be challenged in court by way of trial rather than appeal. The same problems are presented for judicial determination: did the agency act within its jurisdiction, did it correctly follow the law, and was there any rational or factual basis for its action?

In many civil-law countries, the ordinary courts have no control over administrative agencies. Their decisions are reviewed by a special tribunal that is engaged exclusively in that work and that has nothing to do with cases of the type that come into the courts. Its function is solely appellate and solely within the specialized areas entrusted to the administrative agencies. The prototype of this type of tribunal is the Conseil d'État of France.

Enforcement of judicial decisions. The method of enforcing a judicial decision depends upon its nature. If it does nothing more than declare legal rights, as is true of a simple divorce decree (merely severing marital ties, not awarding alimony or the custody of children), or a declaratory judgment (for example, interpreting a contract or a statute), no enforcement is needed. If a judgment orders a party to do or refrain from doing a certain act, as happens when an injunction is issued, the court itself takes the first step in enforcing the judgment by holding in contempt anyone who refuses to obey its order and sentencing him to pay a fine or go to jail. Thereafter, enforcement is in the hands of the executive branch of government, acting through its correctional authorities.

In routine criminal cases and in civil cases that result in the award of money damages, courts have little to do with the enforcement of their judgments. That is the function of the executive branch of government, acting through sheriffs, marshals, jailers, and similar officials. The courts themselves have no machinery for enforcement.

The limits of judicial power

Some judgments are extremely controversial, as was the case with the decision of the Supreme Court of the United States ordering racial desegregation of the schools. When voluntary compliance with such a judgment is refused, forcible methods of enforcement are necessary, sometimes even extending to the deployment of armed forces under the control of the executive branch of the government. The withdrawal of executive support seldom occurs, even when decisions are directed against the executive branch of government itself; but when such executive support is withheld, the courts are rendered impotent. Judges, being aware of their limited power, seldom render decisions that they know to be so lacking in support that they will not be enforced.

COURT STRUCTURE AND ORGANIZATION

There are many different types of courts and many ways to classify and describe them. Basic distinctions must be made between civil and criminal courts, between courts of general jurisdiction and those of limited jurisdiction, and between trial and appellate courts.

Criminal courts. Criminal courts deal with persons accused of crime, deciding whether they are guilty and, if so, determining the consequences they shall suffer. Prosecution is on behalf of the public, represented by some official such as a district attorney, procurator, or a police officer. Courts are also public agencies, but in this instance they stand neutral between the prosecution and the defense, their objective being to decide between the two in accordance with law.

In civil-law countries a more active role is assigned to the judge and a more passive role to counsel than in common-law countries. In the common-law courts, in which the "adversary" procedure prevails, the lawyers for both sides bear responsibility for producing evidence and they do most of the questioning of witnesses. In civil-law countries, "inquisitorial" procedure prevails, with judges doing most of the questioning of witnesses and having an independent responsibility to discover the facts. This difference pertains more to procedure rather than function.

If a man has been found guilty, he is sentenced, again according to law and within limits fixed by legislation. The objective is not so much to wreak vengeance upon the offender as to rehabilitate him and deter others from following his example. Hence the most common sentences are fines, short terms of imprisonment, and probation (which allows the offender to remain at large but under supervision). In extremely serious cases, the goal may be to prevent the offender from committing further crimes, which may call for a long term of imprisonment or even capital punishment. The death penalty, however, is gradually disappearing from the criminal codes of civilized nations.

Criminal proceedings in any nation inevitably have

some educational impact on defendants and upon members of the general public. In Communist nations education is a conscious and primary goal. A basic provision of Soviet law declares:

> By all its activities the court shall educate the citizens of the U.S.S.R. in the spirit of devotion to the Motherland and the cause of communism in the spirit of strict and undeviating observance of Soviet laws, of care for socialist property, of labor discipline, of honesty toward public and social duty, of respect for the rights, honor and dignity of citizens, for the rules of socialist common-life.

Civil courts. Civil courts deal with "private" controversies, as where two individuals (or corporations) are in dispute over the terms of a contract or over who shall bear responsibility for an auto accident. Ordinarily the public is not a party as in criminal proceedings, for it has no interest beyond providing just rules for decision and a forum where the dispute can be impartially and peacefully resolved.

It is possible, however, for the government to be involved in civil litigation if it stands in the same relation to a private party as another individual might stand. Thus, if a postal truck should run down a pedestrian, the government might be sued civilly by the injured man; or if the government contracted to purchase supplies that turned out to be defective, it might sue the dealer for damages in a civil court.

Distinction between civil and criminal actions

The objective of a civil action is not punishment or correction of the defendant or the setting of an example to others but rather to restore the parties so far as possible to the positions they would have occupied had no legal wrong been committed. The most common civil remedy is a judgment for money damages but there are others, such as an injunction ordering the defendant to do or refrain from doing a certain act or a judgment restoring property to its rightful owner.

Civil claims do not ordinarily arise out of criminal acts. A man who breaks his contract with another or who causes him a physical injury through negligence may have committed no crime but only a civil wrong for which he may not be prosecuted criminally by the public.

There are, however, areas of overlap, for a single incident may give rise to both civil liability and criminal prosecution. In some nations, such as France, both types of responsibility can be determined in a single proceeding under a concept known as adhesion by which the injured party is allowed to assert his civil claim in the criminal prosecution, agreeing to abide by its outcome. This removes the necessity of two separate trials. In common-law countries there is no such procedure, even though civil and criminal jurisdiction may be merged in a single court. Two separate actions must be brought, independent of each other.

Courts of general jurisdiction. Although there are some courts that handle only criminal cases and others that handle only civil cases, a more common pattern is for a single court to be vested with both civil and criminal jurisdiction. Such is the High Court of England and such are many of the trial courts found in U.S. states. Often these tribunals are called courts of general jurisdiction, signifying that they can deal with almost any type of controversy, although in fact they may not have jurisdiction over certain types of cases assigned to specialized tribunals. Often such courts are also described as superior courts, because they are empowered to handle serious criminal cases and important civil cases involving large amounts of money.

Even if a court possesses general or very broad jurisdiction, it may nevertheless be organized into specialized branches, one handling criminal cases, another handling civil cases, another handling juvenile cases, and so forth. The advantage of such an arrangement is that judges can be transferred from one type of work to another, and cases do not fail to be heard for having been instituted in the wrong branch since they can be transferred administratively with relatively little effort.

Courts of limited jurisdiction. Specialized tribunals of many kinds exist, varying from nation to nation. Some deal only with the administration of the estates of deceased persons (probate courts), some only with disputes between merchants (commercial courts), some only with disputes between employers and employees (labour courts). All are courts of limited jurisdiction. Deserving of special mention because of their importance are juvenile courts, empowered to deal with misconduct by children and sometimes also with the neglect or maltreatment of children. Their procedure is much more informal than that of adult criminal courts, and the facilities available to them for the pretrial detention of children and for their incarceration, if necessary after trial, are different. The emphasis is on salvaging children, not punishing them.

Traffic courts also deserve mention because they are so common. They process motor vehicle offenses such as speeding and improper parking. Their procedure is summary and their volume of cases heavy. Contested trials are relatively infrequent.

Finally, in most jurisdictions there are what are called, unfortunately and for want of a better term, "inferior" courts. These are often manned by part-time judges who are not trained in the law. They handle minor civil cases involving small sums of money, such as bill collections, and minor criminal cases carrying light penalties, such as simple assaults. In addition to finally disposing of minor criminal cases, such courts may handle the early phases of more serious criminal cases—fixing bail, advising defendants of their rights, appointing counsel, and conducting preliminary hearings to determine whether the evidence is sufficient to justify holding defendants for trial in higher "superior" courts.

Appellate courts. The tribunals described thus far are trial courts or "courts of first instance." They see the parties, hear the witnesses, receive the evidence, find the facts, apply the law, and determine the outcome.

Above them, to review their work and correct their errors, are appellate courts. These are usually collegiate bodies, consisting of several judges instead of the single judge who usually presides over a trial court. The jurisdiction of the appellate courts is usually general; specialized appellate tribunals handling, for example, only criminal appeals or only civil appeals are rare, although not unknown. The Conseil d'État of France and the Constitutional Court of Germany have already been mentioned as examples of specialization.

Appellate review is not automatic. It must be sought by some party aggrieved by the judgment in the court below. For that reason, and because an appeal may be both expensive and useless, there are far fewer appeals than trials and, if successive appeals are available, as is often the case, far fewer second appeals than original appeals. Judicial systems are organized on a hierarchical basis: at the bottom are numerous trial courts scattered throughout the nation; above them are a smaller number of first level appellate courts, usually organized on a regional basis; and at the apex is a single court of last resort.

Varieties of appellate review

There are three basic types of appellate review. The first consists of a retrial of the case, with the appellate court hearing the evidence for the second time, making fresh findings of fact, and in general proceeding in much the same manner as the court that originally rendered the judgment under attack. This "trial de novo" is used in common-law countries for the first stage of review but only when the trial in the first instance was conducted by an "inferior" court—one typically manned by a part-time judge or two or more such judges, empowered to try only minor cases and keeping no adequate record of its proceedings.

The second type of review is based in part on a "dossier," which is a record compiled in the court below of the evidence received and the findings made there. The reviewing court has the power to rehear the same witnesses again or to supplement their testimony by taking additional evidence, but it need not and frequently does not do so, being content to rely on the record already made in reaching its own findings of fact and conclusions of law. This type of proceeding prevails generally in civil-

law countries for the first stage of appellate review, even when the original trial was conducted in a superior court, staffed by professional judges, and empowered to try important or serious cases.

The third type of review is based solely on a written record of proceedings in the court or courts below. The reviewing court does not itself receive evidence directly but concentrates its effort on discovering from the record whether any errors were committed of such a serious nature as to require reversal or modification of the judgment under attack or a new trial in the court below. The emphasis is on questions of law (both procedural and substantive) rather than on questions of fact. This type of review prevails both in civil-law nations and common-law nations at the highest appellate level. It is also used in common-law nations at lower levels when the judgment of a superior court is under attack. The purpose of this type of review is not merely to assure that correct results are reached in individual cases but also to clarify and expound the law in the manner described earlier. Lower courts have little to do with the development of the law, for they ordinarily do not write or publish opinions. The highest appellate courts do, and it is their opinions that become the guidelines for future cases.

Courts in federal systems. Many nations, such as England, France, and Japan, have unitary judicial systems with all courts (that is, regular courts as distinguished from administrative bodies) fitting into a single national hierarchy of tribunals along the lines just described. Other nations, organized on a federal basis, tend to have more complicated court structures, reflecting the fragmentation of governmental powers between the central authority and the local authorities. In the United States, for example, there are 51 separate judicial systems, one for each state and another for the federal government. To a limited extent, the jurisdiction of the federal courts is exclusive of that exercised by the state courts, but there are large areas of overlap and duplication. At the top level is the Supreme Court of the United States, hearing appeals not only from the lower federal courts but also from state courts insofar as they present federal questions arising under the Constitution of the United States or under federal statutes or treaties. If a case in a state court involves only a question of state law— for example, the interpretation of a state statute—the ultimate authority is the state supreme court, and no appeal is possible to the Supreme Court of the United States.

Court structure in a federal form of government need not be as complicated as that in the United States. It is possible to have only one set of courts for the nation, operated by the central government and handling all cases that arise under state law as well as federal law.

Another possibility is for each state or province to have its own system of courts, handling all questions of federal as well as state law, and for the central government to maintain only a single supreme court to decide questions as to the relationship of the central authority and the local authorities or as to the relationship among the local authorities themselves. This is the pattern in Canada and Australia.

Conflict of laws problems

Another complication resulting from a federal form of government is that questions involving conflict of laws arise with great frequency. Such questions concern the choice to be made between the law of one jurisdiction and another as the rule for decision in a particular case. Even in a unitary system, such problems cannot be avoided, for an English court may be called upon to try a case arising from a transaction that took place in France and to decide whether English or French law should govern. Such problems arise much more often, however, in federal systems, where laws differ from state to state and people move about very freely. Their activities in one state sometimes become the subject of a lawsuit in another, requiring the court to decide which law should apply.

JUDGES

A court is a complex institution whose functioning depends upon many people: not only the judge but also the parties, their lawyers, witnesses, clerks, bailiffs, probation officers, administrators, and many others, including, in certain types of cases, jurors. Nevertheless, the central figure in any court is the judge.

Judges vary enormously, not only from nation to nation but often within a single nation. For example, a rural justice of the peace in the United States—untrained in the law, serving part-time, sitting alone in his work clothes in a makeshift courtroom, collecting small fees or receiving a pittance for salary, trying a succession of routine traffic cases and little else—obviously bears little resemblance to a justice of the Supreme Court of the United States—a full-time, well-paid, black-robed professional, assisted by law clerks and secretaries, sitting in a marble palace with eight colleagues and deciding at the highest appellate level only questions of profound national importance. Yet both men are judges.

The levels of legal training

Lay judges. In some civil-law countries, judges at all levels are professionally trained in the law, but in many other nations they are not. In England, part-time, lay judges outnumber full-time professional judges by about 60 to 1. Called magistrates or justices of the peace, they dispose of about 97 percent of all criminal cases in that nation and do so with general public satisfaction and the approbation of most lawyers. Professional judges deal only with the most serious crimes, which are relatively few in number; most of their time is devoted to civil cases. England places unusually heavy reliance on lay judges, but they are far from unknown in the courts of many other nations, particularly at the lowest trial level. This is as true in the U.S.S.R. as it is in the United States. There is considerable diversity in the way laymen are chosen and used in judicial work. In the U.S.S.R. and the United States, for example, lay judges are popularly elected for limited terms, whereas in England they are appointed by the lord chancellor to serve until retirement or removal. In the U.S.S.R. and England the lay judges serve intermittently in panels on a rotating basis for short periods, whereas in the United States they sit alone and continuously. In the U.S.S.R. lay judges (who are called assessors) always sit with professional judges; in England, they sometimes do; and in the United States, they never do. In some underdeveloped nations, few judges at any level are legally trained. They are more likely to be priests, for the law they administer is mainly derived from religious teaching, and religion and secular government are often not sharply differentiated. The vast majority of nations that use lay judges at the lowest trial level, however, insist upon professionally trained judges at higher levels: in trial courts of general jurisdiction and in appellate courts.

Professional judges in the civil-law tradition. Professional judges in civil-law countries are markedly different in background and outlook from professional judges in common-law countries. Both are law trained and both perform substantially the same functions, but there the similarities cease. In a typical civil-law country, a man graduating from law school makes a choice between a judicial career and a career as a private lawyer. If he chooses the former and is able to pass an examination, he is appointed to the judiciary by the minister of justice (a political officer) and enters service in his early 20s. His first assignment is to a low-level court; thereafter, he works his way up the judicial ladder as far as he can until his retirement on a pension. His promotions and assignments depend upon the way his performance is regarded by a council of senior judges, or sometimes upon the judgment of the minister of justice, who may or may not exercise his powers disinterestedly and on the basis of merit. The civil-law judge, in short, is a civil servant.

Professional judges in the common-law tradition. In common-law nations, the path to judicial office is quite different. Upon completion of his formal education, a man typically spends 15, 20, or 25 years in the private practice of law or, less commonly, in law teaching or governmental legal service; then, at about age 50, he becomes a judge. He takes no competitive examination but is appointed or elected to office. In England the appoint-

Political status of the judiciary

ive system prevails for all levels of judges, including even lay magistrates. Appointments are primarily under the control of the lord chancellor, who, although a cabinet officer, is also the highest judge of the realm. They are kept surprisingly free from party politics. In the United States, the appointive method is used in federal courts and some state courts, but it tends to be highly political. Appointments are made by the chief executive of the nation or state and are frequently subject to legislative approval. In many states, judges are popularly elected, sometimes on nonpartisan ballots, sometimes on partisan ballots with all the trappings of traditional political contests. In an attempt to de-emphasize political considerations and yet maintain some measure of popular control over the selection of judges, a third method of judicial selection has been devised and is slowly growing in popularity. Called the Missouri Plan, it involves the creation of a nominating commission that screens judicial candidates and submits to the appointing authority a limited number of names of men considered qualified. The appointing authority must make his choice from the list submitted. The man chosen as judge then assumes office for a limited time, and, after the conclusion of this probationary period, he stands for "election" for a much longer term. He does not run against any other candidate but only "against his own record."

In common-law countries, a man does not necessarily enter the judiciary at a low level; he may be appointed or elected to his nation's highest court or to one of its intermediate courts. He does not look forward to any regular pattern of promotion, nor is he necessarily assured of long tenure with ultimate retirement on a pension. In some courts, life tenure is provided, usually subject to mandatory retirement at a fixed age. In others, tenure is limited to a stated term of years. At the conclusion of his term, if not mandatorily retired earlier, the judge must be re-elected or re-appointed if he is to continue.

While in office, the common-law judge enjoys greater power and prestige and more independence than his civil-law counterpart. He occupies a position to which most members of his profession aspire. He is not subject to outside supervision and inspection by any council of judges or by a minister of justice; nor is he liable to be transferred by action of such an official from court to court or place to place. The only administrative control over him is that exercised by judicial colleagues, whose powers of management are generally slight, being limited to such matters as requiring periodical reports of pending cases and arranging for temporary (and usually consensual) transfers of judges between courts when factors such as illness or congested calendars require them. Only if a judge misbehaves very badly is he in danger of disciplinary sanctions and then usually only by way of criminal prosecution for his misdeeds or legislative impeachment and trial, resulting in removal from office—a very cumbersome, slow, ill-defined, inflexible, ineffective, and seldom used procedure. In parts of the United States, newer and more expeditious methods of judicial discipline are developing in which senior judges are vested with power to impose sanctions ranging from reprimand to removal from office of erring colleagues. They are also vested with power to retire judges who have become physically or mentally unfit to discharge their duties.

Except at the very highest appellate level, common-law judges are no less subject than their civil-law counterparts to appellate reversals of their judgments. But appellate review cannot fairly be regarded as discipline. It is designed to protect the rights of litigants; to clarify, expound, and develop the law; and to help and guide rather than reprimand lower court judges. (D.K.)

Other judicial officials. In most countries there are other officials who serve the court. Court clerks, who are responsible for case records and documents, and bailiffs, who are in charge of keeping order, are found in most judicial systems. Also prevalent are officers who prosecute cases in the government's name: states attorneys and district attorneys in the United States, procurators-general in the U.S.S.R., and *procureurs généraux* in France.

Probation officers are found in many countries including the U.S. and Japan. Notaries in France, Italy, and the U.S.S.R. have greater powers than their counterparts in the U.S. In fact, they perform many services carried out by lawyers in the common-law system such as drafting and verifying wills and contracts and preparing petitions for presentation in court.

Certain countries have officials that are particularly indigenous to their country or legal system. France, for example, has a *juge d'instruction*, who is responsible for the preliminary investigative proceedings prior to a criminal trial.

THE STRUCTURE AND STATUS OF THE JUDICIARY UNDER COMMUNISM

Although the essential legal institutions of the Soviet Union and other Communist countries are based on the civil-law system, certain features are unique. These characteristics are partly the result of the Soviet Union's attitudes toward law that antedate the Soviet system, but mostly they result from the attempt to reconcile Marxist theory with the institutional needs of a modern society.

According to Marx and his followers the legal system, like all other governmental structures and instruments of class oppression, would "wither away" in a Communist society; thus the courts that existed after the Revolution were considered temporary institutions, required only during the transition to Communism. The ordinary and traditional business of the courts was carried on by the so-called people's courts, while "revolutionary tribunals" dealt with individuals the government considered to be political opponents. A nonjudicial body in the hands of the secret police (at first called Cheka, later OGPU and NKVD), operating in the style of an administrative agency, also heard cases and handed out sentences—usually of the severest kind.

The Soviet three-level court system

In 1921 some capitalist measures were temporarily introduced to revive the economy, and this necessitated some stabilization of the legal system and its institutions. A three-level system of courts with civil and criminal jurisdiction was established in 1922 for the Russian Republic, which in the same year formed a federation with the other soviet republics under its jurisdiction, making up the Union of Soviet Socialist Republics. A new constitution created a federal court—the U.S.S.R. Supreme Court—and a federal judiciary act of 1924 established uniform principles for the judiciary throughout the republics, patterned largely after the system adopted by the Russian Republic. The basic structure of the courts laid down at that time has remained essentially the same to the present, with some minor changes and reforms.

The "people's courts" on the local level are courts of original jurisdiction for minor criminal cases and a large number of civil cases. The next level, the provincial courts, receive appeals from the people's courts and have original jurisdiction over political and serious civil and criminal cases. The highest level in each republic is its supreme court, which hears appeals from the provincial courts, disciplines lower courts, and has some original jurisdiction over extremely serious cases.

On all three levels, appellate cases are tried by a court consisting of three full-time judges, whereas one judge and two lay judges, or assessors, preside over cases on first hearing. Judges of the people's courts are popularly elected every five years and judges on the other two levels are "elected" by soviets (bodies combining legislative and executive functions) of the corresponding levels of government. All judges may be recalled before the expiration of their terms by those who elected them.

The federal court system is twofold. There are courts called military tribunals that deal with charges against men in the armed forces and with charges of espionage brought against civilians. The other federal body is the U.S.S.R. Supreme Court—the highest judicial body—which has original jurisdiction in a few special cases relating to the survival of the regime, appellate power over the decisions of the supreme courts of the republics or decisions of the military tribunals, and the right to issue

directives to all inferior courts in matters of administration of justice on the basis of a series of its decisions. Although Soviet legal theory is patterned after that of civil law countries in that it does not recognize judicial lawmaking, these Supreme Court directives function as a source of law and are binding on all courts. The status of the judiciary in the Soviet Union has undergone some changes that parallel the institutional changes since the early days of the Revolution. The system organized in 1922 had the stated purpose of safeguarding the conquests of the Revolution and establishing the dictatorship of the proletariat. Judges were called upon to use their "revolutionary conscience" in deciding cases, and the doctrine of impartiality and independence of the judiciary was repudiated. With the passage of time, however, the Soviet rulers found the need for legal institutions of a stable nature increasing rather than decreasing, and the goal of the legal system was changed from protection of a particular class to protection of the socialist order and the rights of all citizens. Although the role of the judiciary is still conceived of as a political task, there is some acceptance of the idea that judges should be independent and impartial. Marxist philosophy notwithstanding, the U.S.S.R. and other Socialist countries are confronted with a growing need for legal institutions to fulfill many of the same functions as those in the West. One attempt to fill this need in recent years has been the appearance of "social organizations," such as the "comrades' courts," which are described as voluntary organizations using persuasion and social influence to deal with matters that would otherwise come before a court. But these organizations are party controlled, have only limited power to impose sanctions, and do not appear at present to offer an effective alternative to the type of legal institutions that have been developing within Soviet society.

The status of judges

The other Communist countries, both in eastern Europe and Asia, adopted legal institutions largely patterned after the Soviet model. Since Stalin's death, however, there have been some modifications in the eastern European countries, coinciding with the reforms in the civil and criminal codes adopted by the Soviet Union in the late 1950s and early 1960s. Chinese leaders, however, have resisted efforts to codify their laws, preferring flexibility in their courts, and they have abandoned the policy of copying Soviet legal patterns.

BIBLIOGRAPHY. S. BEDFORD, *The Faces of Justice* (1961), discussion of how cases are handled in England, Germany, Austria, Switzerland, and France; J.H. MERRYMAN, *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America* (1969), a summary of the principles and institutions in civil-law countries; D. KARLEN *et al.*, *Anglo-American Criminal Justice* (1967), comparison of two judicial systems in the area of criminal law; OLIVER WENDELL HOLMES, *The Common Law*, ed. by M. DEWOLFE HOWE (1963), classic treatment of the growth of law through the judicial decisions; B.N. CARDOZO, *The Nature of the Judicial Process* (1921), explanation by a distinguished judge of how an appellate court reaches its decisions; and ROSCOE POUND, *Organization of Courts* (1940), detailed treatment of court structure in the U.S.

Overviews of legal institutions within specific countries include R.M. JACKSON, *The Machinery of Justice in England*, 5th ed. (1967); L. MAYERS, *The American Legal System*, rev. ed. (1964); H.J. BERMAN, *Justice in the U.S.S.R.*, rev. ed. (1963); M. CAPPELLETTI *et al.*, *The Italian Legal System* (1967); and H.P. DUBEY, *A Short History of the Judicial Systems of India and Some Foreign Countries* (1968).

Covenant

The concept of covenant—that is, a binding promise—is of far-reaching importance in the relations between individuals, groups, and nations. It has social, legal, religious, and other aspects. This article is concerned primarily with the term in its special religious sense and especially with its role in Judaism and Christianity.

NATURE AND SIGNIFICANCE

Covenants in the ancient world were solemn agreements by which societies attempted to regularize the behaviour of both individuals and social organizations, particularly in those contexts in which social control was either inadequate or nonexistent. Though ancient pre-Greek civilizations apparently never developed a descriptive theory of covenants, analysis of covenant forms and the ancient use of language yields a definition that essentially is the same as that found in modern law. It is a promise or agreement under consideration, usually under seal or guarantee between two parties, and the seal or symbol of guarantee is that which distinguishes covenant from modern contract.

Definition of covenant

The concept of covenant has been of enormous importance in the biblical tradition; from it there is derived the long traditional division of the Bible into the Old and New Testaments (Covenants). In postbiblical Judaism and sporadically in Christianity, the concept of covenant has been a major source and foundation of religious thought and especially of the concept of the religious community, but the nature and content of covenant ideas have undergone an extremely complex history of change, adaptation, and elaboration.

Though both covenant and law in the ancient world were means by which obligation was both established and sanctioned, and are often virtually identified with each other in modern scholarly literature, there are, nevertheless, very important contrasts between the two that should not be obscured. A covenant is a promise that is sanctioned by an oath. This promise in turn was accompanied by an appeal to a deity or deities to "see" or "watch over" the behaviour of the one who has sworn, and to punish any violation of the covenant by bringing into action the curses stipulated or implied in the swearing of the oath. Legal procedure, on the other hand, may be entirely secular, for law characteristically does not require that each member of the legal community voluntarily swear an oath to obey the law. Further, in ordinary legal procedure the sanctions of the law are carried out by appropriate agencies of the society itself, not by transcendent powers beyond the control of man and society.

Because a person can bind only himself by an oath, covenants in the ancient world were usually unilateral. In circumstances in which it was desirable to establish a parity (equivalence) treaty, such as in rare cases in political life, the parity was obtained by the simple device of what might be termed a double covenant, in which both parties would bind themselves to identical obligations, and neither was therefore subjected to the other.

The oath was usually accompanied by a ritual or symbolic act that might take any of an enormous range of forms. One of the most frequent of these was the ritual identification of the promisor with a sacrificial animal, so that the slaughter and perhaps dismemberment of the animal dramatized the fate of the promisor if he were to violate the covenant.

ORIGIN AND FUNCTION OF COVENANTS

Origins. That covenants most probably originated in remote prehistoric times is indicated by the fact that they were already well-developed political instruments by the 3rd millennium BC. To judge from later parallels and from the modern observations of anthropologists, covenants may very well have developed at least in part out of marriage contracts between exogamous tribes or bands; *i.e.*, those groups that stayed within the required patterns of intermarriage. Whether or not this was the case, the most important functions of covenants for 1,000 years before the 13th-century BC Sinai covenant (see below) had to do with the creation of new relationships, both familial and political. Though the old theory of "social contract" —*i.e.*, the basic agreement about the social and political order—as the basis of large social organizations has not for some time been much in favour among social scientists, very early historical evidence seems increasingly to suggest that covenants may have been much more instrumental in society than has been realized.

Early functions of covenants

Typically, so far as existing sources now reveal, a covenant between social groups regularized in advance the relationships between two societies after one had been subjugated by a superior coercive force, usually by mili-

tary action or the threat thereof. In the Mari documents (18th-century-BC archives from the palace at Mari in Syria), such a covenant was called a *salimum*, a "peace," probably because the promises made by the vanquished brought to an end the necessity of military operations against the vassal ruler or state. As is the case throughout so much of human history, ancient states characteristically seem to have regarded their neighbors as either enemies or vassals. Thus it is not surprising that covenants made under duress had little vitality, particularly when the terms of the covenant called for a considerable annual tribute to the overlord state.

Late Bronze Age developments. About the beginning of the late Bronze Age (*c.* 1500 BC), there occurred a major step forward in both the form and the concept of political covenants as is attested by treaties of the Hittite Empire of Asia Minor. Though the realities of political life were probably little changed, since the foreign policy of the Hittite Empire was primarily military, the structure of suzerainty treaties from this time on included rather strenuous efforts to demonstrate that the vassal's obligations to the Hittite overlord were really founded upon the former's self-interest, not merely upon the brute military force of the latter.

Hittite treaties

By far the most evidence for international treaties in the ancient world comes from Hittite sources, which were contemporary with the events that preceded and led up to the formation of the ancient Israelite federation of tribes in Palestine. The treaty form in written texts was highly developed and flexible but usually exhibited the following structure: preamble, historical prologue, stipulations, provisions for deposit and public reading, witnesses, and curses and blessings formulas. (1) The preamble names the overlord who grants the treaty-covenant to the vassal. The titles and laudatory epithets of the Great King are also given. (2) The historical prologue describes the previous relationships between the two parties in some detail, usually emphasizing the benevolent acts of the Great King toward the vassal. Thus the covenant is based upon the demonstrated benefits that have already been received and therefore holds out the expectation of continuing advantage for faithful obedience to the covenant. There is an implication that obedience to obligations is based upon gratitude. (3) The stipulations, which in form are much like those of the ancient Mesopotamian law codes (case law), define in advance the obligation of the vassal in certain circumstances. In addition, there are also generalized statements of obligation of a type that has been called "apodictic law" (regulations in the form of a command). The obligations deal particularly with military assistance, the treatment of fugitives, and foreign policy. Treaty relationships with other independent states are a violation of covenant. (4) Provision is made for deposit of the treaty in the temple and for periodic public reading. Because the temple is the "house of the god," the written document was placed there for the watchful attention of the deity. The treaty obligations, however, were also binding upon the vassal's citizenry, and so at stipulated intervals the text was read to the assembly, both as a reminder and a warning. (5) The list of witnesses included, in addition to the major deities of both states, deified elements of the natural world, such as mountains, rivers, heaven and earth, winds, and clouds. The witnesses were those powers that were believed to be beyond human control and upon which man and society were regarded as completely dependent. They were invoked to apply the appropriate sanctions of the covenant. (6) The curses and blessings formulas are the sanctions that furnish not only negative but also positive motivations for obedience. They include the natural and historical calamities beyond human control, such as disease, famine, death without posterity, and destruction of the society itself. The blessings are of course the opposite: prosperity, peace, long life, continuity of kingship and society.

In view of the obsession with rituals that characterized Hittite culture, some elaborate ceremony probably accompanied the ratification of covenant, such as the account of one preserved in the document known as "The Soldiers' Oath," but it is not described in existing covenant texts.

Scholars in Europe and America in the 20th century have seen an astounding similarity between this treaty structure and the biblical traditions of the Sinai covenant. Publication of texts in the mid-1950s was followed by an enormous amount of scholarly discussion, but as yet no conclusions can be said to represent a scholarly consensus. The formal similarity to biblical traditions cannot be denied, but the problem of what historical conclusions can be drawn from the formal similarities is highly sensitive and controversial. While the following synthesis is a probable, and historically plausible, interpretation, it must be admitted that other possibilities can by no means be excluded.

THE ORIGIN AND DEVELOPMENT OF BIBLICAL COVENANTS: JUDAISM

The Sinai covenant. *Historical background.* The 100 years between 1250 and 1150 BC saw the complete destruction, or reduction to virtual impotence, of every major political state in the eastern Mediterranean region and the beginning of a "dark age" that has yielded very few written materials from which historical conclusions can be drawn. The reasons for the universal catastrophe are far from clear, but the reversion of society to communities of peasants and shepherds with a subsistence level economy can be well illustrated archaeologically. The earliest biblical traditions illustrate the conditions in Palestine at this time, though it is a difficult task to distinguish genuine ancient traditions from the use of the past by biblical writers to give religious validity to social realities or institutions of much later date.

In view of the highly elaborate social structure of the old Bronze Age states—with its apex in the military aristocracy, a highly complex priesthood, and ritual—and the equally complex social structure of the many local enclaves and tribes—each with its particular god—the monotheistic and ethically centred religious ideology of early Israel has been regarded for millennia as a miracle of "revelation," which cannot be explained on the basis of usual historical principles and concepts. Yet, ancient Israel was an historically existent community created, and precariously maintained, by a unity of which the religious ideology was the foundation for two centuries, until military considerations resulted in the formation of a political centralization of power about 1000 BC. The covenant tradition is the only instrument by which the effective functioning of that unity can be understood, and its importance is underlined by the biblical traditions themselves. The structure of the Hittite treaties now makes available an historical precedent that enables scholars to understand the structure of early Israelite thought and consequently its functional operation in history.

The covenant tradition as an instrument of unity

The covenant at Sinai. The Decalogue (The Ten Commandments) given by Yahweh, the God of the Israelites, at Sinai, plus the various traditions associated with earliest Israel yield all of the important elements of the Hittite treaty form but in an extremely succinct and simple form. Yahweh is identified as the covenant giver, and the historical prologue is the only possible one according to the ancient traditions: the announcement that it is this God who delivered the assembled group from bondage in Egypt (in the 13th century BC). This delivery is a free, voluntary act of the deity that forms the basis of the obligations that the community can either accept together with a lasting relationship to that God or reject, thus entailing a permanent hostility (hatred) between the God and human beings. It is the common relationship to a single sovereign God that furnished the basis for a radically new kind of community, which grew with rapidity first in Transjordan, then in Palestine proper, until it included virtually all the nonurban population of the region.

The new community was the answer, temporarily at least, to the old dilemma of civilization: how to maintain peace among a large and diverse population, perform the necessary social functions of cooperation and protection, and control individual attacks upon the security and

property of others without the enormous and expensive paraphernalia of political bureaucracy, military machine, and the ruinous tax collector. It was, for all functional purposes, the Kingdom (or Rule) of Yahweh, which excluded the deification of any other factor in human history or nature that was of importance to human life and well-being. The Sinai covenant marked the beginnings of nearly all the various theological themes that were to be so greatly elaborated upon in the following millennia: the Providence, or Grace, of God; the Kingdom of God; the sin of man and the wrath of God; the Holy People as the community of God; the rewards and punishments of the obedient and the disobedient respectively; and above all, the ethical norms as the essence of divine command over against the universal pagan obsession with proper ritual as the normative expression of man's subjection to the divine will.

Modern functional equivalents of the Ten Commandments

The Sinaitic covenant stipulations may be expressed in modern functional terms in the following manner: (1) The commandment to have no other gods involves the obligation to refuse subjection to all other social and human concerns and their symbolization in art forms so as to give them a position of parity or superiority to Yahweh and his commands. (2) The commandment not to take the name of God in vain emphasizes the unconditional sanctity of oaths that Yahweh was called upon to guarantee and enforce. (3) The commandment to observe the sabbath, the seventh day, the original social function of which is still unknown, could very well have grown out of a common village custom, for even in Rome in the 1st century BC, good farming practice permitted work animals and slaves to rest every eighth day—and this is precisely the interpretation given in Deut. 5:14. (4) The commandment to honour father and mother emphasizes the treatment of parents with respect and deference, which must have been of particular importance in a time of social upheaval and polarization. (5) The commandment not to kill meant that killing of persons by persons, even by accident if it involved negligence, was a usurpation of the divine sovereignty over persons. Contrary to modern reinterpretations, among opponents of capital punishment and pacifists, this could not include execution of persons for crime or killing of the enemy in warfare, for in both cases human beings were acting as the agents of Yahweh under divine command, just as the various officials of states have long carried out similar functions without incurring personal guilt for their acts. (6) Other commandments against theft, adultery, and false witness categorically prohibit acts that call into question the security of property, of family relationships and true lineal succession, and the integrity and therefore the justice of juridical procedures in society. (7) Finally, the prohibition of coveting what one's neighbour has excludes an enormous range of social attitudes and motivations that modern man now takes for granted as normal, if not essential.

The Ten Commandments as ethical obligations

Most, if not all, of the Ten Commandments are ethical obligations of which violations are very difficult if not impossible for society to detect, much less to enforce or punish. The Sinai covenant, therefore, marked the beginnings of a systematic recognition that the well-being of a community cannot be based merely upon socially organized force, nor can the political power structure be regarded, as in ancient pagan states, as the manifestation of the divine, transcendent order of the universe.

Post-Sinai covenants. Traces remain in the biblical traditions to indicate that the new community formed from a "rabble" at Sinai was in very short time joined by a considerable part of the population of Transjordan and Palestine proper. After the destruction (in the late 13th century) of the military chiefdoms ruled by Sihon and Og in the area east of the Jordan River, the Hebrews held a covenant ceremony at Shittim (northeast of the Dead Sea), which has been greatly elaborated upon in tradition as the "second giving of the Law," Deuteronomy. Though it is true that the Book of Deuteronomy from the 7th century BC exhibits the same basic structure as that of the old covenant form, it is at present impossible to reconstruct the original form or content of the Shittim covenant. It may be presumed that entry into the community by covenant was followed by the allotment of land as tenured fiefs from Yahweh and the organization of the population into "tribes." This organization probably was the last event of the Hebrew leader Moses' life, and the sequel in the more important covenant at Shechem (northwest of the Dead Sea) took place under the leadership of Joshua, the successor of Moses.

Shechem evidently had had an important covenant tradition long before Israel existed. The name of its god, Baal Berit ("Lord of the Covenant"), presupposes some kind of covenant basis for the local social structure, just as a considerable segment of the population can be shown to have originated from Anatolia.

The Shechem covenant

The Shechem covenant narrative has been preserved at least in part in Joshua, in which Joshua appeals to the family and clan heads to choose between the new dominion of Yahweh and the continuation of the old ancestral cults of the Amorite tradition "beyond the River." As in the case of the Transjordan covenant at Shittim, this covenant followed the defeat of a coalition of petty kings and evidently the removal of many others according to the list of Joshua. Again, there ensued an allotment of fields and an organization of the population into administrative units called "tribes," each under a *nasi* (literally, "one lifted up").

The entire process from the covenant at Sinai to the unification of perhaps a quarter of a million people by a covenant involving a religious loyalty to a single deity took only a little over one generation. It began with a group of probably considerably less than 1,000 people who left Egypt with Moses.

The subsequent history of the Sinai covenant tradition is very complex. The Book of Deuteronomy preserves slight traces of a covenant-renewal ceremony held every seven years, which is inherently plausible and which would function as a means for obtaining the oath-bound loyalty to Yahweh and his dominion of those who had come into the community from the outside or who had come of age in the intervening period.

The covenants of the Israelite monarchy. (1020–587/586 BC). Since early Israel was a religious confederacy of tribes that bitterly rejected the old military chiefdoms and their religious ideology, which elevated a Baal, or local agricultural deity—the god with the club as a symbol of the supernatural power undergirding the king—to a position of preeminence in the pantheon, it follows that the authentic Yahwist traditions stemming from Moses could not furnish a religious ideology to legitimize the monarchy when it was finally established first under Saul (reigned *c.* 1020–1000 BC) and then successfully under David (reigned *c.* 1000–962). Furthermore, early in David's reign, he had incorporated by military force most of the existing city-states of Palestine and Transjordan into his empire, and that population had never given up the old Bronze Age cults.

It is not surprising, therefore, that this double dilemma of the new political structure should have driven the royal bureaucracy to pre-Mosaic sources as a solution to the problem. One result was the reintroduction of the age-old pagan concept of the king as the "chosen" one of the gods and a radically different—and opposite—concept of covenant, in which it was now Yahweh, not the king or the people, who bound himself by oath. Possibly modelled after old royal covenants by which ancient pagan kings made a grant to their faithful retainers, the Davidic covenant introduced a radically different (and thoroughly pagan) element into the Mosaic tradition, and the two traditions contended with one another for the next 1,000 years.

Since the old Israel-Jacob (pre-Mosaic) traditions also could not furnish an ideological base for unifying the old Israelite and non-Israelite populations under the monarchy, pre-Mosaic epic traditions of Abraham (perhaps 19th–18th centuries BC) were appealed to to furnish the "common ancestor" symbol of unity, and the covenant tradition—no doubt, already a part of that epic—was readapted to bring it up to date. The deity (now identified with Yahweh) bound himself by oath to fulfill certain

promises to Abraham, though the content of the promise, in the form now received, was by and large a description of the historical situation of the Davidic empire. Though it is difficult to see what the social or ideological function may have been, the covenant with Noah (the hero of the Flood) in Genesis exhibits the same structure. The result of all these radical changes in a very short time was a complete confusing of the religious tradition and structure and a permanent deposit of the pre-Mosaic pagan religious ideology into the biblical tradition. It seems virtually certain that the Sinai tradition was itself systematically reinterpreted in the so-called ritual decalogue of Exodus in which it is dogmatically stated that the Sinai obligations were entirely ritual in nature, rather than ethical-functional. The first tables of stone of the Ten Commandments, after all, had been "broken," which in the ancient world was a customary phrase used to indicate the invalidation of binding legal documents.

Reinterpretation of the Decalogue

The next several centuries illustrate the constant battle between the Mosaic and the reintroduced pagan elements. The prophets proclaimed and supported the disintegration (*c.* 922 BC) of the Solomonic empire into a northern (Israel) and a southern (Judah) kingdom as the divine chastisement of Yahweh for gross disobedience. Particularly in the north, which did not retain the Davidic dynasty, the prophets periodically proclaimed the necessity and inevitability of wiping out one royal dynasty after another. Elijah, a 9th-century BC rustic prophet, ridiculed the idea that the Israelites could limp along on both legs—*i.e.*, observe loyalty to both the Yahwistic and the Baal cults. Reforms were carried out occasionally, but not until the time of Josiah, the young king of Judah (late 7th century BC), and the discovery of an old copy of the Mosaic legal-ethical tradition (the Deuteronomic code) in *c.* 621 BC was serious reform undertaken —and there with little permanent success. The preservation of the Mosaic tradition was a function of the destruction of the monarchical state and its religious symbol, the temple, which nearly all the pre-exilic (before 587/586 BC) prophets had predicted.

The post-Exilic covenant tradition. Though the prophet Jeremiah (late 7th century BC) had predicted a "new covenant" written upon the heart (Jeremiah), not until the time of the prophets Ezra and Nehemiah in the 5th century is there another biblical narrative of covenant making, this time one of incalculable importance for the future of both postbiblical Judaism and Christianity and perhaps even for certain aspects of political theory or practice in the West (*e.g.*, "Covenant" of the United Nations, Mayflower Compact, and constitutions).

The account in Nehemiah is not so much that of a covenant as it is of a constitutional convention, the purpose of which was to establish as binding law the complex of traditions that had been preserved and recorded as the "law of God which was given by Moses, the servant of God" (Nehemiah). It is a one-party enactment by the authorities and representatives of the community, in which Yahweh appears only as the deity addressed in the long historical prologue in the form of a prayer. The content is a recapitulation of the Deuteronomic history (interpretations of the 7th-century BC document), narrating the benevolent acts of Yahweh and the sin and punishment of the people. In order to avoid the curses, and obtain the blessings, the community resolved henceforth to observe the "law of God." From this time on, the dominant concept of covenant in Judaism identifies it with circumcision, the ritual by which on the eighth day of his life, the male Jew becomes obligated to obey the law of Moses, the *berit* (covenant). The Sinai covenant had become permanently identified with the accumulation of legal-ritual tradition, and the community was identified not as the complex variety of all those who wished and accepted the rule of God but as the ethnic group of those who were heirs of the promise to Abraham in direct lineal (and fictitious) descent.

ORIGIN AND DEVELOPMENT OF COVENANT: CHRISTIANITY

The New Testament tradition of the covenant. The cup of wine at the Last Supper of Jesus and his disciples before Jesus' crucifixion is identified in all New Testament sources as the (new) covenant by Jesus himself, but in spite of millennia-long controversy, theological elaboration, and discussion, the nature and meaning of the covenant has never been adequately understood historically, and the variety of interpretations regarding covenant in the New Testament itself indicates that very early in the tradition it had become a problem. Here it is possible only to indicate some significant associations that might explain why it was called a "covenant" and how the ancient Sinaitic tradition was radically renewed but the basic structure retained.

First, it has been noted that a most important aspect of covenant traditions common to most ancient cultures was the ritual identification of the oath taker with the sacrificial victim. The identification of the bread and the wine with the body and blood of Christ at the Last Supper apparently was interpreted in this sense, so that the subsequent death of the victim entails the symbolic death—the ultimate curse for breach of covenant—of all those who were thus identified with the victim. Consequently, the curses of the law were nullified. The death of Jesus thus becomes in the Christian proclamation the centre of the historical narrative—the historical prologue of the covenant—leading up to the covenant enactment, or the *sacramentum*, to use the Latin term of the early church, which in secular use at that time meant primarily the soldier's oath of loyalty to the emperor (see above *Late Bronze Age developments*). The Christian covenant was thus a highly complex historical act that brought about a relationship of the believer to Christ whose (normally) unseen Glory was identified with that of God himself, whose Lordship was viewed as operational in history, and whose community (of believers) was identified with the Kingdom (Dominion or Rule) of God. If God in the Old Testament could rule without kings, God could, for the New Testament writers, rule without the elaborate structure of the accumulated legal traditions. They were regarded as valuable for edification and for warning but no longer as having binding validity. The anathema, or curse, was no longer tied to the definitions of legal violation but rather to rejection of God's rule in Christ. The community in turn was no longer the lineal descent group with a parochial ritual tradition but the assembly (*ekklēsia*) of those who had through the covenant accepted a relationship to the dominion of Christ.

The death of Jesus in relation to the covenant

The obligations could not, in the New Testament viewpoint, be again defined in legal terms, nor could they be enforced by social power structures, which could deal only with external formal acts, not with the basic springs of behaviour, such as love or hate. The content of obligation was thus not defined; instead, in the Sermon on the Mount (Matthew) and other New Testament literature, it is the criteria (motivations, ethical norms, personality traits) by which the rule of God is recognized upon which the emphasis falls. The presumption is that anyone who is capable of recognizing the rule of God in his experience in society will also be capable of understanding what the nature of his obligation will be in specific circumstances. The curses and blessings alike are then postponed until the final judgment. The motivations of fear of punishment and hope of reward are irrelevant to the daily routine of ethical choice, which is thus not only possible (*i.e.*, not prescribed in advance by legal definition) but unavoidable and also necessary to make responsible ethical decisions in a world that is characterized by cultural diversity and change.

The post-apostolic church. Covenant concepts in early Christian theology apparently centred on the transference of the Davidic covenant to the Messianic figure—*i.e.*, Christ. The fundamental theological problem of the early church was to validate the authority of Christ against both paganism and Judaism and to maintain the authority of the new religious community. After the great theologian Augustine (354–430), little attention was given to covenants until the Reformation in the 16th century. Though Luther (1483–1546) referred to and discussed the biblical covenants, it was never of particular importance to his theology. It is rather in Reformed the-

ology, particularly that of John Calvin (1509–64) and the later Puritans of the 17th century, that its further elaboration took place. One aspect of the use of covenant may be cited in the famed Mayflower Compact of Nov. 11, 1620 (drawn up by the Pilgrims, Separatists from the Church of England) by which a "civil body politic" was formed that would in turn enact laws and offices for the general good.

Significance of covenants in Reformed Protestant Theology

The theological elaboration of covenant in Puritan and Separatist theology centred on the themes of election, grace, and Baptism. It is curiously ironic that covenant enactment, such as the Mayflower Compact, became historically operative but remained essentially secular, while the religious covenant became predominantly a theological concept associated particularly with Baptism—the ritual means by which a person became a participant in the covenant of grace. The essential elements in the biblical covenant—*i.e.*, that of free, voluntary acceptance of ethical obligation on the basis of and as response to past experience—has virtually always given way to covenant as fixed religious dogma that legitimizes the social structure. Covenant historically has been a means by which new communities are formed, particularly in times of rapid change, social dislocation, or political breakdown. Covenants have rarely been the actual instruments by which societies actually functioned for long, but they are extremely frequent as ideological foundations for sociopolitical legitimacy.

COVENANT IN OTHER RELIGIONS

Islām. Covenants (*mīthāq*, *'ahd*) were of great importance in the formative period of Islām (7th century AD, or 1st century AH—after the Hegira, Muḥammad's flight from Mecca to Medina). More than 700 verses of the Qur'ān, the Muslim sacred scripture, have to do with various aspects of covenant relationships. As one recent Muslim writer, Sayyid Qutb, states, Islām combines both the Old and the New Testaments (covenants) and the Last Covenant, of Islām, as well. All revelation from Adam to Muḥammad is regarded by Muslims as a unit, mediated through a series of prophets, or messengers, with whom God made a covenant: Noah, Abraham, Moses, and Jesus. Though the concept is difficult, it seems that the prophet in each case was given a revelation and a religion to which he covenanted with God to witness faithfully. This concept of a covenant of the prophets conveys the conviction of the unity of revelation as well as the unity of God in past history.

On the second level, the Muslim community itself is often regarded as being composed of those who have accepted the covenant with God. In this connection, the grace, or providence, of God in nature or creation is of great importance. In addition to this view is the repeated emphasis upon the doctrine that God alone is man's sole benefactor, and for these reasons the response of gratitude is an important element in the structure of the covenant. It is also necessary that rewards and punishments are included. These are predominantly, as in the Christian concepts, focussed upon the hereafter, paradise, and hell, though not exclusively so. The recipients of the rewards and punishments are described as those who obey or disobey Allāh's (God's) commands, which include prayer, paying the *zakāt* (head tax: an obligatory charity), belief in the messengers of Allāh, fearing God alone, and refraining from theft, adultery, murder, and false witness. They are further obligated to show kindness to parents and to strive in the cause of God with their persons and property.

Importance of the covenant in the formative period of Islām

On the historical and social level, it seems quite certain that the community of the formative period in Islām was based on covenant acts, in which persons or groups formally proclaimed their acceptance of Muḥammad's message and swore an oath of loyalty, accepting the obligations outlined above. References to the clasp of hands indicate that this was probably regarded as the formal act of commitment and acceptance by the community. In later Muslim theology, as in Christianity, the covenant idea seems to have been of comparatively little importance.

Other religions. It seems that only in the religions stemming from the biblical tradition is covenant of central importance. Though gods are often invoked as guarantors of promises sworn to in Iranian and Indic (Hindu) religious traditions, the covenant with a deity or the community as a covenant-bound one apparently was of relatively little importance, or possibly the concept has not been recovered by modern scholarship. The great importance of Mithra in early Iranian religion as god of the covenant, and Mitrā-Varuṇa in Indic (Hindu) religion suggests that such concepts may have been more important than is now realized. Thus, modern scholarship has yet to indicate the importance of the covenant concept in Indo-Iranian and other religions.

BIBLIOGRAPHY. W. BEYERLIN, *Origins and History of the Oldest Sinaitic Traditions* (Eng. trans. 1966); G.E. MENDENHALL, "Covenant Forms in Israelite Tradition," *Biblical Archaeologist*, 17:50–76 (1954), "Covenant," *Interpreter's Dictionary of the Bible*, 1:714–723 (1962), and *The Tenth Generation: The Origins of the Biblical Tradition* (1971); R.C. DARNELL, *Idea of Divine Covenant in the Qur'an* (1970); D.J. MCCARTHY, *Treaty and Covenant* (1963); "Egyptian and Hittite Treaties," in J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 2nd ed. pp. 199–206 (1955); V. KOROSEC, *Hethitische Staatsverträge: Ein Beitrag zu ihrer juristischen Wertung* (1931); D.R. HILLERS, *Covenant: The History of a Biblical Idea* (1969); M. WEINFELD, "The Covenant of Grant in the Old Testament and in the Ancient Near East," *Journal of the American Oriental Society*, 90: 184–203 (1970); A. GOETZE, *Kleinasien*, 2nd ed. (1957); M. WEIPPERT, *Die Landnahme der israelitischen Stämme in der neueren wissenschaftlichen Diskussion* (1967), an excellent summary of recent scholarly discussion; K. BALTZER, *Das Bundesformular* (1960); A. JAUBERT, *La Notion d'alliance dans le Judaïsme aux abords de l'ère chrétienne* (1963); I. GERSHEVITCH (ed. and trans.), *The Avestan Hymn to Mithra* (1959).

(G.E.Me.)

Covilhã, Pêro da

A Portuguese traveller, Pêro da Covilhã was the first European known to have explored both western India and East Africa. Born at Covilhã in Beira *c.* 1460, he served the duque de Medina-Sidonia in Seville for six or seven years, returning to Portugal with the duke's brother late in 1474 or early in 1475, when he passed into the service of King Afonso V, first as a junior squire and then as squire, serving with horse and arms. He accompanied the king when he claimed the Castilian throne and was proclaimed at Plasencia, and he was present at the Battle of Toro. He also escorted the King on a fruitless journey to France to seek aid from Louis XI. On Afonso's death, Pêro served his son John II as a squire of the royal guard and was employed as a confidential messenger to Spain. He was also sent on two missions to North Africa, one, in the guise of a merchant, to seek the friendship of the ruler of Tlemcen, and the other to Fez to buy horses for Dom Manuel, later king.

Attempt to find Prester John

John II hoped to profit from the spice trade of India and to make contact with the Christian ruler of Abyssinia, identified with the semimythical Prester John. Abyssinians had already visited Rome and even the Iberian Peninsula. John had sent Diogo Cão (Diogo Cam) down the west coast of Africa, and he had discovered the Congo and sailed beyond, but his belief that he had reached or was about to reach the cape proved unfounded. John then ordered Bartolomeu Dias (*q.v.*) to pursue Cão's explorations. He also decided to send travellers by land to report on the location and trade of India and Abyssinia. This move may have resulted from reports received in 1486 in Benin (a kingdom on the west coast of Africa), referring to a great ruler far to the east. Pêro was chosen for the mission to India, and Afonso de Paiva, a squire who spoke Arabic, was to seek Prester John and discover a route from Guinea to Abyssinia. The men left Portugal in May 1487 with letters of credit on Italian bankers; they reached Barcelona and sailed to Naples and Rhodes, where they assumed the guise of honey merchants and sailed to Alexandria. They became ill, and their wares were seized, but they bought other goods and went to Cairo, joining a group of North Africans travelling to

Aden. There they separated, Pêro going to India, reaching Cannanore, Calicut, and Goa. He then returned to Ormuz, in Persia, sometime between October 1489 and March 1490. Meanwhile, Afonso de Paiva had reached Abyssinia. The two had proposed to meet at Cairo. Pêro arrived there about the end of 1490 or early 1491 and received news of his companion's death. Meanwhile, John II had sent two messengers to Cairo to instruct Pêro to return when the mission was completed. Pêro wrote a letter to John about his experiences and continued on to Abyssinia. One of the messengers accompanied him to Ormuz, where they separated. Pêro made his way to the Red Sea. Disguised as a Muslim, he visited Mecca and Medina. He also saw Mount Sinai, reaching Zeila in 1492 or 1493, whence he passed by caravan to Abyssinia, where he was destined to spend the rest of his life.

Visit to Muslim holy places

Pêro was received by the Abyssinian ruler Emperor Eskender and was well treated and made governor of a district. He was not, however, allowed to leave the country. Some years later the Abyssinian regent, Queen Helena, sent an Armenian named Matthew to Portugal. He reached Afonso de Albuquerque at Goa in 1512 and was in Portugal in 1514. It was then decided to send a Portuguese embassy to Abyssinia. The first ambassador died, and his successor, Dom Rodrigo de Lima, and his party left from India in 1517 and finally reached the Emperor's camp in December 1520. They found Pêro old but robust, and he served them as guide and interpreter. When they returned in 1524, Pêro and his wife and family accompanied them for part of the way, and he sent his 23-year-old son with Dom Rodrigo to be educated in Portugal. Pêro died sometime after 1526.

BIBLIOGRAPHY. Most of what is known of Pêro da Covilhã derives from the account of FRANCISCO ALVARES, chaplain to the embassy. His *Verdadera informaçam das terras do Preste Joam* was printed (1540), and also in a different version in Ramusio's collection (1550). See also Alvares' *Prester John of the Indies*, rev. and ed. by C.F. BECKINGHAM and G.W.B. HUNTINGFORD, 2 vol. (1961); and GASPAR CORREA, *Lendas da India*, 4 vol. (1858–64).

(H.V.L.)

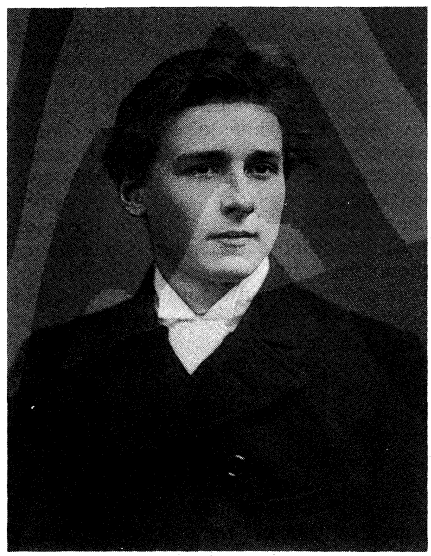
Craig, Edward Gordon

The life and work of Edward Gordon Craig were clearly and closely interwoven. The son of an actress, and "of" the theatre from birth, he worked in three distinct yet overlapping areas of artistic endeavour: as actor, as director-designer, and, most important, as dramatic theorist. Critics often labelled him an idealistic scene designer whose work was impractical in terms of existing theatre architecture. Certainly his visions were incapable of realization at the time of conception because of limited technology and lack of appropriate materials. But most of his life was devoted to a "theatre of the future," and the inspiration behind much of today's theatre practice and theory clearly stems from Craig's pioneering work.

Born in Stevenage, Hertfordshire, in England, on January 16, 1872, Craig was the second child of a liaison between Ellen Terry, who was to become one of England's most celebrated actresses, and Edward William Godwin, architect, theatre critic, and designer. Like Edith (the other child of Godwin and Ellen Terry), their son was given the name Craig, which he adopted as a surname. He made his official debut as an actor on September 28, 1889, as a member of Henry Irving's company at the Lyceum Theatre in London. During his nine-year association with Irving, Craig appeared in a variety of roles, and he also worked in touring companies between Lyceum seasons, all of which theatrical experience was of inestimable value in the formulation of his aesthetic theories.

Debut as an actor

Craig was a handsome young man with somewhat delicate features. Consequently, he was cast in such roles as Romeo, Cassio, Hamlet, Lorenzo, and Malcolm. He complained that he "looked like an actor" and noted with relief that, when he left off acting, he began to lose the "girlish face which I dislike so much . . . and by 40 I had improved it a bit more." As well as performing, Craig also began to try his hand at directing. His early efforts clearly revealed a debt to Irving (who by the 1890s had become the foremost exponent of an historical-archaeological approach to the visual elements of dramatic production). In 1893 he married May Gibson, who bore him four children, but the marriage did not last.



Craig, 1890.
By courtesy of the Mander and Mitchenson Theatre Collection

Meanwhile, however, Craig became friends with other creative artists, among them the painters William Rothenstein, James Pryde, and William Nicholson; the two last, known as the Beggarstaff Brothers, taught Craig how to use wood-engraving tools. Also among his friends were literary men such as Max Beerbohm, and the musician Martin Shaw. These new friends helped to widen Craig's interest in the theatre to embrace other arts. Shaw, for example, introduced Craig into the Purcell Operatic Society, under whose aegis he produced *Dido and Aeneas* in 1900. This production caused a stir by its new approach to theatre art, for Craig avoided both historical-archaeological scenography and the flat, unimaginative use of electric stage lighting. His keynote was one of simplicity: in setting, costume, lighting, and movement. *Dido*, repeated in 1901 with *The Masque of Love*, was followed in rapid succession by Handel's *Acis and Galatea* (1902), Laurence Housman's nativity play, *Bethlehem* (1902), and, under his mother's management, Henrik Ibsen's *Vikings* and Shakespeare's *Much Ado About Nothing*, both in 1903. (It was the last time he was to work with Ellen Terry. In 1875 she had parted from Godwin; much later Craig was to speculate what the three of them together might have achieved in the theatre.)

Success of his *Dido and Aeneas*

These productions, too, were marked by simplicity and unity of concept, with the emphasis being placed on the movement of actors and of light. Soon to come was the idea of moving scenery, which ultimately led to Craig's vision of a theatre without actors or words, in which the movement of scenery and of light would engender an emotional response in the audience. But these productions, despite their artistic impact, were commercial failures; the financial support that would have permitted Craig to develop his ideas and to experiment was not forthcoming in England. Therefore, toward the end of 1904, he took up an earlier invitation from Count Harry Kessler—the arbiter of taste at the influential Weimar Court—to visit Germany. In December of the same year, in Berlin, he met the dancer Isadora Duncan and began a short love affair with her. In a necessarily subjective and often fictitious account of their relationship, Duncan stressed Craig's restless dedication to his work, and the obsession she unwittingly describes has the ring of truth. He never ceased to admire her dancing, however, recognizing in it the pure movement that was for him an essential ingredient of theatre art.

The following few years were a particularly fruitful pe-

riod for Craig. In 1905 the first statement of his developing theories, *The Art of the Theatre*, was published; in 1906 he designed a production of Ibsen's *Rosmersholm* for Eleonora Duse, the world-renowned Italian actress; and in 1907 he returned to his earlier thoughts about movement. By this time he was able to record those ideas through the medium of etching; fascinated by the new technique, he etched some 15 metal plates within two months. His new theatrical concept was that the entire "scene" should be movable in all parts; both the floor and ceiling were to be composed of squares that, under the control of the artist, could be moved up and down independently or in groups within a constantly changing pattern of light. Thus an emotional response might arise in the audience through the abstract movement of these plastic forms. The etchings illustrating these scenographic concepts were exhibited in 1908 in Florence (*Portfolio of Etchings;* 1908, 1910) and collected, with an essay, in *Scene* (1923). They depicted frozen motion rather than a series of individual, separate stage designs.

Although the vision of the "moving scene" with its vertical movement was clear in Craig's imagination, he could find no way to translate the theory into practice because of his limited technical knowledge. With a growing fear of having his idea stolen, he suppressed any further public display of it and substituted a compromise—screens that moved horizontally. After several years of experiments with models using these screens, he allowed the Irish poet and dramatist William Butler Yeats to construct a screen set in 1911 for his Abbey Theatre in Dublin, a set that was used until 1951. In 1912 Craig co-produced *Hamlet* with the Russian director Konstantin Stanislavsky at the Moscow Art Theatre. This was the first and only use of the moving screens under Craig's own supervision. It also was the last time Craig tried to put his pioneer theories into practice.

Because commercial theatre managers, conservative and traditional, ignored Craig's advanced thinking, he received no financial support for productions other than those few sponsored by noncommercial managements. Such rejection proved to be an advantage, however. It provided more and more a limitless stage—his own fertile imagination—upon which to create his theatre magic. To find expression for the ideas and to disseminate them, he turned increasingly to the graphic arts and the written word. His wood engravings and etchings in *Woodcuts and Some Words* (1924) and *Nothing; or, The Bookplate* (1925) are eminently theatrical. He cut dramatic light into the wood blocks and metal plates from which the illustrations were printed as he would into the black void of the stage. Early in his experience as a graphic artist, from 1898 to 1903, he had published *The Page*, a journal devoted to all the arts paralleling his own discovery of them. Following the successful appearance of *The Art of the Theatre*, expanded and retitled *On the Art of the Theatre* (1911–69), Craig discovered his gift for writing.

Late writings

A particularly fecund period commenced that lasted almost to the very end of his life. In addition to writing a number of books (*Towards a New Theatre*, 1913; *The Theatre Advancing*, 1919) and hundreds of periodical articles, he founded *The Mask* in 1908, which, until its demise in 1929, became not only the first theatre journal but a platform for Craig's rebellious ideas. Using almost 100 pseudonyms in addition to his own name, Craig wrote controversial pieces on the theatre of the future in all its aspects, as well as editorials, historical pieces, and book reviews. Among his many avant-garde ideas were the rejection of the academic mannerisms of the actor and the decorative trappings of both costumes and settings. These rejections, Craig argued, would restore art to the theatre so that it could take its proper place with its sister arts. To accomplish these goals he established a school in 1903 to study the art of the theatre. Unsuccessful at first, the school was reopened in 1913 in Florence but closed the following year with the outbreak of World War I. The function of Craig's school was not only to plan for a new theatrical future but, as part of the training program, to study theatre history to promote greater insight and understanding. In terms of physical play production, only one man was to be in artistic control of the production; the visual emphasis was to be on form, colour, the plastic use of light, and, most important, movement; the last was to be achieved through the use of the horizontally moving screens, for Craig never had managed to find the solution to a stage movable in all parts.

After 1914 he lived chiefly in Italy, moving from city to city with Elena Meo (whom he had met in 1900) and their two children. He designed a production of Ibsen's *Pretenders* in 1926 at the invitation of the Royal Theatre in Copenhagen, for which King Christian X of Denmark conferred on him the Order of the Knights of the Dannebrog for his services to the Danish theatre. Craig's last production was Shakespeare's *Macbeth*, produced in New York in 1928. He refused to go to the United States, however, and the execution of his designs and lighting was carried out by others. Neither of these plays involved Craig's earlier, creative use of stage space.

In 1931 Craig went to live in France and in 1948 made his home in the south of that country, where he wrote his memoirs of the years 1872 to 1907, entitled *Index to the Story of My Days* (1957), and continued to arrange and correct his various papers. He died at Vence, France, on July 29, 1966.

Many of Craig's "impractical" ideas have since become common practice; once a "rebel," he is now seen as a prophet. Even in those areas where modern technology has improved upon Craig's vision, the underlying aesthetic concept is still relevant. One measure of recognition for his life-long, singleminded pursuit of theatre art came in 1956 when Queen Elizabeth II conferred upon him the privilege of being named a Companion of Honour.

BIBLIOGRAPHY. EDWARD CRAIG, *Gordon Craig: The Story of His Life* (1968), is the definitive work; an earlier study, based largely on the Craig collection in the Bibliothèque Nationale, Paris, is DENIS BABLET, *Edward Gordon Craig* (1962; Eng. trans., 1966). Important because of the chronology by Craig himself is JANET LEEPER, *Edward Gordon Craig: Designs for the Theatre* (1948). IFAN KYRLE FLETCHER and ARNOLD ROOD, *Edward Gordon Craig: A Bibliography* (1967), records Craig's written and graphic work. ARNOLD ROOD, "'After the Practise the Theory,' Gordon Craig and Movement," *Theatre Research/Recherches Theatrales*, 11:81–101 (1971), traces the development of Craig's theories regarding movement.

(A.Ro.)

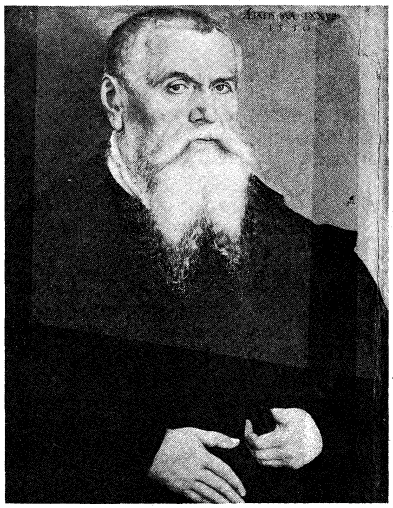
Cranach, Lucas, the Elder

A painter, draftsman, and illustrator, Lucas Cranach the Elder was one of the most important and influential artists in the 16th-century flowering of German painting and graphic arts. Admired by his contemporaries for the speed with which he executed commissions, Cranach, with the help of assistants, worked with such unremitting industry that, despite many losses, some 400 of his paintings have been preserved. The striking individuality of his works and his skill as a portraitist have made him one of the most popular German artists of the period. He exerted a lasting influence on the art of northern and northeastern Germany through his own work, through his numerous pupils, and through his son Lucas Cranach the Younger, himself a famous painter.

The elder Cranach was born in 1472 in Kronach, some 55 miles (90 kilometres) north of Nürnberg. Although only a year younger, he survived Albrecht Dürer, the great genius of German art, by 25 years and, in fact, outlived all the significant German artists of his time.

Cranach's teacher was his father, the painter Hans Müller of Kronach, with whom he worked from 1495 to 1498. He is known to have been in Coburg in 1501, but the earliest of the works that have been preserved date from *c.* 1502, when he was already 30 and ill in Vienna. It was there that he dropped the surname of Müller, calling himself Cranach after his home town, which was then spelled that way. In 1504 he signed "Rest on the Flight into Egypt" with "LC" (Lucas Cranach).

Through his pictures and woodcuts he made an important contribution to the painting and illustrations of the Danube school, the art of the Austrian Danubian region



Cranach, self-portrait, tempera on panel, 1550. In the Uffizi, Florence. 66.36 cm × 48.9 cm
Alinari

Contributions to the Danube school

around Vienna, Passau, and Ratisbon, as well as Salzburg and Innsbruck. In Vienna, then flourishing under the patronage of Emperor Maximilian I, Cranach came in contact with the Humanists teaching at the university, doing portraits of the scholars Dr. Johannes Stephan Reuss (1503, Germanisches Nationalmuseum, Nürnberg) and Dr. Johannes Cuspinian (*c.* 1502–03, Sammlung Oskar Reinhart, Winterthur).

Wittenberg period. Presumably while Cranach was still in Vienna, he received news of his appointment as court painter to the elector Frederick the Wise of Saxony; he must already have been a famous artist, for he was given two-and-a half times the salary paid to his predecessor. In spring 1505 he arrived in Wittenberg, a university town on the Elbe River and seat of the electoral court, where he remained for 45 years, until July 1550. Through Cranach, who received important commissions from three successive electors and caused many young artists to come to Wittenberg, the town became an art centre.

Beginning with a Latin panegyric for him that the schoolmaster Georgius Sibutus published in 1507, Cranach's name is often found in the records, sometimes without commentary presumably because he was so well-known to the public. Dr. Christoph Scheurl, a jurist in the university, prefaced his *Oratio*, published in 1509, with a Latin introduction dedicated to the artist in which he tells of his life and works, characterizing the man Cranach as "friendly, talkative, generous, and obliging" and mentioning his uncommon industriousness that did not let him spend an idle hour. Such qualities made Cranach not only a popular court painter, who maintained himself in the Elector's favour and in that of his two successors, but also an artist who was highly esteemed by the nobility, the wealthy bourgeoisie, and the university professors. It is a further measure of his skill in handling people of all classes and faiths and a token of the popularity of his art that he continued to receive commissions from Catholics even after the onset of the Protestant Reformation.

Vignettes for Luther's writings

The Reformation had begun in 1517 in Wittenberg with Martin Luther's Ninety-Five Theses. Being on friendly terms with Luther, who had been a teacher at Wittenberg University since 1508 and a doctor and professor of Bible exegesis since 1512, Cranach designed, among other things, ornamental vignettes for Luther's writings (1518–19). He also left portraits of Luther, his wife Katherina von Bora, and his parents. Through these and other portraits, Cranach helped form today's image of Luther's circle. Having already created a court art, the portraitist of the leaders of the Reformation thus became also the initiator of a new religious art, the painting of Lutheranism. As such, Cranach did altarpieces and paintings for Lutheran churches.

The output of Cranach's large studio ranged from mere artisan's painting through gilding and decorative work to paintings on building façades and in palace halls. He also designed tapestries, medals, coins, court dresses, escutcheons, decorations on cannon, and painted glass. What was and continues to be important are his altarpieces for Catholic and Protestant churches, his panels with religious and lay scenes, and his portraits. Aside from his numerous paintings, there are more than 100 separate woodcuts by him, as well as woodcuts in the form of book illustrations, and six engravings.

For many years, Cranach's life in Wittenberg ran an even course. In 1508–09 he was at Antwerp, in Flanders, possibly on a diplomatic mission. Cranach probably did not marry until about 1512, when he had reached 40; in 1513, at any rate, he purchased a corner house on the market square that carried with it the lucrative privilege of selling wine. The property was so stately that King Christian II stayed there in 1523. Like a typical entrepreneur in the early days of capitalism, Cranach was active far beyond his duties as court painter. He not only derived an income from the sale of wine but in 1520 bought a pharmacy, which he leased to pharmacist's assistants; he also was, at least temporarily, the owner of a printing press and a book and stationery shop (*c.* 1524). By 1528 he was, after the chancellor Brück, the richest man in Wittenberg.

Success as an entrepreneur

Furthermore, Cranach, whose works were sought after by Protestant and Catholic patrons alike, not only enjoyed the confidence of princes and leaders of the Reformation but was also so highly regarded by his fellow citizens that he sat on the town council from 1519–20 on and served as burgomaster of Wittenberg in 1537–38, 1540–41, and 1543–44.

Fate struck him a few blows, nonetheless. His son Hans, also a painter, died in 1537 in Bologna, and his wife Barbara died in 1541.

Last years. In 1547 the palmy days of Wittenberg came to an end with the defeat of the Protestant princes by the emperor Charles V in the Schmalkaldic War, a religious conflict in which the Elector of Saxony was taken prisoner. The Emperor, who owned one of Cranach's paintings, had sat in 1508 as a child in Flanders for a Cranach portrait. He now had the artist summoned to his headquarters, near occupied Wittenberg, in order to renew acquaintance with him.

Summoned by his own sovereign who was held captive by the Emperor, Cranach joined the Elector in July 1550 at Augsburg, accompanying him the following year to Innsbruck, where he painted continually in order to keep him in good humour. In Innsbruck Cranach met Titian, whom he painted in 1551 (now lost). When John Frederick the Magnanimous was finally able to return to his diminished country in 1552, he established his court in Weimar, where Cranach, at 80, was given a new contract as court painter. He died in Weimar on October 16, 1553, at the age of 81.

Assessment. Cranach did not acknowledge his works with his name. The early ones, before 1504, were unsigned; from 1504 to 1506 his signature consisted of an entwined "LC"; from 1506 to 1509, it consisted of the separated initials "LC"; from 1509 to 1514, it consisted of these spaced initials and his coat of arms, the winged serpent, which became his sole signature in 1515. All works, even those that had issued from his large workshop or studio (in which he often employed ten or more assistants) and had only been checked out by him, henceforth carried this device, which was also used by his son Lucas the Younger, until his death in 1586. This gave rise to many problems of attribution that still remain unsolved. The fact that so few works bear any date further complicates the establishment of a Cranach chronology. The works of the 30-year-old artist were paintings of a profoundly devotional kind set in the emotion-laden landscapes of the Alpine foothills, with ruins and windswept trees. Later, too, he remained as it were a teller of fairytales, even when depicting scenes from classical mythology, sensuous nudes, and very slender figures in the proportions of Mannerism. Along with such paintings as

Painterly style

"Reclining River Nymph at the Fountain," "Adam and Eve," and "Hercules and Antaeus," he painted Biblical beauties, such as Salome or Judith, and such classical themes as Lucretia and the judgment of Paris, subjects to which he returned time and again. He represented female saints as beautiful and elegant ladies in fashionable dress and covered with jewelry. In his works, the borderline between sacred and mundane art is blurred.

Drawing played a minor role in the production of this inventive artist. Only a handful of his drawings are autonomous works, "master drawings" like the eight border drawings of 1515 for the emperor Maximilian's prayer book. The others are sketches for murals, altarpieces, portraits, or preliminary outlines of new projects.

Pictor celerrimus (swiftest of painters) is what Cranach is called on his tombstone, and his contemporaries never ceased to marvel at the speed with which he worked. But this very speed also suggested the limitations of his art, for his strength lay not in reflection, composition, and construction but in an impulsive creativity that was nourished by his imagination and fancy, particularly in unheroic and idyllic scenes. His art was especially popular in that period of great political upheavals, and for good reason: his success may well be due to the fact that his contemporaries, in public life the protagonists of embattled ideologies, yearned for beauty in man and in nature and for a peaceful refuge from the world's turmoil.

No other German painter was given the honour that came to Cranach after his death, when Lucas Cranach the Younger finished in 1555 a large altar in triptych form for the parish church of Wittenberg. On the side panels, the ducal family is shown kneeling; in the centre (an allegory on redemption through Christ's expiatory death), John the Baptist stands below the cross, and between John the Evangelist and the reformer Martin Luther, life-size, the painter Lucas Cranach the Elder, onto whose head Christ's redeeming blood flows from the wound. It is a memorial by which the reigning family immortalized not only itself but also its reformer, Luther, and its confidant Cranach, the court painter who had served it faithfully for almost 50 years.

MAJOR WORKS

"St. Francis Receiving the Stigmata" (*c.* 1502; Gemäldegalerie der Akademie der Bildenden Künste, Vienna); "St. Jerome in Penitence" (1502; Kunsthistorisches Museum, Vienna); "Christ on the Cross" (1503; Alte Pinakothek, Munich); "Rest on the Flight into Egypt" (1504; Staatliche Museen Preussischer Kulturbesitz, Berlin); "St. Catherine Altarpiece," right wing (1506; Gemäldegalerie, Dresden); "The Holy Kinship" (*c.* 1510–12; Gemäldegalerie der Akademie der Bildenden Künste, Vienna); "Duke Henry the Pious" and "Duchess Katharina von Mecklenburg" (1514; Gemäldegalerie, Dresden); "Virgin and Child with St. Anne" (*c.* 1515–20; Alte Pinakothek, Munich); "The Nativity" (*c.* 1515–20; Gemäldegalerie, Dresden); "A Prince of Saxony" and "A Princess of Saxony" (*c.* 1516–18; National Gallery of Art, Washington, D.C.); "Reclining River Nymph at the Fountain" (1518; Museum der Bildenden Kunste, Leipzig); "Martin Luther As St. George" (*c.* 1521; Kunstsammlungen, Weimar); "David and Bathsheba" (1526; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Cardinal Albrecht of Brandenburg as St. Jerome" (1526; John and Mable Ringling Museum of Art, Sarasota, Florida); "Adam and Eve" (1526; Courtauld Institute Galleries, London); "Portrait of Dr. J. Scheyring" (1529; Musées Royaux des Beaux Arts, Brussels); "The Stag Hunt of the Elector Frederick the Wise" (1529; Kunsthistorisches Museum, Vienna); "Paradise" (1530; Kunsthistorisches Museum, Vienna); "The Close of the Silver Age (?)" (*c.* 1530; National Gallery, London); "Apollo and Diana in a Wooded Landscape" (1530; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Venus and Cupid" (*c.* 1530; Staatliche Museen Preussischer Kulturbesitz, Berlin); "The Judgment of Paris" (1530; Kunsthalle, Karlsruhe); "Hercules and Antaeus" (*c.* 1530; private collection, Garmisch-Partenkirchen, Germany); "Feast of Herod" (1531; Wadsworth Atheneum, Hartford, Connecticut); "The Old Lover" (1531; Gemäldegalerie der Akademie der Bildenden Künste, Vienna); "The Payment" (1532; Nationalmuseum, Stockholm); "Venus" (1532; Städelsches Kunstinstitut, Frankfurt); "Self-Portrait" (1550; Uffizi, Florence).

BIBLIOGRAPHY

Biography: CHRISTIAN SCHUCHARDT, *Lucas Cranachs des Aelteren Leben und Werke*, 3 vol. (1851–70); KURT GLASER, *Lukas Cranach*, 2nd ed. (1923), with 117 illustrations; H. LUDECKE, *Lukas Cranach der Ältere: Der Künstler und seine Zeit* (1953), a Marxist interpretation, with 154 illustrations; FRIEDRICH THONE, *Lucas Cranach der Aeltere* (1965); DIETER KOEPPLIN, "Lucas Cranachs Heirat und das Geburtsjahr des Sohnes Hans," *Zeitschrift des Deutschen Vereins für Kunst wissenschaft*, 20:79–84 (1966), important because it correctly dates Cranach's marriage *c.* 1512 and the birth of his son not before 1513.

Paintings and drawings: MAX J. FRIEDLANDER and JAKOB ROSENBERG, *Die Gemälde von Lukas Cranach* (1932), a catalog with 368 illustrations; CHARLES L. KUHN, *A Catalogue of German Paintings of the Middle Ages and Renaissance in American Collections* (1936); THEODOR L. GIRSHAUSEN, *Die Handzeichnungen Lukas Cranachs des Ältern* (1937); JAKOB ROSENBERG, "The Problem of Authenticity in Cranach's Late Period," *Art Quarterly*, 18:165–168 (1955); and *Die Zeichnungen Lukas Cranachs* (1960).

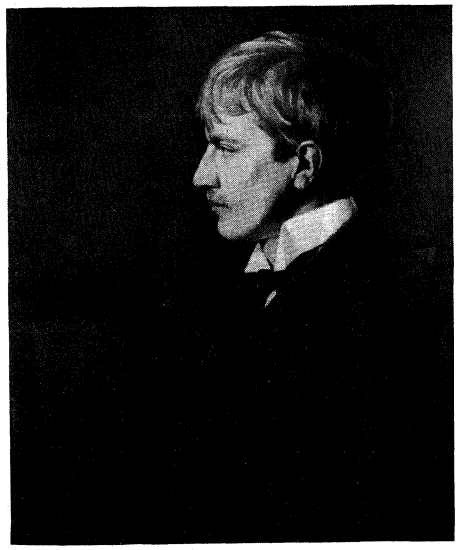
Graphic arts and illustrations: EDUARD FLECHSIG, *Cranachstudien* (1900), an important contribution on the woodcuts of the Cranach circle; W. SCHEIDIG, *Katalog der Lukas-Cranach-Ausstellung Weimar und Wittenberg* (1953), contains Cranach's woodcuts and copperplate engravings as well as woodcut book illustrations of the Cranach circle; JOHANNES JAHN, *Lucas Cranach als Graphiker* (1955), an evaluation of Cranach's illustrations to about 1520; FRIEDRICH THONE, *Wolfenbüttel*, 2nd ed. (1968), a discussion of Cranach the Elder's drawing of the model for the woodcut view of Wolfenbüttel and for the two coats of arms.

(F.Th.)

Crane, Stephen

Although Stephen Crane died at the age of 28, he left enough outstanding fiction to place him solidly among the half-dozen best U.S. novelists and short-story writers of the 19th century. He first broke new ground in *Maggie: A Girl of the Streets* (1893), which evinced an uncompromising (then considered sordid) realism that initiated the literary trend of the next generation—*i.e.*, the sociological novels of Frank Norris, Theodore Dreiser, and James T. Farrell. In the social realism of "The Monster" (1898), Crane anticipated Ralph Ellison's *Invisible Man* (1952) and other writers on the plight of the Negro in society. Crane intended *The Red Badge of Courage* to be "a psychological portrayal of fear," and reviewers rightly praised its psychological realism. The first nonromantic novel of the Civil War to attain widespread popularity, *The Red Badge of Courage* turned the tide of the prevailing convention about war fiction and established a new, if not unprecedented, one. The secret of Crane's success as war correspondent, journalist, novelist, short-story writer, and poet lay in his double point of view, by which he achieved tensions between irony and pity, illusion and reality, or the double mood of hope contradicted by despair. A great stylist, Crane was a master of the contradictory effect.

By courtesy of University of Virginia Library, Barrett Library of American Literature



Crane, painting by C.K. Linson, 1896.

Methodist heritage and education

Born on November 1, 1871, in the parsonage of the Central Methodist Church in Newark, New Jersey, Stephen was the 14th child of Mary Helen Peck and the Reverend Dr. Jonathan Townley Crane. The year they were married Dr. Crane published his *Essay on Dancing* (1848)—he was against it. "Upon my mother's side," Stephen once remarked, "everybody as soon as he could walk became a Methodist clergyman—of the ambling-nag, saddle-bag, exhorting kind." Mrs. Crane, the daughter of the Reverend George Peck, an eloquent minister and editor of the Methodist *Christian Advocate*, was the niece of Bishop Jesse Truesdell Peck, one of the founders of Syracuse University and author of *What Must I Do To Be Saved?* (1858). A woman of broad culture, although not university educated, Mrs. Crane was active in temperance programs, served as vice president of the Women's Christian Temperance Union in New Jersey, and reported Methodist revivalist meetings in the *New York Tribune*.

Stephen's ancestors included not only clergymen but also soldiers, including the four sons of the Stephen Crane who served in the Continental Congress (1774–89). As one of Stephen's friends declared in 1896, "It is an interesting study in heredity to note the influence of these two professions in Mr. Crane's literary work, the one furnishing the basis of style, the other of incident."

Stephen grew up in full rebellion against the dogmas of his clerical forebears and came to revel in all the vices his father had preached against in *Popular Amusements* (1869) and *The Arts of Intoxication* (1870)—the vices of the novel, the theatre, baseball, tobacco, and alcohol. Stephen not only read novels but also wrote them. Admonished by his father to "Read your Bibles," Stephen was intimately versed in the Bible, and his writings are haunted by his religious background.

Stephen attended school in Asbury Park, New Jersey, where his brother Jonathan Townley operated a news agency from 1882 to 1892, when the *New York Tribune* fired him because of an article Stephen wrote for his column, "On the New Jersey Coast." The Junior Order of United American Mechanics took offense at Crane's description of their Asbury Park parade. From September 1885 to December 1887, Stephen attended the Pennington Seminary, of which his father had been president from 1849 to 1858. He next attended Claverack College (from January 1888 to June 1890), a semi-military and coeducational school, where he was a strict disciplinarian as drillmaster, though he pretended not to like drilling. The school's leading tenor, he pretended not to like singing. Gambling, drinking, smoking, he had "a bully time," the happiest in his life; the faculty thought he would come to a bad end. Yet he was a voracious reader of all 19th-century English writers, Shakespeare, Plutarch's *Lives of Noble Grecians and Romans*, and the classics of Greece and Rome.

Military drill at Claverack accelerated his interest in warfare, as did the Civil War reminiscences of his history teacher, the Rev. Gen. John Bullock Van Petten, who as true Christian and brave soldier personified the two strands of Crane's heredity. They are fused again in Jim Conklin of *The Red Badge of Courage*. Although Stephen took pride in his military ancestors, he was "bitterly ashamed" to be named after the biblical martyr.

Crane next attended Lafayette College (fall 1890). Flunking out, he then survived one semester (spring 1891) at Syracuse University, where again he spent more time on the baseball diamond than in the classroom. Although he was acclaimed by his teammates as the best catcher and infield man of his time, he decided against being a professional baseball player in favour of a literary career. While at Syracuse he wrote his first draft of *Maggie* and "Great Bugs in Onondaga," his first but not his last literary hoax.

To write *Maggie* Crane explored the Bowery (where he claimed to have received his "education"), bumming beds in studios of artist friends, which he described in his fourth novel, *The Third Violet*. These friends acquainted him with the French Impressionists and Neo-Impressionists, which may account for the prose pointillism of his *Red Badge of Courage*, which is composed of seemingly disconnected images that coalesce like the blobs of colour in French Impressionist paintings. Strikingly opposite in style is the sober realism of *George's Mother*, a study of fear in the squalor of the Bowery. He began this novel after publishing *Maggie* and wrote it (simultaneously with *The Red Badge of Courage*) on the back sides of sheets of "The Holler Tree" in May–June 1893. The *Tribune* in 1892 had published a dozen of his Sullivan County tales and sketches. The fact that these writings were all composed within a span of two or three years (1892–94) and that the styles differed from one another prompted the influential critic William Dean Howells to remark, "Here is a writer who has sprung to life fully armed."

Recognition as a writer

Crane published *Maggie* at his own expense by selling his brother William shares of coal-mine stock inherited on the death of their mother in 1891; the unknown printers did him in for $869. Rejected by editors as "not nice," Crane issued *Maggie* under the pseudonym "Johnston Smith." Reform ministers ignored the novel, and it did not sell; but *Maggie* did win Crane the friendship of Howells, and Hamlin Garland, who was already forceful in literary affairs, favourably reviewed it. He urged Crane to explore the demimonde of the city, which Crane did in the New York *Press* sketches and stories from 1894. Curiosity was Crane's ruling passion.

Early in 1895 Crane journeyed to the West and Mexico, writing sketches for the Bacheller and Johnson Syndicate. His best western stories, however, were to be written in England. Late in 1895 Irving Bacheller, who had syndicated *The Red Badge of Courage* in a shortened newspaper version in 1894, sent Crane to Jacksonville, Florida, with $700 in gold to report the insurrection in Cuba. But, on New Years' Day 1897, the filibustering "Commodore" sank with $5,000 worth of ammunition, and Crane —reported drowned—finally rowed into shore in a dinghy with the captain, cook, and oiler, Crane scuttling his money belt of gold before swimming through dangerous surf. The result was one of the world's great short stories, "The Open Boat."

War correspondent in Cuba and Greece

Unable to get to Cuba, Crane went to Greece to report the Greco-Turkish War for the New York *Journal*, accompanied by Cora Taylor, the madame of the Hotel de Dream in Jacksonville who had fallen in love with the famous author and had sold her "house of joy" to become the "first" woman war correspondent. At the end of the war, they settled in England in a villa at Oxted, Surrey, and in April 1898 Crane departed to report the Spanish–American War in Cuba, first for the New York *World* and then for the New York *Journal*. When the war ended, Crane hid out in Havana, writing the first draft of *Active Service*, a novel of the Greek war. He finally returned to Cora and England nine months after his departure and settled in an almost uninhabitable and costly to maintain 14th-century manor house at Brede Place, Sussex. Here Cora, a silly woman with social and literary pretensions, contributed to Crane's ruin by encouraging his own social ambitions. Financially they bled themselves to death by entertaining hordes of spongers, as well as close literary friends—including Joseph Conrad, Ford Madox Ford, H.G. Wells, Henry James, and Robert Barr, who completed Crane's Irish romance *The O'Ruddy*.

Friendship with Conrad, James, and Wells

Crane lived in desperation against time and debts (they totalled $5,000 at his death). His indifference to money coupled with his constant need of it patterned his life and also Cora's. On borrowed money, Cora deposited the dying Crane in a sanitarium at Badenweiler, Germany, where he died on June 5, 1900, of tuberculosis compounded by the recurrent malaria fever he had caught in Cuba. Crane's body, after lying in a London horse stall, was brought back to New York City for services and then internment at Hillside, New Jersey.

MAJOR WORKS

NOVELS: *Maggie: A Girl of the Streets* (privately printed 1893, published 1896); *The Red Badge of Courage* (1895); *George's Mother* (1896); *The Third Violet* (1897); *Active Service* (1899); *The O'Ruddy* (1903).

SHORT STORIES: *The Little Regiment, and Other Episodes of the American Civil War* (1896); *The Open Boat, and Other Tales of Adventure* (1898); *The Monster and Other Stories* (1899); *Whilomville Stories* (1900); *Wounds in the Rain: A Collection of Stories Relating to the Spanish-American War of 1898* (1900); *Last Words* (1902).
POETRY: *The Black Riders and Other Lines* (1895); *War Is Kind* (1899).

BIBLIOGRAPHY. Three biographical works, all entitled *Stephen Crane*, are those by R.W. STALLMAN (1968), which locates Crane in his historical background; THOMAS BEER (1923); and JOHN BERRYMAN (1950), similar in style and theme to Beer. LILLIAN GILKES, *Cora Crane* (1960), defends Cora by denigrating Crane. R.W. STALLMAN, *Stephen Crane: A Critical Bibliography* (1972), incorporates and extensively enlarges the Crane bibliography by AMES W. WILLIAMS and VINCENT STARRETT (1948), and reports contemporary reviews in quotation and critically annotates more than 2,000 writings on Crane. See also THEODORE L. GROSS and STANLEY WERTHEIM, *Hawthorne, Melville, Stephen Crane: A Critical Bibliography*, pp. 210–295 (1971).

The definitive texts are collected in *The Works of Stephen Crane*, ed. by FREDSON BOWERS (1969–). Comprised of tampered texts, *The Works of Stephen Crane*, ed. by WILSON FOLLETT, 12 vol. (1925–27), remains useful solely now for its introductions, notably by HERGESHEIMER, AMY LOWELL, and WILLA CATHER. *Stephen Crane: Letters*, ed. by R.W. STALLMAN and LILLIAN GILKES (1960), contains 184 new letters by Crane, Cora, and their friends. *Uncollected Writings*, ed. by OLOV W. FRYCHKSTEDT (1963), does not include new Crane manuscripts; *The War Dispatches* (1964), and *The New York City Sketches* (1966), ed. by R.W. STALLMAN and E.R. HAGEMANN, do contain some new ones. *The Collected Poems*, ed. by JOSEPH KATZ (1966), supersedes Follett's 1930 edition. *The Complete Short Stories and Sketches* (1963), and *The Complete Novels* (1967), were edited by THOMAS A. GULLASON; the former book lacks 40 pieces and thus is not complete, nor are its texts the first periodical publication as claimed. *Sullivan County Tales and Sketches*, ed. by R.W. STALLMAN (1968), adds seven new pieces to MELVIN SCHOBERLIN, *Sullivan County Sketches* (1949). Stallman's *Stephen Crane: An Omnibus* (1952), presented the first collection of Crane letters, brought together for the first time the manuscripts of *The Red Badge of Courage*, and critically scrutinized Crane's best works.

Among the best short scrutinies ever written on Crane are the essays by JAMES B. COLVERT and SERGIO PEROSA in *A Collection of Critical Essays*, ed. by MAURICE BASSAN (1967); and Daniel Knapp, "Son of Thunder," in *Nineteenth Century Fiction*, 24:253–291 (1969). A landmark in Crane criticism was the Crane number of *Modern Fiction Studies*, vol. 5 (1959). Other collections of criticism include: SCULLEY BRADLEY (ed.), *The Red Badge of Courage: Text and Essays in Criticism* (1962); M. BASSAN (ed.), *Maggie: Text and Context* (1966); J. KATZ (ed.), *The Blue Hotel* (1969); STANLEY WERTHEIM (ed.), *Studies in Maggie and George's Mother* (1970); and R.W. STALLMAN (ed.), *The Art of Stephen Crane: A Critical Symposium* (1974).

An excellent study of the poetry is DANIEL G. HOFFMAN, *The Poetry of Stephen Crane* (1957). Other studies of Crane and his works are EDWIN CADY, *Stephen Crane* (1962); and JEAN CAZEMAJOU, *Écrivain journaliste* (1969). His pamphlet *Stephen Crane* (1969), is an excellent short introduction. ERIC SOLOMAN, *Stephen Crane in England* (1964), refashions Crane's literary relationships, which were outlined in *Letters* and narrated in *Cora Crane* (1960); and which obtain rather exhaustive portraits in *Stephen Crane: A Biography* (1968). Solomon's *From Parody to Realism* (1966), reduces Crane's works to parodies of popular romantic literature. DONALD B. GIBSON, *The Fiction of Stephen Crane* (1968), is at its best on "The Blue Hotel." The best short studies of Crane's poetry are HARLAND S. NELSON, "Stephen Crane's Achievement As a Poet," *Texas Studies in Literature and Language*, 4: 564–582 (1963); MAX WESTBROOK, "Stephen Crane's Poetry: Perspective and Arrogance," *Bucknell Review*, 11:24–34 (1963); and RUTH MILLER, "Regions of Snow: The Poetic Style of Stephen Crane," *Bulletin of the New York Public Library*, 72:328–349 (1968).

(Ro.W.S.)

Cranmer, Thomas

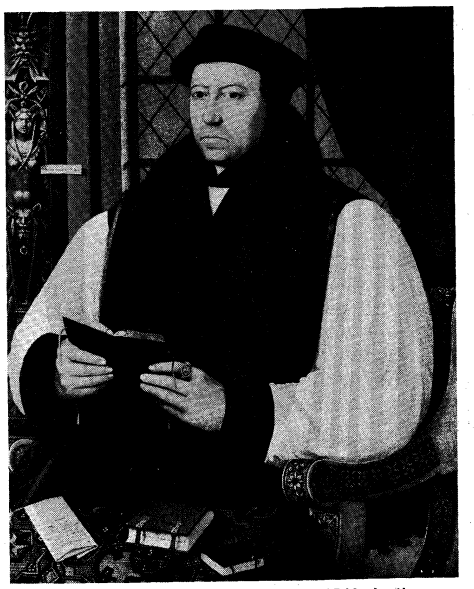
Thomas Cranmer was the English prelate who found in canon law a rationale that enabled King Henry VIII to justify invalidating his marriage to Catherine of Aragon without recourse to the Pope and who subsequently was made the first archbishop of Canterbury of the reformed Church of England. As archbishop, he put the English Bible in parish churches, drew up the Book of Common Prayer, and composed a litany that remains in use today. His belief in the divine right of kings to rule the clergy as well as the people and his emphasis on the Bible made him the characteristic Anglican of his day. Denounced for promoting Protestantism by the Catholic Mary I, he was tried for treason, convicted of heresy, and burned at the stake.

Born at Aslacton, Nottinghamshire, on July 2, 1489, he was the second son of Thomas Cranmer and Agnes (née Hatfield). His father seems to have belonged to the lowest rank of the gentry; at any rate, he had only enough property to endow his eldest son, John, so that Thomas and his younger brother, Edmund, were destined for the church. After experiencing the teaching of a "marvellous severe and cruel schoolmaster," whose ministrations Cranmer later maintained instilled in him a permanent uncertainty and pliability, the boy went on to Cambridge in 1503. In 1510 or 1511 he was elected to a fellowship at Jesus College but was soon compelled to vacate because he married a relative of the landlady of the Dolphin Inn. During this time he earned his living by teaching at Buckingham (later Magdalene) College, leaving his wife to lodge at the Dolphin; out of this arrangement grew a later story that he had started out in life as a hostler.

His wife died in childbirth soon after their marriage, however, and Jesus College restored Cranmer to his fellowship. He now entered the church and threw himself into his studies, becoming one of the outstanding theologians of his time, a man of immense, though not very original, learning. From about 1520 he belonged to a group of scholars who met regularly at the White Horse Inn to discuss the theological and ecclesiological problems raised by Luther's revolt; known to be inclined to the new way of thinking, they were dubbed "Little Germany." Among the group that was to lead the first generation of the English Reformation were William Tyndale, Robert Barnes, Thomas Bilney, and, above all, Cranmer, who by 1525 included among his prayers one for the abolition of papal power in England.

Entry into royal service. Cranmer's ambitions for reform would have remained academic had it not been for the political events into which he was soon drawn, however contrary they were to his upbringing and tastes. From 1527 onward, Henry VIII pursued his desire to be freed from his first wife, Catherine of Aragon, in order to marry Anne Boleyn, and in 1529 the wheels of the "divorce" seized also upon Cranmer. In August a plague known as the sweating sickness swept the country and was especially severe in Cambridge; to escape the sickness, Cranmer left the town with two of his pupils, named Cressy and related to him through their mother, and went to their father's house at Waltham in Essex. The King was visiting in the immediate neighbourhood at the time, and two of his chief councillors, Stephen Gardiner and Edward Fox, soon afterwards met Cranmer in those lodgings. Not surprisingly, they were led to discuss the King's meditated divorce. The old idea that Cranmer suggested consulting the canonists and the universities about the legality of a marriage with a deceased brother's widow was mistaken, but he evidently demonstrated his belief that the King had the right of it and advocated reliance on the services of home-bred theologians. Henry, who was willing to secure the help of any likely head and hand, however obscure, summoned Cranmer for an interview and commanded him to lay aside all other pursuits in order to devote himself to the question of the divorce. Cranmer accepted a commission to write a propaganda treatise in the King's interest, stating the course he proposed and defending it by arguments from Scripture, the Fathers, and the decrees of general councils. His material interests certainly did not suffer by compliance. He was commended to the hospitality of Anne Boleyn's father, the Earl of Wiltshire, in whose house at Durham Place he resided for some time; was appointed archdeacon of Taunton; became one of the King's chaplains; and also held a parochial benefice, the name of which is unknown. When the treatise was finished, Cranmer was called upon to

The divorce question



Cranmer, oil painting by G. Fliccius, 1546. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

defend its argument before the universities of Oxford and Cambridge; but in the end the debates, which on the whole endorsed his position, took place in his absence. He had already been sent to plead the cause before a more powerful if not a higher tribunal. An embassy, with the Earl of Wiltshire at its head, was dispatched to Rome in 1530, and Cranmer was an important member of it. He was received by the Pope with marked courtesy and was appointed grand penitentiary of England, but his argument, if he ever had the opportunity of stating it, did not lead to any practical decision of the divorce question.

In 1532 he was sent to Germany, officially as ambassador to the Emperor Charles V, but with instructions to establish contact with the Lutheran princes. At Nürnberg he made the acquaintance of Andreas Osiander, whose theological position midway between Luther and the old orthodoxy appealed to Cranmer's cautious temperament, while Osiander's niece Margaret appealed even more strongly to the instincts of one who had for too long remained in uncongenial celibacy. Despite his priest's orders he married her in 1532; at the same time, his theological views underwent a further decided change in the direction of reformed opinion.

Archbishop of Canterbury. The year 1532 proved to be a critical one altogether, for in August William Warham, the aged archbishop of Canterbury, died. At first the usual practice of extending the vacancy for the benefit of the King's finances was followed, but by the end of the year it was apparent that the see would have to be filled because the divorce question was coming to a head. Thomas Cromwell's arrival in power as chief adviser in ecclesiastical matters had heralded a more energetic policy, and by January 1533 the act against appeals to Rome was being drafted and Anne Boleyn was pregnant. Since Stephen Gardiner, the obvious candidate for the archbishopric, was out of favour, the King chose Cranmer; by March 1533 he was consecrated and instituted at Canterbury, with the assistance of confirmatory papal bulls and after a declaration that he took the obligatory oath to the Pope without feeling bound by it. He proceeded to do what was expected of him. In May he convened his court at Dunstable, declared the King's marriage to Catherine of Aragon void from the beginning and pronounced the second marriage to Anne Boleyn valid.

In 1536, convinced by the dubious evidence of Anne's alleged adulteries, he in turn invalidated that marriage; in 1540 he assisted in the freeing of Henry VIII from his fourth wife, Anne of Cleves; and in 1542 he was forced to be prominent in the proceedings that resulted in Catherine Howard's execution for treasonable unchastity. There is no question that in these matrimonial politics he did as he was told, though it is entirely improbable that his private opinions on the issues in question in any way contradicted his public doings.

More significant are his activities as archbishop in the reconstructed church. Cranmer had not sought high promotion. His marriage just before his elevation to the archbishopric is fair proof that he expected no such career in the priesthood, in which a necessarily unacknowledged wife would be nothing but an embarrassment. Not until 1548 was he able publicly to recognize her. A story of his carrying her about with him in a chest with air holes is, however, part of the scurrilous legend that grew up around him. Once put in power, however, he could not avoid the consequences; a convinced reformer with leanings toward a succession of continental theological changes, he found himself assisting at the shaping of the Church of England under a master who on the whole had no taste for change. In cooperation with Cromwell he promoted the publication of an English Bible, made compulsory in the parishes by Cromwell's Injunctions of 1538. Even before Henry VIII died (1547), Cranmer had drifted far in the direction of Protestantism; in 1545 he had composed a litany for the reformed Church of England, one of his masterpieces, still in use; and by 1538 had abandoned the traditional Roman Catholic belief in transubstantiation—that Christ is rendered substantially present by the Eucharist (although the properties of bread and wine remain the same)—but retained his belief in the real presence of Christ in the Eucharist. As early as 1536 he was recognized by the northern religious rebels as the leading innovator. His position was in consequence far from comfortable after the Act of Six Articles (1539), which attacked those advocating marriage of the clergy and those denying transubstantiation, and Cromwell's fall in 1540. During Henry's last years, Cranmer's enemies laid at least three elaborate plots to destroy him by convicting him of heresy, but on each occasion they were foiled by Henry's curious attachment to him. In Cranmer, this king, who as a rule kept himself entirely free from personal feelings for his servants and advisers, found a man whom he both trusted and liked. Unlike the rest of them, the Archbishop was neither greedy nor devious; he sought nothing for himself, alone was willing to plead for those who fell into disfavour (a service he performed with equal courage and futility for Sir Thomas More, Anne Boleyn, Thomas Cromwell, and others), and miraculously retained Henry's goodwill throughout. The King regarded him with that mixture of awe and amusement that the worldly and selfish bestow on those who appear simple in affairs; he liked him, listened to him, protected him, but allowed him no political influence whatsoever. It was not surprising that he turned to Cranmer when death came.

Achievements as archbishop

Achievements under Edward VI. With the accession of Edward VI (Henry's only child by his third wife, Jane Seymour) in 1547, Cranmer's time really arrived. From the first, the young king's guardian, Edward Seymour, duke of Somerset, demonstrated his intention to transform the Church of England into a Protestant church. When he fell in 1549 the expected Catholic reaction did not take place because John Dudley (later the duke of Northumberland), who had ousted Seymour, decided to introduce an even more extreme brand of reformed religion. In the doctrinal labours demanded by these changes, Cranmer took the chief and directing part. In 1547 he was responsible for the publication of a *Book of Homilies* designed to meet the notorious grievance that the unreformed clergy did not preach enough. The first Prayer Book, moderately Protestant, appeared in 1549, to be followed in 1552 by the second, which was more outspokenly Protestant. Cranmer was personally responsible for much of the work, but he had the assistance of a number of foreign theologians for whom Edward VI's England acted as a magnet. The most influential of these was probably Martin Bucer from Strassburg, whose position on the Eucharist is reflected especially in the communion service of the second Prayer Book. It was not so

much Bucer, however, who persuaded Cranmer away from the vague Lutheranism, which seems to have been his position in 1547, as either the Pole Jan Łaski the Younger or the Englishman Nicholas Ridley, both men possessed of a more determined and unquestioning temper than was the Archbishop. The ferment of those years also produced Cranmer's 42 Articles (1553), a set of doctrinal formulas defining the dogmatic position of the Church of England on current religious controversies. All clergy, schoolmasters, and degree candidates in the universities were compelled to subscribe to the articles, which were later reduced to 39 and officially accepted by the Anglican church. At this time, Cranmer also attempted to revise the canon law of the English Church, a proposal never enacted but published in 1571 as the *Reformatio Legum Ecclesiasticarum*. Though still deprived of any serious influence in affairs of state, Cranmer dominated and guided the religious revolution of the reign by his learning, authority, and unwavering diligence. He settled in turn the doctrine, ritual, and law of his church in a manner that was to remain fundamental to later developments; above all, the Church of England owed to him the beauty of its liturgy, which shows him to have been not only a theologian but something of a poet.

The 42 Articles

Degradation and martyrdom. Edward VI's approaching death (July 1553) at long last involved Cranmer fatally in politics. After prolonged resistance he was forced by the dying King to subscribe the document by which Northumberland hoped to upset custom, statute law, and the will of Henry VIII in order to transfer the succession from the princess Mary (Henry's daughter by Catherine of Aragon) to his daughter-in-law, the great-niece of Henry, Lady Jane Grey. Although proclaimed queen, she was deposed and beheaded nine days later, and Mary I acceded to the throne. The failure of the plot brought charges of treason against Cranmer, and he was condemned by Mary's government in November 1553. It had in any case become obvious before this that his future held no more bright promises. Mary's accession temporarily destroyed the English Reformation; Cranmer's embittered enemy Stephen Gardiner was at once released from imprisonment and promoted to the chancellorship, and in November 1554 Cardinal Reginald Pole arrived to occupy Canterbury and direct the extirpation of heresy. Cranmer's trial for treason was but a pretext; the Queen and her advisers did not intend him to die for the technical offense of having supported Northumberland's insane conspiracy but meant to destroy him for his long-standing offense in promoting Protestantism. They had to wait until they could get Parliament to repeal the acts of Henry VIII and Edward VI and to reintroduce the laws that enabled the secular arm to burn heretics. With Ridley and Hugh Latimer, a Protestant who had formerly been bishop of Worcester, Cranmer in March 1554 was removed to Oxford where the Counter-Reformation felt safer than in Cranmer's own university. Late in that year the heresy laws were revived, and in September 1555, after enfeebling imprisonment, Cranmer was subjected to a long trial in which he stoutly defended himself against the charge of having unjustifiably departed from his own earlier position on the sacraments and the papacy. The foregone conclusion was arrived at after a variety of technical processes; on February 14, 1556, in a ceremony full of carefully designed humiliation, he was degraded from his episcopal and sacerdotal offices and handed over to the state.

Trial for heresy

But Mary's government had not done with him yet. The burning of the arch-heretic would be an even more useful deed if he could be made to renounce his errors in public, and so a number of ways were tried to break him down. The previous October he had been forced to witness the martyrdom of Ridley and Latimer; now he was temporarily removed from prison into more pleasant surroundings while government agents tried to stir up his doubts. In fact, Cranmer signed five so-called recantations of which the first four did no more than record his consistent belief that what monarch and Parliament had decreed must be obeyed by all Englishmen. His convictions on this point logically forced him to accept the Marian Counter-Reformation as valid, and this acceptance, in turn, in his weak and uncertain state, not unaffected by the delay of death and the faint hope of mercy, finally induced him to make an abject recantation (the sixth) of his whole religious development. The government had every reason to hope that the publication of Cranmer's defection would wreck Protestantism in England. No mercy was to be extended to the old man. The vengeful Gardiner had died, but Queen Mary and Cardinal Pole were quite determined that the sentence must be carried out. Thus on March 21, 1556, Cranmer was taken out to be burned, being first required to make his recantation public. The proximity of death, however, restored both his faith and his dignity. With nothing to lose and only peace of soul to gain, he shocked his enemies by disavowing his recantation and emphatically reasserting that the Pope's power was usurped and transubstantiation untrue. At one blow Cranmer undid all that government propaganda had achieved and restored heart to the surviving reformers. Then he went to his death. As he had promised, he steadfastly held his right hand—which "had offended" by signing the false recantations—into the flame until it was consumed, and soon afterward the fire killed him. His brave and dignified end made an enormous impression.

Recantation and disavowal

Assessment. Cranmer was a very human man who in consequence has attracted a good deal of obloquy from those who have not had to share his tribulations and temptations. Essentially he was and remained a scholar who lacked the strength that singlemindedness and fanaticism instill into the less reflective. He has sometimes been thought of as infirm in moral purpose, but this is to misjudge him. His doubts at the last were cleverly induced by mental torture, and his gradual development away from traditional orthodoxy into more and more definitely Protestant views during the Reformation represents fairly the spiritual career of a man who obeyed reason rather than instinct. Cranmer was always learning and was never ashamed to admit it; his was essentially a humble temper. He had not sought high office and did not particularly enjoy it, though he valued his place for the chance it gave him to promote the changes that he came to regard as essential to the establishment of God's truth. He refused to bear malice or to punish those who traduced him. When Cromwell once told him, in some exasperation, that the "popish knaves" would have his eyes and cut his throat before he would do something about it, Cranmer turned the prophecy with a shrug. In a persecuting age he stood out for his clemency, though in 1550 he did take part in the trial and burning of Joan Bocher. It should be remembered, however, that she was condemned for open blasphemy in denying the Trinity, the one offense that all the church had regarded as unforgivable ever since the struggle with Arianism. For his order and the authority of the church Cranmer had a high respect, which, for instance, appears in his revision of the canon law.

It was part of his religious beliefs that he owed obedience to the king; though he did not worship the state, he served it without hesitation and as a matter of principle. This position did not, as is sometimes alleged, make him servile; alone of Henry VIII's councillors, Cranmer time and again spoke up for the unpopular victim of the moment, and his tart criticism of the King's theology and grammar in the debates over the King's Book of 1543 speaks well for his courage. He stood up to Northumberland when everyone else quailed before that half-demented demon.

These occasional disputes only underline the fact that with him submission to royal authority was ordinarily a fundamental, indeed a doctrinal, tenet. Though he may have been more consistent in this than most, he only stressed more heavily what nearly everybody held at the time. His other guiding star was his study of theology, in which he early discarded the arid aftermath of late medieval Scholasticism and turned instead to Scripture and the early Fathers. His belief in the divine right of

kings to rule the church as well as the state and his biblical theology made him the characteristic Anglican of his day: the intellectual and in part the spiritual founder-father of the reformed church in England.

BIBLIOGRAPHY. The most complete biography is JASPER G. RIDLEY, *Thomas Cranmer* (1962), solid but not subtle and insufficiently sympathetic; A.F. POLLARD, *Thomas Cranmer and the English Reformation, 1489–1556* (1904), is still worth reading. On Cranmer's theology, see C.W. DUGMORE, *The Mass and the English Reformers* (1958); and P.N. BROOKS, *Thomas Cranmer's Doctrine of the Eucharist* (1965); on his place in the history of the English Church, see A.G. DICKENS, *The English Reformation*, rev. ed. (1967).

(G.R.E.)

Creation, Myths and Doctrines of

Doctrines of creation are philosophical and theological elaborations of the primal myth of creation within a religious community. The term myth here refers to the imaginative expression in narrative form of what is experienced or apprehended as basic reality. The term creation refers to the beginning of things, whether by the will and act of a transcendent being, by emanation from some ultimate source, or in any other way.

NATURE AND SIGNIFICANCE

The myth of creation is the symbolic narrative of the beginning of the world as understood by a particular community. The later doctrines of creation are interpretations of this myth in light of the subsequent history and needs of the community. Thus, for example, all theology and speculation concerning creation in the Christian community are based on the myth of creation in the biblical book of Genesis and of the new creation in Jesus Christ. Doctrines of creation are based on the myth of creation, which expresses and embodies all of the fertile possibilities for thinking about this subject within a particular religious community.

The basis for man's placement and orientation in the world

Myths are narratives that express the basic valuations of a religious community. Myths of creation refer to the process through which the world is centred and given a definite form within the whole of reality. They also serve as a basis for the orientation of man in the world. This centring and orientation specify man's place in the universe and the regard he must have for other humans, nature, and the entire nonhuman world; they set the stylistic tone that tends to determine all other gestures, actions, and structures in the culture. The cosmogonic (origin of the world) myth is the myth *par excellence*. In this sense, the myth is akin to philosophy, but, unlike philosophy, it is constituted by a system of symbols; and because it is the basis for any subsequent cultural thought, it contains rational and nonrational forms. There is an order and structure to the myth, but this order and structure is not to be confused with rational, philosophical order and structure. The myth possesses its own distinctive kind of order.

Myths of creation have another distinctive character in that they provide both the model for nonmythic expression in the culture and the model for other cultural myths. In this sense, one must distinguish between cosmogonic myths and myths of the origin of cultural techniques and artifacts. Insofar as the cosmogonic myth tells the story of the creation of the world, other myths that narrate the story of a specific technique or the discovery of a particular area of cultural life take their models from the stylistic structure of the cosmogonic myth. These latter myths may be etiological (*i.e.*, explaining origins); but the cosmogonic myth is never simply etiological, for it deals with the ultimate origin of all things.

The ritual expression of creation: sacred time

The cosmogonic myth thus has a pervasive structure; its expression in the form of philosophical and theological thought is only one dimension of its function as a model for cultural life. Though the cosmogonic myth does not necessarily lead to ritual expression, ritual is often the dramatic presentation of the myth. Such dramatization is performed to emphasize the permanence and efficacy of the central themes of the myth, which integrates and undergirds the structure of meaning and value in the culture. The ritual dramatization of the myth is the beginning of liturgy, for the religious community in its central liturgy attempts to re-create the time of the beginning.

From this ritual dramatization the notion of time is established within the religious community. To be sure, in most communities there is the notion of a sacred and a profane time. The prestige of the cosmogonic myth establishes sacred or real time. It is this time that is most efficacious for the life of the community. Dramatization of sacred time enables the community to participate in a time that has a different quality than ordinary time, which tends to be neutral. All significant temporal events are spoken of in the language of the cosmogonic myth, for only by referring them to this primordial model will they have significance.

In like manner, artistic expression in archaic or "primitive" societies, often related to ritual presentation, is modelled on the structure of the cosmogonic myth. The masks, dances, and gestures are, in one way or another, aspects of the structure of the cosmogonic myth. This meaning may also extend to the tools man uses in the making of artistic designs and to the precise technique he employs in his craft.

Mention has been made above of the fact that the cosmogonic myth situates man in a place, in space. This centring is at once symbolic and empirical: symbolic because through symbols it defines the spatiality of man in ontological terms (of being) and empirical because it orients him in a definite landscape. Indeed, the names given to the flora and fauna and to the topography are a part of the orientation of man in a space. The subsequent development of language within a human community is an extension of the language of the cosmogonic myth.

The primordial structure of culture

The initial ordering of the world through the cosmogonic myth serves as the primordial structure of culture and the articulation of the embryonic forms and styles of cultural life out of which various and differing forms of culture emerge. The recollection and celebration of the myth enable the religious community to think of and participate in the fundamentally real time, space, and mode of orientation that enables them to define their cultural life in a specific manner.

TYPES OF COSMOGONIC MYTHS

The world as a structure of meaning and value has not appeared in the same manner to all human cultures. There are, therefore, almost as many cosmogonic myths as there are human cultures. Until quite recently, classification of these myths on an evolutionary scale, from the most archaic cultures to contemporary Western cultures (*i.e.*, from the assumedly simplest to the most complex) was the most dominant mode of ordering these myths. Recent 20th-century scholars, however, tend to look at the various types of myths in terms of the structures that they reveal rather than on an evolutionary scale extending from the so-called simple to the complex, for, in a sense, there are no simple myths regarding the beginning of the world. The beginning of the world is simultaneously the beginning of the human condition, and it is impossible to speak of this beginning as if it were simple.

Creation by a supreme being. The 19th-century scholars who took an evolutionary survey of human culture and religion (*e.g.*, Sir James George Frazer and Edward Burnett Tylor) held that the notion of the creation of the world by a supreme being occurred only in the highest stage of cultural development.

Primordial supreme beings or creator gods

Andrew Lang, a Scottish folklorist, challenged this conception of the development of religious ideas, for he found in the writings of anthropologists, ethnologists, and travellers evidence of a belief in a supreme being or high god among cultures that had been classified as the most primitive. This position was taken up and elaborated by an Austrian priest-anthropologist, Wilhelm Matthäus Schmidt, who reversed the evolutionary theory, holding that there was a primordial notion of a supreme being, a kind of original intellectual and religious conception of a single creator god, that degenerated in subsequent cultural stages. Though Schmidt's theories of cultural historical stages and diffusion and an original primordial revelation have for the most part been discredited and

abandoned, the existence of a belief in a supreme being among primitive peoples (a notion discovered by Andrew Lang) has been proven and attested to over and over again by investigators of numerous cultures. This belief has been found among the cultures of Africa, the Ainu of the northern Japanese islands, Amerindians, south central Australians, the Fuegians of South America, and in almost all parts of the globe.

Though the precise nature and characteristics of the supreme creator deity may differ from culture to culture, a specific and pervasive structure of this type of deity can be discerned. The following characteristics tend to be common: (1) he is all wise and all powerful. The world comes into being because of his wisdom, and he is able to actualize the world because of his power. (2) The deity exists alone prior to the creation of the world. There is no being or thing prior to his existence. No explanation can therefore be given of his existence, before which one confronts the ultimate mystery. (3) The mode of creation is conscious, deliberate, and orderly. This again is an aspect of the creator's wisdom and power. The creation comes about because the deity seems to have a definite plan in mind and does not create on a trial-and-error basis. In Genesis, for example, particular parts of the world are created seriatim; in an Egyptian myth, Kheper, the creator deity, says, "I planned in my heart," and in a Maori myth the creator deity proceeds from inactivity to increasing stages of activity. (4) The creation of the world is simultaneously an expression of the freedom and purpose of the deity. His mode of creation defines the pattern and purpose of all aspects of the creation, though the deity is not bound by his creation. His relationship to the created order after the creation is again an aspect of his freedom. (5) In several creation myths of this type, the creator deity removes himself from the world after it has been created. After the creation the deity goes away and only appears again when a catastrophe threatens the created order. (6) The supreme creator deity is often a sky god, and the deity in this form is an instance of the religious valuation of the symbolism of the sky.

Rupture of primordial harmony and perfection

In creation myths of the above type, the creation itself or the intent of the creator deity is to create a perfect world, paradise. Before the end of the creative act or sometime soon after the end of creation, the created order or the intent of the creator deity is thwarted by some fault of one of the creatures. There is thus a rupture in the creation myth. In some myths this rupture is the cause of the departure of the deity from creation.

An African myth from the Dogon peoples of West Africa illustrates this point. In this myth the creator deity first creates an egg. Within the egg are two pairs of twins, each pair consisting of one male and one female. These twins are supposed to mature within the egg, becoming at maturation androgynous (both male and female) beings, the perfect creatures to inhabit the earth. One of the twins breaks from the egg before maturation because he wishes to dominate the creation. In so doing he carries a part of the egg with him, and from this he creates an imperfect world. The creator deity, seeing what he has done, sacrifices the other twin to establish a balance in the world. The creation is sustained by this sacrifice, and it is now ambiguous, instead of the perfect world intended by the god.

This myth not only shows how a rupture takes place within the myth itself but also points out the fact that the characteristics of the supreme creator deity noted above seldom exist apart from other mythological contexts. The widespread symbols of dualism (the divine twins), the cosmic egg, and sacrifice are basic themes in the structure of this African creation myth. In myths of this kind, however, prominence must always be given to the might of a powerful creator sky deity under whose aegis the created order comes into being.

Creation through emergence. In contrast to the creation by a supreme sky deity, there is another type of creation myth in which the creation seems to emerge through its own inner power from under the earth. In this genre of myth, the created order emerges gradually in continuous stages. It is similar to a birth or metamorphosis of the world from its embryonic state to maturity. The symbolism of the earth or a part of the earth as a repository of all potential form is prominent in this type of myth. In some myths of this type (*e.g.*, the Navajo myth of emergence), the movement from a lower stage to a higher one is initiated by some fault of the people who live under the earth, but these faults are only the parallels of an automatic upper movement in the earth itself.

From the embryonic or inchoate to the mature and definite

Just as the supreme-creator-deity myth forms a homology to the sky, the emergence myth forms a homology to the earth and to the child-bearing woman. In many cases the emergence of the created order is analogous to the growth of a child in the womb and its emission at birth. This symbolism is made clear in a Zuni myth that states,

> Anon is the nethermost world, the seed of men and creatures took form and increased; even as in eggs in warm places speedily appear . . . Everywhere were unfinished creatures, crawling like reptiles one over another, one spitting on another or doing other indecencies . . . until many among them escaped, growing wiser and more manlike.

The underworlds prior to the created order appear chaotic; the beings inhabiting these places seem without form or stability, or they commit immoral acts. The seeming chaos is moving toward a definite form of order, however, an order latent in the very forms themselves rather than from an imposition of order from the outside.

From another perspective the emergence myth is homologous to the seed. When the homologue of the seed is referred to, the meaning of fertility and death are at once introduced. The seed must die before it can be reborn and actualize its potentiality. This symbolism is dramatically presented in a wide range of funerary rites: one is buried in the earth in hope of a renewal from the earth, or the earth is the repository of the ancestors from whom the new generation emerges. In every case, emergence myths demonstrate the latent potency immanent in the earth as a repository of all life-forms.

Creation by world parents. Closely related to the above type of myth is the myth that states that the world is created as the progeny of a primordial mother and father. The mother and father are symbols of earth and sky, respectively. In myths of this kind, the world parents generally appear at a late stage of the creation process; chaos in some way exists before the coming into being of the world parents. In the Babylonian myth *Enuma elish*, it is stated,

> When on high the heaven had not been named
> Firm ground below had not been called by name,
> Naught but primordial Apsu, their begetter,
> (And) Mummu-Tiamat, she who bore them all,
> Their waters comingling as a single body;

The Maori make the same point when they state that the world parents emerge out of *po*. *Po* for the Maori means the basic matter and the method by which creation comes about. There is thus some form of reality before the appearance of the world parents.

From inertness to activity

Even though the world parents are depicted and described as in sexual embrace, no activity is taking place. They appear as quiescent and inert. The chthonic (underworld) structure of the earth as latent potentiality tends to dominate the union. The parents are often unaware that they have offspring, and thus a kind of indifference regarding the union is expressed. The union of male and female in sexual embrace is another symbol of completeness and totality. As in the African myth from the Dogon referred to above, sexual union is a sign of androgyny (being both male and female) and androgyny, in turn, a sign of perfection. The indifference of the world parents is thus not simply a sign of ignorance but equally of the silence of perfection. The world parents in the Babylonian and Maori myths do not wish to be disturbed by their offspring. As over against the parents, the offspring are signs of actuality, fragmentation, specificity; they define concrete realities.

The separation of the world parents is again a rupture within the myth. This separation is caused by offspring who wish either to have more space or to have light, for they are situated between the bodies of the parents. In

some myths the separation is caused by a woman who lifts her pestle so high in grinding grain that it strikes the sky, causing the sky to recede into the background, thus providing room for the activities of mankind. In both cases an antagonistic motive must be attributed to the agents of separation. In the Babylonian and Maori versions of this myth, actual warfare takes place as a result of the separation.

Over against the primordial union of the world parents, there is the desire for knowledge and a different orientation in space. After the separation, lesser deities related to solar symbolism take precedence in the creation. The sun and light must be seen in these myths as representing the desire for a humanizing and cultural knowledge as over against the passive and inert forms of the union of the parent deities. From the point of separation, the mythic narrative of the world-parent myths states how different forms of cultural knowledge are brought to man by the offspring, the agents of separation. The separation of the world parents is the sign of a new cosmic order, an order dedicated to the techniques, crafts, and knowledge of culture.

Creation from the cosmic egg. In the Dogon myth referred to above, the creation deity begins the act of creation by placing two embryonic sets of twins in an egg. In each set of twins is a male and female; during the maturation process they are together thus forming androgynous beings. In a Tahitian myth, the creator deity himself lives alone in a shell. After breaking out of the shell, he creates his counterpart, and together they undertake the work of creation.

A Japanese creation narrative likens the primordial chaos to an egg containing the germs of creation. In the Hindu tradition the creation of the world is symbolized in the Chandōgya *Upaniṣad* by the breaking of an egg, and the universe is referred to as an egg in other sources. The Buddhists speak of the transcending of ordinary existence, the realization of a new mode of being, as breaking the shell of the egg. Similar references to creation through the symbol of the egg are found in the Orphic texts of the Greeks and in Chinese myths.

The egg is a symbol of the totality from which all creation comes. It is like a womb containing the seeds of creation. Within the egg are the possibilities of a perfect creation (*i.e.*, the creation of androgynous beings). The egg, in addition to being the beginning of life, is equally a symbol of procreation, rebirth, and new life. In a version of the Dogon, one of the twins returns to the egg in order to resuscitate the other.

The cosmogonic water and the diver animal or diver devil

Creation by earth divers. Two elements are important in myths of this type. There is, first, the theme of the cosmogonic water representing the undifferentiated waters that are present before the earth has been created. Secondly, there is an animal who plunges into the water to secure a portion of earth. The importance of the animal is that the creature agent is a prehuman species. This version of the myth is probably the oldest version of this genre. This basic structure of the earth-diver myth has been modified in central Europe in myths that relate the story of the primordial waters, God, and the devil. In these versions of the earth-diver myth, the devil appears as God's companion in the creation of the world. The devil becomes the diver sent by God to bring earth from the bottom of the waters. In most versions of this myth, God does not appear to be omniscient or omnipotent, often depending on the knowledge of the devil for certain details regarding the creative act—details that he learns through tricks he plays upon the devil.

In still different versions of this myth, the relationship between God and the devil moves from companionship to antagonism; they become adversaries, though they remain as co-creators of the world. The fact that the devil has had a part in the creation of the world is one way of explaining the origin and persistence of evil in the world.

Mircea Eliade, a noted historian of religions, has pointed to another theme in certain Romanian versions of this myth. After God has instructed the devil to dive to the bottom of the waters and bring up the earth, the devil obeys, diving several times before he is able to bring up and hold on to a small portion of earth. After the creation of the world from this small portion of earth, God sinks into a profound sleep. This sleep is a sign of mental exhaustion, for only the devil and a bee know the solution to certain details of the creation, and God must, with the help of the bee, trick the devil into giving him this vital information. God's sleep, according to Eliade, is a sign of his passivity and disinterest in the world after it has been created, and it harks back to certain archaic myths in which the supreme deity retires from the world after its creation, becoming disinterested and passive in relationship to his work.

God's post-creation sleep

DOCTRINES OF CREATION

Some of the major types of creation myths have been presented above. It is from myths of this sort and their dominant themes that theological and philosophical speculation have been developed in the various religious communities.

Basic mythical themes. *Primordiality.* In several myths it is stated that the primordial stuff of creation was some form of undifferentiated matter (*e.g.*, water, chaos, a monster, or an egg). It is from this undifferentiated matter that the world evolves or is made. In the case of the egg and monster symbols, there seems to be a notion of a definite original form, but the egg is undifferentiated; for its form is vague and embryonic, and the monster figure —containing all of the forms of chaos in a terrible way— expresses the theme that chaos is not only passive (as is water) but resists creation. Although creation results as a modification of the primordial matter, however, it is this matter that determines and sets the limits to the extension of the world in space and time. Thus, in communities in which myths of this type find their expression, there are periods of mythical-ritual renewal at certain cyclical periods in which the world returns to its original chaos to rise again out of this initial state.

The creator deity as sole determiner of creation's nature and destiny

When it is stated that the supreme being created the world and that there was no primordial matter prior to his being, then the determination of the world is in the mind and will of the deity. This leads to distinctive conclusions regarding the destiny of the world and man. The end (and meaning) of the world is thus not determined by the primordial matter but by the deity who created the world. It is he alone who determines the preservation, maintenance, and end of the world.

Dualisms and antagonisms. In emergence myths there seems to be an easy movement from one stage of creation to the next, but, as has been shown in the Navajo myth, at each subterranean level there is some type of antagonism among the developing embryonic creatures. This is one of the reasons for the separation of the creatures and the movement to another level. Though the emergence myths portray the mildest form of this antagonism, it is still present in myths of this sort.

In the world-parent myths there is antagonism between the offspring and the parents. This is a conflict between generations, expressing the desire of the children to determine their own place and orientation in existence against the passivity of the parents.

A dualism and antagonism is found again in the cosmic-egg myths, especially in the myths in which the egg contains twins. One twin wishes to take credit for the creation of the world alone, interrupting the harmonious growth within the egg before maturation. The faulty creation by this evil twin accounts for the ambiguous nature of the world and the origin of evil.

This observation applies equally to the dualistic structure in some versions of the earth-diver myths. The devil moves in the various versions of this myth from companion to antagonist of God, possessing power to challenge the deity.

Sacrifice of primordial beings

Creation and sacrifice. In many cosmogonic myths, the narrative relates the story of the sacrifice and dismemberment of a primordial being. The world is then established from the body of this being. In the myth *Enuma elish*, the god Marduk, after defeating Tiamat, the primeval mother, divides the body into two parts, one part forming the heavens, the other, the earth. In a West

African myth, one of the twins from the cosmic egg must be sacrificed to bring about a habitable world. In the Norse *Prose Edda*, the cosmos is formed from the body of the dismembered great Ymir, and, in the Indian Ṛgveda, the cosmos is a result of the sacrifice of man.

In these motifs of sacrifice, something similar to the qualification of the undifferentiated matter of creation is suggested, for, just as the primal stuff of creation must be differentiated before the world appears, the sacrifice of primordial beings is a destruction of the primal totality for the sake of a specific creation.

When the victim of the sacrifice is a primal monster, the emphasis is on the stabilization of the creation through the death of the monster. The monster symbolizes the strangeness and awesomeness occurring when a new land or space is occupied. The "monster" of the place is the undifferentiated character of the space and must be immobilized before the new space can be established.

In a myth from Ceram (Molucca Islands), a beautiful girl, Hainuwele, has grown up out of a coconut plant. After providing the community with their necessities and luxuries, she is killed and her body cut into several pieces, which are then thrown over the island. From each part of her body a coconut tree grows. It is only after the death of Hainuwele that mankind becomes sexual; that is, the murder of Hainuwele enables mankind to have some determination in the process of bringing new life into the world.

Theological and philosophical doctrines. Myths and poetic renderings in legends, sagas, and poetry express the basic cultural insights into some of the elements involved in the human consciousness about creation. Theological, philosophical, and scientific theory are types of rationalizations of these basic insights in terms of the particular culture and historical periods of the cultures in question.

The attempt to integrate the meanings of primordiality, dualisms and antagonisms, sacrifices, and ruptures and to meet demands of some kind of logical order and, at the same time, keep alive the meaning of these structures as religious realities, objects of worship, and a charter for the moral life, has led to the development of doctrines.

In "primitive" and archaic societies, the correct ritual enactment of mythical symbols ensures the order of the world. These rituals usually take place at propitious moments (*e.g.*, at the birth of a child, marriage, the founding of a new habitation, the erection of a house or temple, the beginning of a new year). In each case, the seemingly practical activities imitate the mythic structure of the first beginning.

Theological and philosophical speculations and controversies centre within and between religious communities over the issues of the primordial nature of reality, dualisms, the process of creation, the nature of time and space. A doctrine of creation must contain or suggest the manner in which all cultural meanings, empirical and abstract, constitute an integral totality. Speculations based on the initial insights of a mythical theme explicate some principle in the myth as a basis for generalization and logical form on which all elements and themes may be ordered.

Transcendence and otherness. Doctrinal positions may be modelled around any or all of the themes of the cosmogonic myth. If the emphasis falls upon creation by a high god through his thought, word, or other mode, the problem of the otherness and difference between creator and creature becomes a source of theological discussion and philosophical speculations. In Judaism, Christianity, and Islām, the classical locus of this issue is found. All of these religions have theological traditions that raise this problem. Related to this issue is the transcendence and arbitrary action of the creator deity. Because he is prior to the world and its creatures, the question arises whether there are modes of creaturely knowledge or apprehension capable of knowing him; of whether he is subjected to the same categories of being as his creatures; of whether his time and space are the time and space of his creation.

The problem of difference between creator and creature

To some extent, the a priori nature of this type of deity creates an apparent dualism between the creator and the world and creatures. This dualism is mediated in various forms in the traditions. In Judaism it is mediated through nature and the covenant Yahweh has with his people; in Christianity through the mediatorship of his son, Jesus Christ; and in Islām through the sacred word of the Qur'ān by the prophet Muḥammad. Even within these traditions, however, the transcendent nature of the deity and his mediatorship through some other being or principle does not settle the doctrinal issue, for different cultural-historical periods of these traditions offer a variety of theological speculation concerning the nature and meaning of the deity, the world, and the mediator. The traditions offer a structure through which such speculation is ordered and clarified.

Creation through emanations. The theme of emergence is related to theological and philosophical notions of emanations from a single principle and the idea of the transmutation of being. Ideas of this kind are found in "primitive" religion (Dogon, Polynesian), in Taoism, and in the Pre-Socratic philosophers Thales and Anaximander.

In one version of the Dogon myth, creation proceeds from a small seed. Within the seed spontaneous movements begin. These movements, which burst from the shell of the seed and make contributions in space, create all forms of beings and the universe. Similarly, in the Polynesian myth Ta-aroa develops the world out of himself and the shell in which he lived.

The notion of the universe in a perpetual flux

A pervasive theme in Chinese thought is that of a universe in a perpetual flux. This flux follows a fixed and predictable pattern either of eternal oscillation between two apparently opposed poles or of a cyclical movement in a close orbit. The oscillation pattern is expressed by the Yin–Yang doctrine of Taoism. In the five element doctrine, a cyclical movement is correlated with the five elements, earth, wood, metal, fire, and water; these in turn form an equivalence with the third month of summer and with spring, autumn, summer, and winter, respectively. These parallelisms then form equivalences with the five directions, and they in turn with the five primary colors. Ancient Chinese thinkers never discuss an initial conscious act of creation. The cyclical movement itself produced the empirical and abstract form of the cosmos. The oscillation between the Yin and the Yang forms a correlation in all phenomena extending to the realms of time, space, number, and ethics.

Thales thought that the fundamental principle of cosmos was water. The earth floated on water; water was the natural cause of all things. Anaximander taught that there was an eternal undestructible something out of which everything arises and everything returns. In other words, the fundamental substratum of the world could not be an element of the world. The importance of Anaximander was in his use of the term *archē* ("beginning" or "rule") to refer to a principle unlike any other principle or element in the world to explain the cause of all other things in the universe.

Dualisms. Dualistic conceptions of creation come to the fore in the theme of earth-diver myths, in which there is an antagonism between the co-creators of the universe. This conception is present again in myths of divine twins and in Zoroastrianism where the Ormazd and Ahriman represent the creative and destructive principles in creation. In some sense this is not an ontological dualism for the first creative act of Ormazd was the limitation of time and thus the limitation of the power of Ahriman to carry out his destruction. Doctrines of this kind are related to the origin of evil in the world.

SKEPTICISM REGARDING CREATION

Alongside the various myths and doctrines regarding creation, there are equally skeptic positions concerning the unknowability of creation. This critique is present in several religious and philosophical traditions. It may be correlated with the mythical meaning of *deus otiosus*, the deity who retires from the world after his creation, or with the mythic theme from some earth-diver myths that emphasize the physical and intellectual fatigue of the deity after creation. In the first case, the removal of the

deity from creation leaves no access to his plan or will; in the other case, because of the fatigue of the deity who has exhausted all of his knowledge in creation, there is thus nothing for man to learn from him.

In the Indian tradition the Ṛgveda, an ancient sacred text, expresses skepticism in this manner:

> He, the first origin of this creation, whether he formed it all or did not form it,
> Whose eye controls this world in highest heaven, he verily knows it, or perhaps he knows not.

The Buddha declared certain cosmological and metaphysical questions unanswerable. His refusal to answer questions of this kind gave rise to the "silence of the Buddha" as a philosophical style in Buddhism. They included such questions as: whether the world is eternal or not or both; whether the world is finite (in space) or infinite or both or neither.

In the Chinese tradition Kuo Hsiang (died AD 312) questioned the origin of the basic oscillation of the Taoist movement. For Hsiang there is no such thing as Non-Being for Being is the only reality. Being could not have evolved from Non-Being nor can it revert to Non-Being. As Kuo Hsiang put it,

> I venture to ask whether the Creator is or is not? If He is not, how can He create things? If He is, then (being one of these things), He is incapable of creating the mass of bodily forms. . . . The creating of things has no Lord; everything creates itself. Everything produces itself and does not depend on anything else. This is the normal way of the universe.

Skepticism of this same kind is expressed by Parmenides, a Pre-Socratic, and in the modern tradition of Western philosophy from Immanuel Kant's *Kritik der reinen Vernunft* (1st ed. 1781; Eng. trans., *Critique of Pure Reason*, 1929) to Ludwig Wittgenstein's *Tractatus Logico-Philosophicus* (1922). Skepticism of this kind about the nature of the cosmic order and especially about the ultimate origin of the universe places limitations on the possibility of the rational consciousness to authentically ask these questions. In some instances theologians have agreed and held to a notion of revelation as a response to these unanswerable questions. In other cases, the questions themselves have been labelled nonsensical.

The work of Hartshorne and Reese

Charles Hartshorne and William Reese, 20th-century U.S. philosophers, have attempted to clarify and criticize all possible rational reflections concerning the relationship of deity to the universe. They state two opposed positions. The first is that of classical theism in which there is the admission of plurality, potentiality, becoming, as a secondary form of existence outside of God. The other position, that of classical pantheism, says that though God includes all within himself, he cannot be complex or mutable, for such categories only express human ignorance and illusion. They attempt to overcome this dilemma by combining these contrary poles into a dipolar conception of the meaning of deity. Because classical theism is primarily a Western approach to the problem and classical pantheism an Eastern approach, the dipolar conception of the relationship of deity to the created order is at the same time a synthesis of Western and Eastern thought. In addition to this, these philosophers set forth a method of analyzing all conceptions of deity and world according to basic religious and rational categories. As metaphysicians they go far in refuting the skepticism regarding rational knowledge of the relationship between the deity and the universe.

BIBLIOGRAPHY

Cosmogonic myths: CHARLES H. LONG, *Alpha: The Myths of Creation* (1963), gives examples of various types of cosmogonic myths from different cultures. *La Naissance du Monde* (1959), gives myths from several cultures with a commentary by MIRCEA ELIADE. For ancient Near Eastern myths, see *Ancient Near Eastern Texts Relating to the Old Testament*, ed. by JAMES B. PRITCHARD, 3rd ed. with suppl. (1969). JOHANNES PEDERSEN, *Israel*, 4 vol. (Eng. trans. 1926–40), is a cultural-religious study that shows the relationship between the creation myth, the land, and kinship system (see especially the chapter on "World of Life and Death"). For the nature and structure of myths and symbols, see ERNST CASSIRER, *Philosophie der symbolischen Formen*, 4 vol. (1953–56; Eng. trans., *The Philosophy of Symbolic Forms*, 3 vol., 1953–55); MIRCEA ELIADE, *Le Mythe de l'éternel retour* (1949; Eng. trans., *The Myth of the Eternal Return*, 1954), *Traité d'histoire des religions* (1948; Eng. trans., *Patterns in Comparative Religion*, 1958), and *Myth and Reality* (1963).

The development and structure of Greek myths: JOHN BURNET, *Early Greek Philosophy*, 4th ed. (1930, reprinted 1963), is a well-written interpretation of the pre-Socratic myths of creation. ARNOLD EHRHARDT, *The Beginning* (1968), shows the common structure of the cosmologies of the Gospel of John and pre-Socratic thinkers. BENJAMIN FARRINGTON, *Science and Politics in the Ancient World*, 2nd ed. (1946), attempts to place the development of scientific thought in ancient Greece within the context of the social history of the period. W.K.C. GUTHRIE, *In the Beginning* (1957), is limited to a study of Greek notions of the origins of man and the state. JEAN-PIERRE VERNANT, *Mythe et pensée chez les grecs*, 2nd ed. (1966), explores the interrelationships of time, space, memory, and technique in Greek myths.

Christian doctrine: For a theological history of the Christian doctrine of creation in its variety and continuity, see JAROSLAV PELIKAN, *Development of Christian Doctrine* (1969), *The Christian Tradition* (1971), and *Historical Theology: Continuity and Change in Christian Doctrine* (1971). JOHN MACQUARRIE, *Principles of Christian Theology* (1966), presents a structural and systematic analysis of the elements of Christian theology, showing how the doctrine of creation fits into theological systems.

Islām: DE LACY O'LEARY, *Arabic Thought and Its Place in History*, rev. ed. (1939, reprinted 1963), deals with the internal and external sources of Arabic philosophy and cosmology. SEYYED HOSSEIN NASR, *An Introduction to Islamic Cosmological Doctrines* (1964), explicates a tradition in Arabic thought that expresses creation in symbolic and cosmological images.

Zoroastrianism: Several Zoroastrian myths and doctrines of creation are found in R.C. ZAEHNER, *The Dawn and Twilight of Zoroastrianism* (1961).

Chinese: ARTHUR F. WRIGHT (ed.), *Studies in Chinese Thought* (1953), brings together ten essays on various aspects of Chinese thought. Most valuable is DERK BODDE, "Harmony and Conflict in Chinese Philosophy," pp. 19–80. For a history of Chinese philosophical speculation as it relates to cosmogony and cosmology, see FUNG YU-LAN, *A History of Chinese Philosophy*, 2nd ed., 2 vol. (Eng. trans. 1952–53).

Indian: Philosophical speculations about creation in the various schools of Indian philosophy can be found in SURENDRANATH DAS GUPTA, *A History of Indian Philosophy*, 5 vol. (1922–55); and HEINRICH ZIMMER, *Philosophies of India* (1951), especially the chapter on philosophies of time. ALAIN DANIELOU, *Le Polythéisme hindou* (1960; Eng. trans., *Hindu Polytheism*, 1964), is a description and interpretation of the gods of Hinduism in relationship to their philosophical meaning. T.R.V. MURTI, *The Central Philosophy of Buddhism* (1955), is an explication of the Mādhyamika system of Buddhist philosophy that denies creation.

Comparative works: HAJIME NAKAMURA, *Ways of Thinking of Eastern Peoples* (1964), is a comparative work showing the similarities and contrast between Indian, Chinese, Tibetan, and Japanese modes of thought especially as it concerns creation. C.F. VON WEIZSACKER, *The Relevance of Science: Creation and Cosmogony* (1964), deals with the evolution of thought about creation from myth to scientific theory. The same ground is covered in MILTON MUNITZ, *Space, Time, and Creation* (1957). CHARLES HARTSHORNE and WILLIAM REESE (eds.), *Philosophers Speak of God* (1953), is a text in comparative philosophy that explores the rational bases for several conceptions of God and creation in Eastern and Western thought.

(C.H.Lo.)

Creed and Confession

Creeds and confessions of faith are authoritative formulations of the beliefs of religious communities (or, by transference, of individuals). The two terms are sometimes used interchangeably, but when distinguished, a "creed" refers to a brief affirmation of faith employed in public worship or initiation rites, while a "confession" is generally a longer, more detailed, and systematic doctrinal declaration. Both creeds and confessions were historically called symbols, and the teachings they contain are termed articles of faith or, sometimes, dogmas.

Role of belief

The role of belief within religion is interpreted differently in the various empirical disciplines and by the proponents of particular theological or philosophical positions. Traditionally, it has been considered the primary

factor in religion, but some modern scholars often regard beliefs as rationales for ritual, secondary expressions of religious experience or ideological sanctions for social and cultural patterns. The present article follows a current anthropological and sociological tendency to define religion as a symbolic system in which ideas and their concomitant attitudinal aspects and actions provide to an individual or group a model of itself and its world. From this perspective, every religion involves distinctive views or beliefs regarding the nature of ultimate reality.

Origins and functions of creeds. These beliefs, however, need not be explicitly articulated, but, may be wholly embedded and transmitted in rituals, myths, and social structures and practices. This is especially true in primitive religions. Even when differentiated from other factors, beliefs are frequently not stated in creedal form but are diffusely expressed in sacred writings, legal codes, liturgical formulas, and theological and philosophical reflection. This was true in the ancient cultural religions of Egypt, Mesopotamia, Greece, and Rome, and in traditional Hinduism, Confucianism, and Taoism. When, however, a religion is transmitted from one culture to another (as from Semitic to Hellenistic; *i.e.*, Palestine to Rome) or claims some degree of universal or exclusive truth, formal creeds often develop as aids in maintaining continuity and identity. They serve this purpose because the relative abstractness, comprehensiveness, and concentration of the verbal expressions of beliefs enable them to serve better than most other forms of religious symbolism as stable identifying marks in pluralistic, changing, proselytizing, and missionary situations.

Purpose of creeds

Creeds in the full sense are therefore found only in so-called universal religions, such as Zoroastrianism, Buddhism, Judaism, Christianity, Islām, and certain modern Hindu movements (*e.g.*, Brahmo Samaj). Even here they are of variable importance, with some groups rejecting all formal creeds. Confessions are less common. They function to define the distinctive beliefs of opposing or uniting groups within a given religion or to formulate doctrines appropriate to new circumstances, and are chiefly a Christian phenomenon during the period from the Reformation to the present.

Religions of the East. Related to creeds in the full sense are certain words and phrases which have partially creedal functions. Terms like *tao* (literally, the "way") in Taoism or *li* (rules of propriety) and *hsiao* (filial piety) in Confucianism summarize fundamental emphases of the religious systems of which they are a part. The endlessly repeated *mantra* (evocatative sacred syllables) of magic invocation, *Oṃ mani padme hūṃ* ("O, the jewel in the lotus"), especially popular in Tibetan Buddhism, is in one sense a profession of belief in the Avalokiteśvara (jewel's) presence in the world (lotus). Various Hindu *mantras*, most notably the Gāyatrī prayer from the Ṛgveda (3.62.10) that is learned as part of the initiation rites of Brahmin youth, also serve in part as professions of faith. Indeed, it is primarily through liturgical utterances (*e.g.*, the Lord's Prayer in Christianity), that religious identity is signalized and faith confessed in most religions.

More specifically creedal is the early thrice-repeated *tri-ratna* of Hīnayāna Buddhism: "I take my refuge in the Buddha. I take my refuge in the *dharma* (doctrine). I take my refuge in the *saṅgha* (monastic community)."

Religions of the West. Even earlier perhaps are such Zoroastrian formulations as "I profess myself a Mazdāh-worshipper, a Zarathustrian, enemy of the demons, servant of the Lord" (*Yasna* 12,1), whereby the believer declared himself a monotheist, a member of a specific community, and a dualist.

Islām. The intensely anti-polytheistic faith of Islām is summed up in the *shahādah:* "there is no God but God; Muḥammad is the Prophet of God." This is proclaimed in the daily calls to prayer from every mosque, and every Muslim must recite it aloud with full comprehension and assent at least once in his life, and profess it without hesitation until his death. Doctrinal disputes have contributed to the development of additional creedal formulations called *'aqā'id* (singular, *'aqīdah*), but these do not divide Islām into clearly marked confessional groupings or denominations such as exist in Christianity.

Judaism. In Judaism, the central affirmations of belief are parts of worship; *e.g.*, the confessions of the oneness of God in the Shema (Deut. 6:4 "Hear, O Israel: The Lord our God is one Lord") and of the resurrection of the dead in the *amidah* (standing prayer). Of the various medieval attempts to formulate creeds, the most enduring has been Maimonides' Thirteen Principles of Faith, but these have never become formally binding. The Reform movement's doctrinal declarations, such as the Pittsburgh Platform (1885), have been without lasting influence. The reason for this paucity of creeds is that Jewish identity has been chiefly defined in terms of the observance of the commandments and of the Oral Law, not the acceptance of doctrines.

Christianity. In Christianity, in contrast, there are over 150 officially recognized creeds and confessions. In part this is because the church was from the beginning doctrinally oriented, making the acceptance of a specific kerygma (proclamation) a condition for membership. The faith of the community was expressed in acclamations such as "Jesus is Lord" (*e.g.*, Rom. 10:9, I Cor. 12:3) and in longer, partly stereotyped summaries of essential beliefs (*e.g.*, I Cor. 15:3 ff.) For the New Testament community, in contrast to some Christian groups in later times, a creedless Christianity was inconceivable.

Fully formed creeds first developed for use in baptismal rites and catechetical instruction. They generally had three sections concerned with God the Father, Jesus Christ, and the Holy Spirit, but were variable in wording and content and only gradually became standardized.

The Ecumenical Creeds: Apostles', Nicene, and Athanasian

This process culminated in the West in the Apostles' Creed, which is now almost universally recognized by Western churches, and is still used in baptismal rites as well as public worship by Catholics and most Protestants. This creed is wholly derived from New Testament affirmations, but the 5th-century legend that the Twelve Apostles were its authors is without foundation. Not until the 8th century is it quoted in its present wording. Its sources, however, are to be found in earlier baptismal creeds, most probably in the Old Roman Symbol, which appears to go back in its essentials to the 2nd century. As is true of other creeds, it is in part intended to exclude heretical views. For example, against Gnosticism and Marcionism (dualistic heresies), it emphasizes that God, not an evil demiurge, is the creator of the world, and against docetic views that Jesus was a heavenly being with a phantom body, it insists that he was born of the Virgin Mary and actually suffered and died and was buried.

The Nicene Creed exists in two versions and represents a new type of doctrinal statement. It was first formulated at Nicea in 325 by the first of the universal, or ecumenical, councils, after Christianity became the official religion of the Roman Empire, and was designed not as a baptismal confession but as a binding standard of orthodox teachings. Its second version has become the most fully ecumenical of Christian creeds, accepted in East and West alike, including the major Protestant bodies. In Eastern churches, it is regularly employed in both Baptism and eucharistic worship; in the West, only in the Eucharist, and chiefly by Roman Catholics, Anglicans, and Lutherans.

The first version of this formulary is that promulgated at the Council of Nicea in 325, but the second version, the "Niceno-Constantinopolitan Creed," which has everywhere become standard and is generally referred to as the Nicene Creed, was affirmed at the Council of Chalcedon (451) as the Nicene "faith of the 150 fathers" (*i.e.*, the Council of Constantinople of AD 381). In 4th- and 5th-century usage, "the Nicene faith" did not refer to the creed of Nicea as such, but rather to its teaching.

Both versions make the same fundamental affirmations against the Arian heresy that denied the equality of the Father and the Son, asserting that Jesus Christ, the Son of God, is *homoousion* ("of one substance") with the

Father. They are also both derived from Eastern baptismal formulas, though which ones is in dispute.

The *filioque* clause, affirming that the Spirit proceeds "from the Son" as well as the Father, was inserted into the text in Spain during the 6th century and gradually spread to all Western churches, but was probably not used in Rome itself until 1014. Eastern Christians continue to reject this addition, though now they do not generally regard it as heretical, especially if it is understood in the sense of "through the Son."

The Athanasian Creed, also called the *Quicumque vult* from its initial words, is the last of what in the West are regarded as the three catholic or ecumenical creeds. It has received some slight recognition in the East, but only since the 16th century. While officially accepted in the Roman Catholic, Anglican, and Lutheran communions, its liturgical use has greatly declined in recent centuries. In part this is because it is in form more a theological exposition than a creed, and in part because of the damnatory clauses that exclude from salvation all those who do not accept every detail of its teaching. The main themes are the nature of Christ and the Trinity, and these are developed in opposition not only to Arianism but also apparently to later heresies such as Nestorianism and Eutychianism. While its doctrine can in general be attributed to the 4th-century Church Father Athanasius, he was not its author. It probably originated in southern France about 450–500, although there is no scholarly consensus on this point.

Origins and functions of confessions. Official doctrine has chiefly developed during later periods of church history by the formulation of confessions of faith, rather than new creeds. This process did not begin, however, until the 16th-century Reformation. During the Middle Ages, dogmas evolved slowly, almost unconsciously, and then were ratified from time to time by decisions of the church councils, such as the decision on the seven sacraments at the Council of Ferrara-Florence in 1439. The Protestant Reformers, however, were confronted with the need to define and make legitimate their views over against the established system, and thus issued comprehensive manifestos that, much more than the early creeds, were not only catalogues of beliefs but also interpretations and apologies for them. The Roman Catholic and Eastern Orthodox churches responded with their own confessional statements.

The Augsburg Confession

Lutheran confessions. The Augsburg Confession (1530) was the first of these statements, and still remains the most authoritative standard in Lutheran churches. It (as well as the Apology of the Augsburg Confession of 1531) was written by Philipp Melanchthon and approved by Martin Luther, and presents an irenic statement aiming to show that the pope and his allies, not the Reformers, had departed from Scripture and the tradition of the early Fathers. Luther's Small Catechism also enjoys official status in all Lutheran churches and has been determinative for most Lutheran preaching and instruction. The Formula of Concord (1577) further defined the Lutheran position in reference to controversies both within and outside the ranks. These four writings, together with the Large Catechism (1529), the Schmalkald Articles, and the Treatise were assembled into the Book of Concord (1580), which has official status in many Lutheran churches.

The Heidelberg Catechism

Reformed churches confessions. In the Reformed tradition stemming from John Calvin (1509–64) and Huldrych Zwingli (1484–1531), each national church produced its own confessional documents. No one of these is authoritative for all, though some (*e.g.*, the Heidelberg Catechism; 1563) are widely esteemed and used. In Switzerland, the First (1536) Helvetic Confession and the Second (1566) Helvetic Confession are the most generally accepted. The French Gallican Confession of 1559 is much admired, and in the Low Countries, the Belgic Confession of 1561 is important. The Netherlands was also the site of the international Synod of Dort (1619) that presented an especially rigid statement of Calvinism against Arminianism (a view that asserted the compatibility of God's sovereignty and man's free will). This same emphasis, combined with Puritan covenental theology, is reflected in the English Westminster Confession of 1646 that in Scotland replaced the Scots Confession in 1560, was adopted with modifications by Congregationalists and many Baptists, and still remains standard for American Presbyterian churches, though with some revisions.

The Anglican Communion. The Thirty-Nine Articles (1563) is the only doctrinal formulation other than the early creeds recognized in the Church of England and its offshoots, but its authority is not great. In the Anglican Communion, *The Book of Common Prayer* plays the identity-sustaining role served by confessions in Lutheran and Reformed churches. The Thirty-Nine Articles, abbreviated to 25, are also the chief doctrinal standard in the Methodist churches, but their authority is uncertain.

Confessions of other Protestant groups. Confessional documents are of little significance for most of the radical groups (*e.g.*, Anabaptists) coming out of the Reformation. To be sure, the Anabaptist Schleitheim Confession (1527) was historically important, the Dordrecht Confession (1632) still has some standing in Mennonite churches, and various Baptist and Congregationalist statements could also be mentioned. The general tendency in these churches, however, has been to oppose formal creeds and confessions for fear of stifling the workings of the Holy Spirit or imperilling the sole authority of the Bible or, in theologically liberal circles, endangering freedom of thought and conscience.

Roman Catholic doctrinal statements. Roman Catholic doctrinal statements are not usually called confessions, but the presentation of the distinctive points of Catholic dogma in the Decrees and Canons of the Council of Trent (1564) is as fully elaborated as are Protestant confessional writings. The dogmatic constitutions of the first Vatican Council (1869–70) and papal definitions of the dogmas of the Immaculate Conception (1854) and of the Assumption (1950) also have some of the character of confessions.

The Decrees and Canons of the Council of Trent

Eastern Orthodox doctrinal statements. Eastern Orthodoxy responded to Protestant and Roman Catholic challenges with the confessions of Peter Mogila, Metropolitan of Kiev, in 1643 and of Dosítheos, the Patriarch of Jerusalem, in 1672, both adopted by the Synod of Jerusalem (1672), as well as with the Catechism of Philaret, Metropolitan of Moscow, revised and approved by the Holy Synod in 1839. The Orthodox, however, place little emphasis on these documents, for they regard only the Nicene Creed with its Chalcedonian additions as fully authoritative, and in practice also treat their historic liturgies as doctrinally more important than later statements.

Creeds and confessions today. Recently new types of confessions have begun to emerge. With the decline of state churches, confessions are no longer legally established norms and can once again regain their original function of witnessing to basic convictions. Especially notable in this respect is the Barmen Declaration, formulated in 1934 by a group of Reformed and Lutheran churchmen in opposition to the Nazi-influenced "German Christians." Because of the advance of the ecumenical movement, recent confessional statements have usually been unitive rather than divisive. The doctrinal basis of the World Council of Churches is limited to the affirmation that it is "a fellowship of churches which accept our Lord Jesus Christ as God and Savior" (1961). Preparation of joint Protestant and Roman Catholic official translations into English of the Apostles and Nicene Creeds commenced in 1969. Another characteristic of contemporary doctrinal statements, such as those of the Roman Catholic second Vatican Council (1962–64) and the Presbyterian (U.S.A.) Confession of 1967, is the attempt to reformulate traditional beliefs in ways appropriate to modern circumstances.

The Barmen Declaration

Despite these developments, creeds and confessions are losing influence in both Christian and non-Christian

groups. They are, among other things, often attacked as obstacles to the individual's freedom of thought. This objection applies with special force against a fideistic attitude, such as is illustrated in extreme form by the well-known saying attributed traditionally, though not altogether correctly, to the 2nd-century North African Church Father Tertullian, *credo quia absurdam est*, "I believe because it is absurd." It is less applicable to another ancient and theologically more common approach summed up in the 11th- and 12th-century theologian Anselm's (and, in a somewhat different wording, Augustine's) classic phrase, *credo ut intelligam*, "I believe in order that I may understand." The latter view claims that true faith promotes rather than suppresses inquiry and intellectual liberty.

Yet, whatever the merits of such views, doctrinal convictions are clearly weakening, even in traditionally creedal and confessional bodies. The search for creedless religion is widespread. There is the possibility, however, that this trend may be eventually reversed because the quest for religious community is also strong, and may require the formation or re-affirmation of community-identifying beliefs; *i.e.*, of creeds or confessions.

BIBLIOGRAPHY. Still the most comprehensive treatment of creeds in all religions is "Creeds and Articles," in the *Encyclopaedia of Religion and Ethics*, 4:231–248 (1912). For anthropological, sociological, and phenomenological considerations, see the *International Encyclopedia of the Social Sciences*, 13:398–414 (1968); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation*, 2 vol., with new appendices, (1963); and J. WACH, *Sociology of Religion* (1944). Works devoted specifically to creedal and confessional formulations are rare for most religions, but see S. SCHECHTER, "The Dogmas of Judaism," *Studies in Judaism*, pp. 147–181 (1896); and A.J. WENSINCK, *The Muslim Creed: Its Genesis and Historical Development* (1932). For Christianity, the fullest collection of texts remains P. SCHAFF, *The Creeds of Christendom*, 3 vol., 6th ed. (1919); for Roman Catholicism, H.J.D. DENZINGER and A. SCHONMETZER, *Enchiridion Symbolorum* (1963); and W.M. ABBOTT (ed.), *The Documents of Vatican II* (1966); for Protestantism, T.G. TAPPERT (ed. and trans.), *The Book of Concord* (1959); and A.C. COCHRANE (ed.), *Reformed Confessions of the 16th Century* (1966). On the Ecumenical movement, see L. VISCHER (ed.), *A Documentary History of the Faith and Order Movement 1927–1963* (1963). Brief but representative collections are B.A. GERRISH, *The Faith of Christendom: A Source book of Creeds and Confessions* (1963); and J.H. LEITH (ed.), *Creeds of the Churches* (1963). Secondary works on early creeds include: O. CULLMANN, *Die ersten christlichen Glaubensbekenntnisse* (1943; Eng. trans., *The Earliest Christian Confessions*, 1949); J.N.D. KELLY, *Early Christian Doctrines*, 2nd ed. (1960); A.E. BURN, *The Athanasian Creed*, 3rd impression (1930); D.L. HOLLAND, "The Earliest Text of the Old Roman Symbol," *Church History*, 34: 262–281 (1965), and "The Creeds of Nicea and Constantinople Reexamined," *Church History*, 38:248–261 (1969). On later confessions, the fullest recent treatment with good bibliographies is E. MOLLAND, *Christendom* (1959), but this is usually supplemented by W.A. CURTIS, *A History of Creeds and Confessions of Faith in Christendom and Beyond* (1911); and C.A. BRIGGS, *Theological Symbolics* (1914).

(G.A.L.)

Cretaceous Period

The Cretaceous Period is the youngest period of the tripartite Mesozoic Era. Its name, proposed by Omalius d'Halloy in 1822, referred to chalk (*creta* in Latin), the characteristic rock formed during the period in many parts of Europe (*e.g.*, the White Cliffs of Dover, England). The Cretaceous Period is an internationally accepted time unit, which is geologically well defined regardless of the presence or absence of chalk; it spans 71,000,000 years, from approximately 136,000,000 to 65,000,000 years ago. The Cretaceous was preceded by the Jurassic Period and succeeded by the Tertiary Period of the Cenozoic Era.

The rocks formed during the Cretaceous Period are called the Cretaceous System. They are distributed in various parts of the world and show considerable variation in their lithological character and the thickness of the sequences. Because a large-scale marine inundation took place during this period, lower areas of various continents were extensively covered by the fine-grained marine sediments, namely chalk, marl, and clay. On the other hand, orogenies, or mountain-building episodes, accompanied by volcanism and plutonic intrusion, took place in the circum-Pacific region and in the site of the Alpine system. In these geosynclinal areas predominantly clastic sediments, including conglomerates, sandstones, shales, and tuffs accumulated in sinking troughs and basins (geosynclines) and were later folded. In many parts of the circum-Pacific area igneous rocks of Cretaceous age are widely exposed.

The Cretaceous Period was a time of great inundation by shallow seas that created swamp conditions favourable for the accumulation of fossil fuels. Many important oil and gas fields produce from subsurface Cretaceous (and also overlying Tertiary) deposits; among them are those of western Siberia, southwestern Asia, the Persian Gulf, northern Africa, Mexico and other areas in North America, Venezuela, and Argentina. Coal measures are also found in some parts of Cretaceous sequences in Siberia, Australia, New Zealand, Mexico, and western United States.

Certain porous, permeable sandstones in the Cretaceous sequence are important for underground water resources (*e.g.*, the Dakota Sandstone of western United States), and certain others are used as glass material. Bauxite produced from the Cretaceous weathering of limestone is an aluminum source in Europe. Cretaceous limestones are quarried for cement and building stone. Smaller pits of clay, shale, and bentonite are worked for ceramic material. The widespread phosphatic beds in the Upper Cretaceous–Lower Tertiary sequence of northern Africa and southern Soviet Union are of economic importance.

In the circum-Pacific mountain systems and island arcs a variety of metals, such as gold, silver, copper, lead, zinc, molybdenum, tungsten, tin, iron, and manganese, were concentrated as ore deposits of various dimensions associated with the igneous activity of the Late Mesozoic.

This article treats the rocks, life, and environments of Cretaceous time. For information on the preceding and subsequent intervals of geological time, see JURASSIC PERIOD; and TERTIARY PERIOD; for an overview of the Cretaceous Period, see MESOZOIC ERA; and for additional detail of relevance on stratigraphy and paleontology, see STRATIGRAPHIC BOUNDARIES; and FOSSIL RECORD.

WORLD STRATIGRAPHY AND PALEOGEOGRAPHY

Stages and stratigraphic units

The Cretaceous System, the rock sequences formed during the 71,000,000-year Cretaceous time interval, is stratigraphically divided into the Lower and the Upper Cretaceous subsystems, each of which is subdivided into six stages. The 12 stages are indicated in the left column of Table 1. Their names were taken from localities of rock exposures in France and adjacent areas in Switzerland and The Netherlands. The lower four stages are called the Neocomian Series and the upper four the Senonian, although some authors may exclude the Maastrichtian Stage from the latter. Each stage is defined by the rock sequence in the type of exposures and also by a particular fossil content of characteristic species. Ammonites, among other fossils, are important for the definition and the international time correlation of the stages. The indices of the ammonite zones in the type sequences in France and related areas in western Europe are shown in the second column of Table 1. The third column cites the belemnites, bivalves, and other ammonites selected as zonal indices for northern Europe, which was part of the Cretaceous boreal province. The stages, thus defined, are now internationally used as time-stratigraphic units, although the characteristic species may differ in separated regions. Tables 1 and 2 show the correlation of the better studied sequences among selected regions of the world.

Some geologists have employed the Danian as the uppermost stage of the Cretaceous from a type section at the top of the chalk sequence in Denmark. The section con-

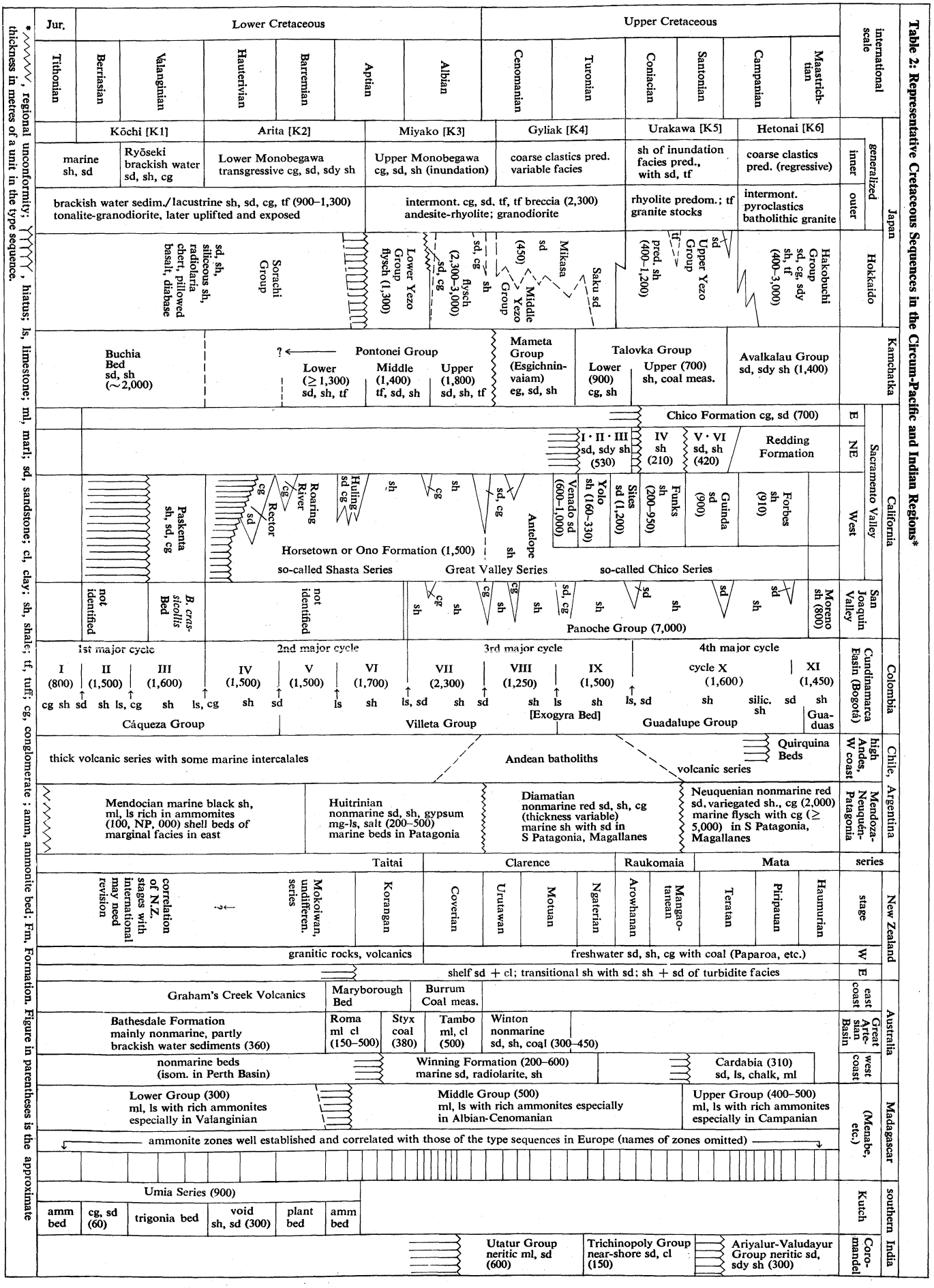
Table 1: Stratigraphic Subdivisions of the Cretaceous System in Europe and North America*

	series	stages (types in France and adjacent areas)	ammonite zones in the type sequences of western Europe (France, England, etc.)	zones in the boreal province and related areas (Soviet Union, etc.)		England	Germany (Westphalia, Emsland, etc.)	Eastern Alps	North American western interior (Wyoming, Kansas)	North American Gulf Coast (Texas, Mexico)
Upper Cretaceous		Maastrichtian	*Sphenodiscus binkhorsti* *Hoploseaphites constrictus* *Pachydiscus neubergicus*	*Belemnella arkhangelskii* *Belemnella lanceolata*	*Acanthoscaphites tridens*	Trimingham Ch	Lanceolaten-senon (ch, 120)	Zwieselalm Bds (500) Nierentaler Beds (pelagic ml, 400)	Lance (nonmarine) (300–50) Fox Hills sd (200–50)	Navarro Group ml, sd, cl (130–200)
	Senonian	Campanian	*Bostrychoceras polyplocum* *Hoplitoplacenticeras vari*	*Belemnitella langeri* *B. mucronata senior*	*Trachyscaphites spiniger*	Norwich Ch	Mucronaten-senon (sd, etc., 50)	Upper (300)	sd Pierre Shale (900)	Taylor Marl (100–400)
			Delawarella delawarensis *Diplacmoceras bidorsatum*	*B. mucronata alpha-* *Gonioteuthis quadrata*	*Scaphites hippocrepis*	Upper Chalk (350)	Quadraten-senon (mp, 200)	Middle (300)		
		Santonian	*Stantonoceras syrtale*	*Inoc. patootensis, Gonioteuthis granulata*			Granulaten-senon (sd, etc., 400)	Gosau Beds	Telegraph Creek sd	
			Texanites texanus	*Inoceramus cardissoides*	*Actinocamax westfalicus*		Emscher (ml, 500)	Lower (700)	Niobrara fm: Smoky Hills Chalk (170) Ft. Hays ls (20)	Austin Chalk (105)
		Coniacian	*Parabevahites emscheris*	*Inoceramus involutus*						
			Barroisiceras haberfellneri	*Inoceramus wanderi*	*Actinocamax lundgreni*		Schloenbachlipläner (50–100) Scaphitenpläner (30) Lamarckipläner (100) Labiatusmergel (30)		Carlile sh (90)	
		Turonian	*Subprionocyclus neptuni*	*Inoceramus costellatus*		Middle Ch (70)		ml (100)	Greenhorn ls (30)	Eagle Ford sh (140)
			Collignoniceras woollgari	*Inoceramus lamarcki*						
			Mammites nodosoides *Metoicoceras whitei*	*Inoceramus labiatus*	*Actinocamax plenus*				Belle Fourche (25)	Woodbine Sand (160)
		Cenomanian	*Dunveganoceras pondi* *Calycoceras naviculare*	*Scaphites aequalis*		Lower Chalk (70)	-kalk Cenoman -pläner (250) -mergel	ml, sd, dolomite breccia (200)		
			Acanthoceras rhotomagense	*Holaster subglobosus*						
			Mantelliceras mantelli *Submantelliceras martimpreyi*	*Schloenbachia varians*	*Neohibolites ultimus*					
Lower Cretaceous		Albian	***Stoliczkaia dispar*** *Mortoniceras inflatum*	*Neogastroplites mclearni* *Hysteroceras orbignyi*		Upper Greensand (30)	Flammenmesyel (50) Gault, sd (20)	dark ml (150)	Mowry sh Newcastle sd Skull Creek sh Fall River sd	Washita Group ls, ml, cl (70–150)
			Dipoloceras cristatum *Hoplites dentatus*	*Anahoplites daghestanensis* *Hoplites dentatus*		Gault cl (20–100)				Fredericksburg Gr ls, ml, cl (100)
			Douvilleiceras mammillatum *Leymeriella tordefurcata*	*Leymeriella tardefurcata*				variegated ml (50)		
		Aptian	*Diadochoceras nodosocostatum* *Parahoplites nutfieldensis* *Cheioniceras martinoides*	*Hypacanthoplites jacobi/ Parahoplites melchioris* *Cheloniceras tschernyshewi*		Lower Greensand (80–200)				Trinity Group sd, cg (200)
			Ch. meyendorfi-Tropaeum bowerbanki *Deshayesites deshayesi/D. forbesi* *Prodeshayesites tenuicostatus*	*Dufrenoyia furcata, Deshayesites dechi,* *Deshayesites weissi, Matheronites ridzewskyi*			Dunkler Neokom-shieferton (sh with sd wedges) (800–950)	black sand, red ml (100)	Fuson sh Dakota sd	
	Neocomian	Barremian	*Silesiles seranonis-Heteroceras astieri*	*Oxyteuthis jasykowi*		Wealden (300 ± 100): cl				marginal facies / offshore facies
			Nicklesia pulchella-Holcodiscus caillеaudianus			sd				
		Hauterivian	*Pseudothurmania angulicosta* *Subsayнella sayni*	*Simbirskites decheni*		cl	Speeton Clay (100)	Neocomian-aptychus ml	Rossfeld Beds (300)	rudistid ls, sd, red beds / cephalopod ls, ml
			Crioceratites duvali *Leopoldia castellonensis*	*Speetoniceras versicolor* *Homolsomites bojarkensis*		sd				
		Valanginian	*Saynoceras vericosum*	*Dichotomites petschorensis* *Polyptychites polyptychus*	*Buchia crassicollis*		Valendis sd (50–25)			
			Kilianella roubaudiana	*Polyptychites michalskii* *Temnoptychites hoplitoides*	*B. sublaevis*	cl				(1,000) / (400–600)
		Berriasian [Ryazanian]	*Subthurmannia boissieri*	*Tollia srenomphalia* *Paracraspedites spasskensis* *Riasanites rjasanensis*	*Buchia okensis*	Darlston beds	Wealden (200–250)	Schrambach Beds (100)		

Montana Group; Colorado Group; Gannett Group (nonmarine sd, sh); Chalk (300 – 450); Wealden (300 ± 100); Gulf Series; Comanche Series; Coahuila Series

* ~~~, regional unconformity; ⌒⌒⌒, hiatus; ls, limestone; ml, marl; sd, sandstone; cl, clay; sh, shale; tf, tuff; cg, conglomerate. Figure in parentheses is the approximate thickness in metres of a unit in the type sequence.

Table 2: Representative Cretaceous Sequences in the Circum-Pacific and Indian Regions*



tains a nautiloid genus, *Hercoglossa*, but no ammonites or belemnites. Recent investigation of microfossils have led many authors to ascribe the Danian to the Paleocene at the base of the Tertiary System, however. Seas have invaded the continents a number of times in geological history. The Cretaceous inundation was one of the greatest because of the extent and duration of the invasions on various continents. The marine transgression, however, may not be synchronous everywhere. The accompanying paleogeographic map indicates the regional differences in the mode of inundation. Major geotectonic conditions (regional deformation of the Earth's crust) also are shown.

Early Cretaceous time

In the Neocomian Epoch of Early Cretaceous, the marine invasion was extensive around the Arctic Ocean. The Neocomian marine beds, mainly consisting of black bituminous clays, were deposited on the Russian Platform, West Siberian Plain, part of Far East Siberia, Alaska, and adjacent areas in Canada. They contain *Buchia* and particular assemblages of ammonites of the boreal fauna. The boreal fauna extended as far south as western Germany and eastern England in Europe and to California in North America. Neocomian marine beds are present in the Andean region of South America, some parts of southern Africa, Madagascar, and western India. The Neocomian of the Tethys region (principal east–west trending geosyncline from Europe across southern Asia), as exemplified by well-known sequences in southeast France and adjacent areas, and also in Mexico, consists mainly of marls of offshore sedimentary facies (*q.v.*), and, in lesser amount, of limestone and calcareous sandstone of near-shore environmental character. Each facies is characterized by a particular biota.

In the same epoch, lacustrine and deltaic deposits, called the Wealden, were formed in lakes and rivers in southern England and adjacent areas of Europe. These contain dinosaur and other fossils. Nonmarine Neocomian beds of a similar type occur on the Atlantic coast of North America, the coastal area of Africa, various parts of Australia, and shoreward of the Neocomian Seas that invaded Siberia. The predominance of coarse clastics in some parts of these nonmarine Neocomian deposits may reflect the crustal instability at or immediately before the beginning of the Cretaceous Period. Earth movements would produce high and low areas, thereby increasing stream gradients and allowing coarser material to be transported to basins. In addition to topography, factors such as climate, of course, affect the nature of the sediment yield of drainage systems (*q.v.*).

Aptian and Albian stages

In the succeeding times of the Early Cretaceous, in Aptian and Albian Stages, the sea around the Arctic Ocean retreated. Although there are marine beds of those stages in the southern part of Russia, they are unrelated to the boreal provinces and are intimately connected with those of the Mediterranean (Tethys) region. In western Europe, the sea invaded the Wealden plain resulting in the deposition of glauconite-bearing green sands of near-shore facies. Above the sands lies black clay, called the Gault, of somewhat offshore but not deepwater facies; it is characterized by a wealth of ammonites. The cycle ended with deposition of the upper sand of regressive facies, the Upper Greensand of England (Table 1). In the Mediterranean region, marls and limestones continued to be deposited in Aptian and Albian times, although local variation in facies occurred from place to place, and Aptian glauconitic sandstone may rest disconformably on older rocks in some areas.

At approximately the same time, the sea transgressed the Gulf Coast of North America, depositing a sequence of sandstones, shales, marls, and limestones that are called the Comanche Series (Table 1). In the western interior a roughly contemporaneous transgression from the north occurred. Its maximum extent in the Albian, with *Gastroplites-Neogastroplites* faunas, reached as far south as Montana, whereas the invasion from the south, with *Oxytropidoceras* ammonite fauna, reached Kansas.

In Australia the basins that were filled with Neocomian nonmarine deposits were invaded extensively by an Aptian to Albian shallow sea, depositing marls or calcareous mudstones. Coastal swampy sediments formed marginally. In Late Cretaceous times the sea retreated almost entirely from Australia, except for several narrow areas. Fluviolacustrine sediments accumulated in the basins until the area became dry land.

Late Cretaceous time

Generally speaking, however, the marine inundation in the Late Cretaceous was of a grand scale in its global extension and its long duration. In Europe, for instance, almost the entire area, except for the Baltic Shield and several islands, was covered by the sea in which was laid down a stratigraphic sequence from Cenomanian to Maastrichtian (to Danian in some places). The rocks characteristically are chalk, with some interbedded marl, sandstone, and siliceous rock. A wide area of western Siberia was under the sea that transgressed from the north, depositing clays and silts and siliceous sediments. Still more extensive areas in the present-day deserts—Libya, Egypt, Arabia, and Central Asia—were invaded by a shallow-sea extension of the Tethys. Calcareous sediments predominated in these regions; clastics were only marginally distributed, as exemplified by the Nubian Sandstone in northeastern Africa. The littoral zone is often represented by phosphatic beds. The maximum inundation in most of the above regions occurred in Turonian to Senonian times.

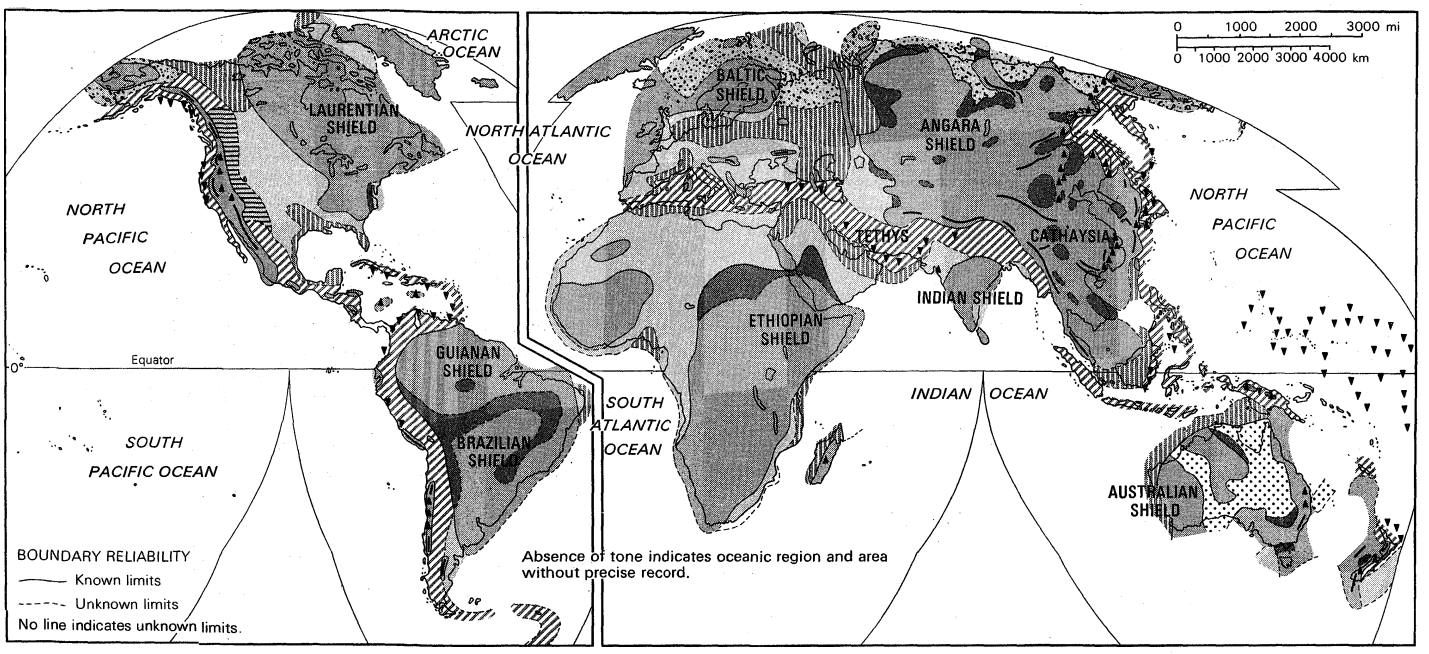
In North America (see map), the Late Cretaceous sea invaded deeply into the interior province, forming a broad seaway joining the Gulf of Mexico and the Canadian Arctic Ocean. In that area shale was the main constituent of the sequence. Chalk and limestones were deposited in the southern part of the area during the time of extensive inundation. From the rising cordillera to the west, clastics were transported to the sedimentary basins, where they formed sandstone wedges in the shaly sedimentary sequence. From those basins the present Rocky Mountains rose. The clastic deposits accumulated as thick units (about 4,500 metres [15,500 feet] maximum) in the subsiding basins in front of the rising cordillera, generally thinning out eastward. In this vast seaway, faunas of the Gulf series in the south and the Colorado–Montana groups in the north were dissimilar. The northern groups were closely allied to the Upper Cretaceous fauna in the interior provinces of Canada and in Greenland.

The Atlantic coast of the United States was also under a fluctuating shallow sea in Late Cretaceous times. Toward the end of the Cretaceous Period the sea retreated from North America. Nonmarine beds of the Lance Formation rest on the Montana Group. Remains of the dinosaur *Triceratops* have been found in that younger sequence.

Marine formations from Albian to Maastrichtian in age are exposed in the South Atlantic coastal regions of Brazil and Argentina and of Africa from Nigeria to Angola. In Turonian time this South Atlantic seaway was connected through Nigeria to an extension from the Tethys in North Africa, as shown by the affinity of the faunas.

In several regions facing the Indian Ocean—the eastern coast of South Africa and Mozambique, coasts of Madagascar, peninsular India, and western and northern Australia—shallow-sea Cretaceous sediments containing common or allied faunas are found. Three times of major transgression are recorded: (1) the older Neocomian, already noted from South Africa, Madagascar, and Kutch (western India); (2) the Albian to Coniacian; and (3) the Santonian to Maastrichtian. These cycles of sedimentation are recognized by examining the prolific fossil faunas and sedimentary sequences of Indo-Pacific affinity (see PALEOGEOGRAPHY; STRATIGRAPHIC BOUNDARIES). Marine Upper Cretaceous rocks are also present on the southern coast of Australia, facing the Antarctic Ocean, and in Antarctica.

A continuous depositional sequence from Cretaceous to Tertiary is observable in a few areas, such as Venezuela, Trinidad, and Egypt, but the precise definition of the Mesozoic–Cenozoic boundary is under debate among the specialists of several geologic disciplines. The retreat of the sea and the rise of the continents at the end of the pe-



Land and sea distribution in the Cretaceous Period.

riod or immediately after are of global scale and have significance for the revolutionary change in the history of life.

MOUNTAIN-BUILDING AND IGNEOUS ACTIVITY

In the circum-Pacific and Alpine zones, indicated as marine geosynclines and mountain systems (see map), remarkable crustal movements occurred during the Mesozoic and Cenozoic eras (see EARTH, GEOLOGICAL HISTORY OF). The movements began with earlier phases of geosynclinal sinking and sedimentation associated with submarine basaltic volcanism and continued during later phases of orogeny, which included folding, thrusting, regional metamorphism, plutonic intrusion, andesitic to rhyolitic volcanism, and uplift of mountains. The details in the time and mode of activity were dissimilar among provinces. There were a number of such major events of the Cretaceous Period.

Circum-Pacific events

In the North American cordillera, the so-called Nevadan orogeny took place in the Sierra Nevada and Klamath mountains in Late Jurassic to Neocomian times; the Santa Lucia, in the Coast Ranges, in Middle to Late Cretaceous times; and the Laramide, in the Rocky Mountains and Sierra Madre Orientale in Late Cretaceous to Early Tertiary times. In the South American Andean system, where marine geosynclinal conditions prevailed in Early Cretaceous times, the mountain building reached its climax in the Mid-Late Cretaceous. In New Zealand the most remarkable orogeny, called the Rangitata, took place in Early Cretaceous times. In Japan the Sakawa orogeny proceeded through a number of phases during the Cretaceous Period. In Far East Siberia the Verkhoyansk and other ranges were formed mainly by Late Jurassic orogeny, referable to the Cimmerian folding, but the crustal mobility continued into Cretaceous times.

In typical examples of these circum-Pacific orogenic systems, the regional metamorphism of high-temperature type and large-scale granitic emplacement occurred on the inner, continental side, whereas geosynclinal sinking, rapid sedimentation, and regional metamorphism predominated on the outer, oceanic side. The voluminous intrusion of granitic rocks, accompanied in some areas by extrusion of volcanic rocks, is one of the most significant events in geological history. The upheaval of the mountains, subject to intense erosion, kept pace with the subsidence of the trough where a thick pile of clastics was deposited. This is well exemplified by the upheaval of the Sierra Nevada, with intermittent emplacement of granitic bodies and the deposition in the Great Valley of California of thick units of Cretaceous shales and sandstones with many conglomerate tongues.

Another well-studied example is found in southwest Japan, where the inner (northern) belt is characterized by volcanic series, with some nonmarine beds, and also by granitic bodies, which can be grouped into three major cycles from Early Cretaceous through Late Cretaceous to Early Tertiary. The outer (southern) belt comprises the depositional basins of mainly marine Cretaceous to Lower Tertiary sequences of clastics, which can again be grouped into three major cycles. The accumulations are extremely thick in certain belts, in which sedimentation accumulated at the rate of 300 to 500 metres (1,000 to 1,600 feet) per 1,000,000 years (approximately ten times the present average marine rate). Similarly, rapid subsidence, with clastic sedimentation keeping pace with the upheaval of volcanic and granitic mountains, is suggested by the Andean system of South America. These sedimentary groups are often of flysch type (rapidly deposited clastic sediments) and frequently show features of sediments deposited by turbidity currents.

Cretaceous geosynclines

In addition to the mountain systems that were derived from the embedded geosynclines of preceding ages, new geosynclines appeared in Cretaceous times. The new geosynclines, with which submarine basaltic volcanism was associated, went on to subside with more sedimentation and finally became mountain systems or island arcs (*q.v.*) in Cenozoic times. Good examples are: the trough of the Franciscan Group on the Pacific side of California, that of Upper Cretaceous to Lower Tertiary sequence in the Antillean belt encircling the Caribbean Sea, that of the Upper Jurassic to Cretaceous Sorachi and Yezo groups in the meridional belt of Hokkaido and Sakhalin, that of the corresponding groups in Kamchatka, and the so-called Luzon Geosyncline in the eastern Philippines.

The Late Mesozoic crustal movements were not only active in the geosynclines and marginal mountain systems

that encircle the Pacific Ocean floor but also affected some inland areas extensively. In the continental area of China, Korea, and Far East Siberia, a crustal disturbance called the Yengshan movement occurred intermittently in the period from Jurassic to Cretaceous; it gave rise to inland basins of various dimensions, thrust ranges, faulted blocks, and even volcanoes and granite masses. The old, consolidated, continental mass was thus tectonically rejuvenated in Late Jurassic to Cretaceous times. The Maryborough Trough, where thick Cretaceous clastic and volcanic rocks were accumulated and then folded, may represent a similarly rejuvenated movement in eastern Australia.

Alpine–Himalayan events

Along the Alps–Himalaya orogenic system, geosynclinal conditions generally prevailed during Jurassic to Cretaceous time. In the earlier phases of the history of the geosynclines, submarine eruptive rocks from basic to ultrabasic magma (ophiolites) were predominant in some belts associated with radiolarian cherts (see SILICEOUS ROCKS). White pelagic limestones were formed, probably on offshore submarine mounts of moderate depth. Reef limestone accumulated on near-shore, shallower banks. Ammonite-bearing marls may have been deposited at intermediate depths. In the later phases, the major Alps–Himalaya Geosyncline was differentiated into geanticlinal ridges and sinking troughs, which migrated as thrusting proceeded. A particular type of sediment called flysch, typically consisting of alternating sandstones and shales with characteristic current marks and displaced fossils, was deposited in the trough. Conglomerates and breccias (*q.v.*) were deposited around the rising ridges, whereas bathyal marls were carried to the distal part of the trough. These and other tectonic and sedimentary features have been studied intensively in the Alps, Carpathians, Pyrenees, and other areas (see SEDIMENTARY FACIES).

In the Alpine system of Europe, orogenic phases are known at the middle of the Cretaceous, immediately before the Senonian Gosau Beds, and at the end of the period, although Alpine orogeny did not reach its climax until Early Tertiary times. The history in the Himalayas and other orogenic systems in southern Asia has not been investigated so precisely as has the European Alpine system, but the events in Cretaceous times seem to have been important in the tectonic development of those areas also.

Volcanic activity in the Pacific Ocean

In an extensive area of the central and western Pacific Ocean are found groups of volcanic seamounts of basaltic rock with summit depths of 1,300 to 21,000 metres (4,300 to 69,000 feet). Some of them are flat topped, with shelves on their flanks on which rest reef deposits or gravels indicating a shallow-water environment. Some of the deposits contain recognizable Cretaceous fossils. Although the seamounts were formed at various times during the Late Mesozoic and Cenozoic eras, a great number of them were submarine volcanoes building to the sea surface during the Cretaceous. They sank to their present deep levels some time after the age date of their youngest shallow-water fossil. In the northwestern Pacific between that submarine volcanic region and the Japanese islands, chalky sediments containing Cretaceous microfossils have been recovered in deep-sea coring operations. The extent of those Cretaceous deposits can be determined by seismic studies. Another such seamount off the southern coast of Hokkaido contains evidence of Cretaceous shallow-water gastropods, indicating that it too must have existed in Cretaceous times, although it may have moved to its present position resting on the westward-spreading sea floor.

CRETACEOUS LIFE

Invertebrates

Chalk and other offshore fine-grained calcareous sediments consist of the remains of nannoplankton (floating forms). The tiny coccolithophorids and other organisms that often are grouped under the name *Protista* are so minute that they must be examined by electron microscope. Smaller benthonic foraminiferids (protozoans) are also common in fine-grained sediments, and planktonic forms, which also contribute to chalk, are widespread and useful for correlating the ages of strata between distant places. Diatoms, radiolarians, ostracods, pollen, and spores also are identified microscopically in some Cretaceous sediments. Detailed study of these microfossils is helpful in the exploration for oil and gas.

On the very shallow banks of tropical to subtropical seas, reef-building organisms such as colonial corals, bryozoa, stromatoporids, and other calcareous algae flourished. Large foraminiferans (*Orbitoidaceae*); sessile (anchored) bivalves resembling corals, called rudistids; thick-shelled gastropods (nerineids); and certain kinds of echinoids were associated with them. Reefs were distributed mainly in the equatorial Tethys region, including the so-called American Tethys (Caribbean), and they extended to Japan and some central Pacific seamounts at certain times.

Ammonites were predominant among the marine invertebrates of the Cretaceous Period, which also was the final stage in the evolutionary history of this group of cephalopods, which has no Cenozoic survivors. Although the families that lived in the Jurassic persisted up to Early Neocomian times, new families became dominant after the middle of the Neocomian, and other renewed differentiations occurred at a number of times during the Cretaceous. Ammonites of various aberrant forms occur fairly commonly, exemplified by *Nipponites* and *Madagascarites*. Examples of gigantism are found in certain families, with shells as much as six feet in diameter. Pseudoceratitic sutures (superficially resembling a Triassic simpler type) are present in certain other Middle and Late Cretaceous families.

Belemnites, a group of coleoid cephalopods that left their cigar-shaped internal shells as fossils, occur in Cretaceous rocks but are extremely rare in the Late Cretaceous strata of the Pacific region. Why both ammonites and belemnites became extinct at the end of the Cretaceous Period still lacks explanation. Nautiloid cephalopods slowly and gradually evolved during that period and persist to the present without great change.

Bivalves (clams and relatives) and gastropods (snails) of the Cretaceous Period show a variety of forms, adapting themselves to various environments. Most Trigonians, for instance, lived in shallow, open sea; oysters (including *Ostrea*, *Amphidonta*, *Exogyra*, and *Gryphaea*) in the littoral zone; rudistids and nerineiids in Mediterranean reef facies; and some special groups (*e.g.*, *Trigonioides*, *Plicatounio*, *Nipponaia*) in nonmarine basins of eastern Asia. Certain thin-shelled bivalves, such as *Inoceramus*, *Aucellina*, *Maccoyella*, and *Didymotis*, were widely distributed in offshore sediments and are useful as time indicators because of their rapid evolution and wide dispersal. They disappeared at or before the end of the period.

Echinoderms, asteroids, crinoids, sponges, brachiopods, bryozoas, simple corals, and decapods (lobsters and crabs) occupied various habitats in the sea. Some of these groups are beautifully preserved and are good material for paleobiologic and evolutionary studies. Estherians (see BRANCHIOPODA), tiny bivalved crustaceans, lived in abundance in ponds and lakes and were preserved under certain favourable conditions in continental sediments.

Vertebrates

Cretaceous fish faunas generally appear more modern than those of the Jurassic because of the presence of groups that are closely allied to present-day rays, porbeagle sharks, herrings, and other bony fishes. *Macropoma*, a coelacanth, has been found fossilized in chalk.

In addition to the ichthyosaurs (fishlike reptiles), the mosasaurs (marine lizards), which seem to have eaten fishes and ammonites, were abundant. The pterosaurs, winged reptiles, evidently lived on the seashores or cliffs, soaring to the sea to catch fishes. Their remains are more common in marine sediments. Some pterosaurs in the Cretaceous Period (*e.g.*, *Pteranodon*) were very large with maximum wingspread of eight metres (27 feet). Although the fossils of birds are rare, *Hesperornis* (a flightless diver) and *Ichthyornis* (small winged form), whose remains were discovered from the Upper Cretaceous of

Kansas, represent fish-eating seabirds that had long jaws bearing teeth.

Dinosaurs were predominant among the land animals. Characteristic examples include *Tyrannosaurus*, the largest flesh-eating dinosaur; *Iguanodont*, a plant eater that walked on hind feet; *Trachodont*, which had numerous rough teeth; *Triceratops*, with three peculiar horns on the head; and *Struthiomimus*, which probably ate insects or seeds, judging from the elongated bill-like mouth. It is one of the most remarkable events in the history of life that the reptiles (see REPTILIA), which flourished during the Mesozoic Era, declined at the end of the Cretaceous Period, and the dinosaurs became extinct. The mammals (see MAMMALIA), on the other hand, were quite indistinct in Mesozoic times but burgeoned with multiple divergence and development early in the Cenozoic Era.

Plants

The land plants in the Early Cretaceous differed little from those in the Jurassic. The main constituents of the flora were cycadeoids (cyad-like plants), conifers, ginkgos, and ferns. Toward the middle of the period, angiosperms increased their dominance. In the Late Cretaceous the flora became more like those of the Cenozoic Era; they included figs, magnolias, poplars, plane trees, and willows. With the increasing predominance of flowering plants, insects also may have developed in the period; but their fossil record can exist only under special, rare conditions of preservation (see FOSSIL RECORD).

CRETACEOUS CLIMATES

On the evidence of fauna and flora distribution, climatic zones are roughly outlined between the tropical to subtropical equatorial Tethys region and the warm or somewhat cooler boreal and austral regions at higher latitudes. Expanded seas of the period should have produced an equitable climate. In the Cretaceous System, evaporites are comparatively few. Coal seams, indicating relatively high humidity, are intercalated in a number of places. Bauxite occurs at the unconformity between limestone sequences of the Cretaceous in some areas of Europe (Hungary, Yugoslavia, southern France, etc.), probably indicating that weathering (*q.v.*) took place under warm, humid conditions.

Although the quantitative data of paleotemperature are not sufficiently numerous, the available measurements by an oxygen-isotope method on some shells (belemnites and others) indicate warmer seawater even in the boreal region (see CLIMATIC CHANGE; DATING, RELATIVE AND ABSOLUTE). They also suggest some decline of the temperature in the Maastrichtian Stage and possibly also in the Cenomanian. On the basis of the paleomagnetic study of some Cretaceous rocks (see ROCK MAGNETISM), the North Pole is presumed to have been somewhere to the south of the present one. Evidence of glaciers is almost entirely absent in the period, except for the mountain glaciers, which might have existed in the southern part of the then-rising, high Andean orogenic system. So far as the available evidence indicates, the marine Late Cretaceous (Senonian) fauna of the Antarctic Peninsula (Graham Land) was essentially similar to that of nearby Chile and to New Zealand and also had some species in common with the Senonian fauna of India and Japan.

BIBLIOGRAPHY. P. ALLEN, "The Wealden Environments: Anglo-Paris Basin," *Phil. Trans. R. Soc.*, ser. B, 242:283–346 (1959), a fine example of a study of Early Cretaceous environments by stratigraphical and sedimentological analyses; R. BOWEN, "Oxygen Isotope Paleotemperature Measurements on Cretaceous Belemnoidea from Europe, India and Japan," *J. Paleont.*, 35:1077–1084 (1961), seawater temperatures during the Cretaceous Period, derived from oxygen-isotope analyses; G. COLOM, "Jurassic-Cretaceous Pelagic Sediments of the Western Mediterranean Zone and the Atlantic Area," *Micropaleont.*, 1:109–123 (1955), a Jurassic-Cretaceous paleogeographical reconstruction; L.B. KELLUM (ed.), *El sistema cretácico; un symposium sobre el cretácico en el Hemisferio Occidental y su correlación mundial* (1959), a comprehensive description of Cretaceous stratigraphy and correlation problems in various parts of the world; T. MATSUMOTO (ed.), "Age and Nature of the Circum-Pacific Orogenesis," *Tectonophysics*, vol. 4, no 4–6 (1967), 23 papers treating the tectonic activity, regional metamorphism, granitic intrusions, and volcanism of Late Jurassic to Cretaceous time in the circum-Pacific region; D.P. NAIDIN, "On the Paleogeography of the Russian Platform during the Upper Cretaceous Epoch," *Stockh. Contr. Geol.*, vol. 3, no. 6 (1959), a concise account of the paleogeography of this area and its changes through time; W.P. POPENOE, R.W. IMLAY, and M.A. MURPHY, "Correlation of the Cretaceous Formations of the Pacific Coast (United States and Northwestern Mexico)," *Bull. Geol. Soc. Am.*, 7:1491–1540 (1960), a good correlation chart, with annotations and bibliography, that shows an intimate relation between Cretaceous sequences in the Pacific Coast region and those of Japan and Alaska; J.B. REESIDE, JR., "Paleoecology of the Cretaceous Seas of the Western Interior of the United States," *Mem. Geol. Soc. Am.*, 67:505–542 (1957), a summary account of the changes in paleogeography on the basis of stratigraphic correlation and lithofacies and biofacies analyses; A.P. VINOGRADOV (ed.), *Atlas of the Lithological-Paleogeographical Maps of the USSR*, 3 vol. (1968), modern compilation of a series of maps that presents the Mesozoic history of this extensive region.

(T.M.)

Crete

The fifth-largest island in the Mediterranean and the largest of the many islands that form part of modern Greece, Crete (Kríti) is officially merely an administrative subdivision, but its extraordinary history has earned the island a status accorded many independent entities with areas much larger than its 3,189 square miles (8,260 square kilometres) and population of about 500,000. The island is relatively long and narrow, stretching for some 152 miles on its east–west axis and varying from 35 to 7½ miles in width.

Lying on Europe's southern fringe, Crete is halfway between Asia Minor and mainland Greece and is twice as far from Libya and Egypt; it also helps enclose the Aegean Sea, a geographical factor that has had considerable influence on its history and culture. Crete's political and economic affairs may be domestically linked with Greece, but the island is an international archaeological and tourist attraction. Crete, moreover, has survived so many challenges to its individuality that there will probably always be people who consider themselves first and foremost Cretans. (For related historical information, see AEGEAN CIVILIZATIONS; see also GREECE; AEGEAN SEA; MEDITERRANEAN SEA.)

History. Crete's history is often a part of the eastern Mediterranean's, yet it can boast distinct moments of its own. There is no evidence that man arrived on the island before 6000–5000 BC, and the first inhabitants undoubtedly came from somewhere in Asia Minor or the Levant (possibly from Egypt or Libya). They, their descendants, and subsequent groups of migrants introduced the full range of Neolithic culture—stone tools, cultivated plants, domesticated animals, weaving, pottery, houses, and, by about 3000 BC, copperworking. Whatever the various origins of these peoples, their fusion with the Mediterranean environment produced a Bronze Age culture, which is called the Minoan civilization after the island's legendary ruler Minos. The first centuries (2600–2000 BC, the Early Minoan, or Pre-Palace [Prepalatial] Period) produced nothing more spectacular than fine stone-carved vases and circular vaulted tombs. But, about 2000 BC, "palaces" began to be built on the sites of Knossos, Phaestos, and Mallia, inaugurating the Middle Minoan, or Protopalatial Period. Economic, political, and social organization began to flourish, with increased trade in the eastern Mediterranean, while stone carving, goldwork, jewelry, and pottery demonstrated aesthetic progress.

Minoan Bronze Age civilization

About 1700 BC, one of Crete's periodic earthquakes destroyed parts of the three major palaces, but there was no break in the continuity of Minoan culture. The palaces were reconstructed and even enlarged, introducing the Middle Minoan III, or New Palace (Neopalatial) Period. These ambitious complexes, with a medley of sculpture, fresco painting, pottery, and metalwork, are still visible today. A rich ceremonial life included snake goddesses and bull-leaping. The Minoans' ships, meanwhile, ranged even farther, possibly as far west as Spain; but whatever power Cretan rulers exercised in those areas was economic. Indeed, since the same Linear B script (as

philologists call this early Greek writing) recording the same Greek language was more widespread at Achaean-Mycenaean sites on the Greek mainland than at Minoan sites, it has been conceded that, by about 1500 BC (Late Minoan I), Mycenaean Greeks had assumed an influential, perhaps dominant role in Minoan affairs.

Then, about 1450 BC, Knossos and many other centres suffered another earthquake, possibly related to the catastrophic explosion at Thíra, the volcanic island north of Crete. This ushered in the Late Minoan II, or Post-Palace (Postpalatial) Period, completing the Mycenaean Greek ascendancy in Mediterranean commerce. Minoan civilization did not become definitely stagnant, however, until the Iron Age, which commenced about 1200 BC. Eventually the Dorians, another Greek-speaking people, moved in and organized Crete, while some Minoans retreated into the mountains, to become known later as Eteocretans—"true Cretans."

Post-Minoan developments

Crete still played a part in the transfer of various cultural forms from the Near East to Greece: a Cretan variation of the Orientalizing phase of Aegean art is known as Daedalic (about 700–600 BC). During Athens' heyday, Crete fascinated Greeks as a source of myths, legends, and laws. Eventually, the Romans appeared and by 67 BC had completed their conquest of Crete, converting it into Cyrenica, a province linked with North Africa; after their empire had been divided, Crete passed to Byzantium (the Eastern Roman Empire) in AD 395. Christianity, traditionally introduced by St. Paul, who was driven ashore on Crete around AD 47, gathered momentum under the Apostle's appointee, Bishop Titus, but the island later subsided into the Dark Ages. After 824, a group of Arabs controlled at least parts of Crete, but Christianity triumphed in 961 under Nicephorus Phocas, subsequently a Byzantine emperor.

In the aftermath of the Fourth Crusade, which turned aside to sack Constantinople, Crete was sold to the Venetians in 1204; they called both the island and its main city Candia and fitted Crete into their commercial empire. Native Cretans, however, never abandoned the Orthodox religion, the Greek language, and their popular lore. In 1648 the Ottoman Turks, already possessing parts of Crete, began attempts to take Candia. After one of the longest sieges in history, Candia, and with it Crete, fell to the Turks in 1669. With the economy stagnant and many Cretans perforce nominal Muslims, native culture nevertheless survived. But uprisings were always frustrated, including the one that accompanied the Greek revolution of 1821 and another of 1866, which involved an explosion and massacre at Arkadhi, a monastery in central Crete, the symbol ever since of the island's motto, "Freedom or Death!" The Turks were finally ejected in 1898, and the island was granted autonomous status under a high commissioner, Prince George, the younger son of the King of Greece. But nothing short of union with Greece would satisfy many Cretans. Among them was Eleuthérios Venizélos, the tempestuous, Cretan-born politician who eventually forced Prince George out; and in 1913, as Greek premier, he presided over the official union of Crete with Greece.

Crete's moment in World War II

Since then, Crete has shared most of Greece's history. Nevertheless, a unique moment came in 1941 during World War II when the Greek government, along with British, Commonwealth, and some Greek troops, were forced by the advancing Germans to retreat from mainland Greece to Crete. Shortly thereafter, on May 20, the Germans launched history's first—and still the only successful—purely airborne invasion, putting down all organized resistance within ten days. Most of the Allied forces were evacuated from the southern coast, but the Cretans were left to another occupation till the last German troops surrendered in May 1945. Postwar Crete made a slow recovery. Since the 1950s it has benefitted from growing international commerce, though its basic social and political patterns yield less readily to change.

The landscape. *Natural topography.* Crete is dominated by harsh mountains rising out of the sea, stark evidence of its geological origins, for the island is a remnant of a block thrown up in Tertiary times (2,500,000 to 65,000,000 years ago). Much of Crete's 650 miles of rocky coastline slopes down from the major mountains of Crete's east–west axis, a spine that breaks naturally into four main groups: the westernmost Lévka Óri (White Mountains); the central Ídhi (or Psilorítis) Mountains, with Crete's highest point, the summit of Mt. Ídhi, Stavros, 8,058 feet (2,456 metres) high; the east central Dhíkti (or Lasíthi) Mountains; and the far eastern Thriftí Mountains. Another range, the Asterousia (or Kófinos) Mountains, runs along the south central coast between the Mesará Plain and the Libyan Sea (Libikon Pélagos). The more gradual slope of the northern coast provides several natural harbours as well as coastal plains, where major cities have grown up: Khaniá, Réthimnon, and Iráklion (the Candia of history). The major flatland, however, is the Mesará Plain, which extends along the south central region for about 18 miles, averaging three miles wide. Also, on its northern side, Crete has several upland basins, including the Omalos in the Lévka Óri and the Nidha in the Ídhi Mountains; the most notable is the Lasíthi Plain, an almost perfect stadium, measuring about 50 square miles ringed by mountains. Cretans have lived mostly on the edges of these plains to avoid seasonal flooding, while taking maximum advantage of the arable land.

Soils and drainage. Nomadic grazing occurs on 48 percent of Crete's total area, while 20 percent of the land is entirely unproductive. Over the centuries the islanders have so stripped the once thickly wooded slopes that the earth has eroded, leaving largely bare limestone. As a result, the surface is so porous and honeycombed that much of the water goes underground. This accounts for the springs and the many merely seasonal watercourses. There are only about six rivers on Crete, including the Platanías (near Khaniá), the Milopótamos (north central region), the Anapodháris (south central), and the Yeropótamos (in the Mesará).

Limited surface-water resources

The soil is largely rocky, with little alluvium or loam; it lacks nitrogen and phosphorus, though potassium and calcium carbonate are plentiful. Crete also is marked with many fissures and ravines, such as the gorge of Samaria, which extends about 11 miles inland from the southwestern coast. There are also caves, the source of many mythological and historical episodes. One freshwater lake, Kournás, lies west of Réthimnon. The island experiences occasional earth tremors, but only two serious earthquakes have occurred in the last 100 years.

Climate. Crete's climate varies between temperate and tropical. The mountains are colder and wetter than the lowlands, and snow remains throughout winter above 1,600 feet and is almost permanent on the highest peaks. Annual rainfall averages about 25 inches, mostly from October to March. In the hot, dry summer, prevailing northeasterly sea breezes (the *meltemi*, or etesian winds) keep the coastal regions pleasant but cannot break the drought of the interior. A dusty haze often pervades the atmosphere, with occasional sirocco winds blasting in from the Sahara. But by late October eastward-moving cyclones passing to the north or south bring more variable and tempestuous winds, with rain at sea level and snow in the mountains.

Frost is practically unknown on the coast; the south shore enjoys a particularly mild winter, with an average January temperature at Iráklion of 54° F (12° C). In summer, Iráklion's daily maximum averages 84° F (29° C).

Vegetation and animal life. Despite its demanding climate and man's abuse of the land, Crete supports a varied vegetation. Characteristic Mediterranean scrub (maquis or garigue), unproductive but flowery, dominates the landscape. Something is usually blooming, be it phlomis (a type of mint), thorny broom, spurges (cactus-like plants), asphodel, thyme, heath (a type of evergreen shrub), burnet (a kind of herb), or rockrose (a local shrub). Generally, Crete's flora is similar to that of the Peloponnese and Asia Minor, but the island is also noted for several native species, including *Acer creticum*, a dense, spiny maple shrub, and *Berberis cretica*, a barberry (a red-berried shrub). The quince (*Cydonia ob-*

longa) is said to be indigenous to Crete. Perhaps the most prized local plant species is the Cretan dittany; related to marjoram, it is a small, perennial, gray-green plant with ruddy-pink flowers and clings to rocky cliffs and sparse patches of soil.

Crete once nurtured valued cypress and cedar forests, which, as late as the 16th century, supplied wood for the Venetian fleet. Now only small clumps of wild cypress survive on the Lévka Óri, and only about 2 percent of the total land area bears such trees as Aleppo pine, ilex, holm oak, chestnut, and plane trees. Olives, carobs (evergreens whose pods have a sweetish pulp), and orange trees are cultivated, along with several curiosities such as almond trees, wild palm, bananas, and the black mulberry.

The Cretan wild goat

The most spectacular of Crete's fauna is known as the *agrímí* ("the wild one"), a wild goat (*Capra aegagrus*) related to the ibexes that range across Asia Minor and down into Iran and Pakistan. This wild goat was hunted down to the 20th century but became nearly extinct and was confined to the gorge of Samaria; now three offshore islets serve as natural preserves. The smaller animals and birds are the ones usual in such an environment. There are no poisonous snakes, credit for which is traditionally awarded to St. Titus. The surrounding sea yields shellfish, squid, sponges, and edible fishes such as red mullet, surprisingly underexploited.

Traditional regions. Crete has contained varied lifestyles. Réthimnon is the "intellectuals' city"; Anóyia retains distinctive popular traditions. Perhaps the most special "pocket" of all, Sfakákia, the southwesterly region isolated by mountains, has a particular heritage of independence, boasting men taller and stronger than other Cretans. In any case, Crete divides traditionally into two regions: coast and mountain. This split is accentuated by the fact that coastal dwellers live mostly along the northern shore. The basic modern difference, however, is between those dependent on the land and those engaged in urban pursuits.

Land use. Only about 30 percent of Crete's total area can be actively cultivated. There are some regional specialties—citrus fruits in the west, carob trees in the east, for instance—but farming patterns are similar throughout the island: small patches of land, cultivated with little use of machinery, and women helping in the fields, some land still irrigated by hand-operated wells, though draft animals, gasoline pumps, and windmills are taking over that task. An exception is the Mesará Plain; it is relatively well watered and one of the few parts of Crete that lends itself to modern agricultural practices, where large machinery is increasingly employed. Two deeply embedded traditions have hindered agricultural development: one is the Cretan preference for living together in villages, which incurs long journeys to the fields; the other is the splitting up of land in legacies, so that tiny lots are walled in stone and individual ownership of the small plots is commonly quite scattered.

Urban settlement. The eight or nine largest cities on the northern coast account for a third of the island's population, yet most of them are overgrown villages. Iráklion and Khaniá are the two exceptions, with almost one-quarter of Crete's population. Iráklion has become

City life in Iráklion

quite cosmopolitan, with tourists, hotels, restaurants, shops, and similar enterprises. Its broad streets, cafés, cinemas, museum, and market make it most lively, while the port is also bustling. Modern problems such as traffic congestion, pollution, and urban blight are also appearing on the heels of haphazard civic growth. Khaniá, half the size of Iráklion, is correspondingly more provincial.

The people. Although claims have been made that Cretans are taller than other Greeks or have different head measurements, due to Dorian or other heritages, such differences are unverifiable. All Cretans speak Greek, albeit with variations of dialect, especially in rural or mountain areas.

Religious affiliations. Virtually all Cretans belong to a special branch of the Orthodox Church, directly responsible to the patriarchate of Constantinople. The archbishop of Crete has his seat in Iráklion, and most Cretans are devout. There are a few Roman Catholics, but Crete's old Jewish population moved out long ago, as did the sizable Muslim Turkish community in the early 1920s as part of the population exchange between Greece and Turkey, when thousands of Greeks from Asia Minor were resettled on Crete.

Demographic trends. More and more Cretans share in such modern amenities as improved medical facilities, agricultural aids, communications and transport, and more are born with the chance to live longer, healthier lives. Crete's small population nevertheless is larger than the island can support. Along with a gradual flow of population to the larger towns, working class Cretans are leaving for mainland Greece or Europe, educated, ambitious youth are going to Athens or abroad, and families are emigrating to such traditional destinations as the United States and Australia.

The economy. *Agriculture and natural resources.* Most of the economy rests on agriculture: Crete is one of Greece's leading regions in the production of olives and

Major crops

olive oil, grapes (including seedless sultanas, raisins, and wine), citrus fruits, and the carob, or locust, bean, all exported mainly to Greece. For itself, Crete grows fresh vegetables, fruits, nuts (almonds and acorns), and some grains (barley and oats, but insufficient wheat). It also raises sheep and goats for meat, cheeses, wool, and hides. Fine as many of these products are, none carries much weight in modern commerce.

Cretans also mine talc, lignite, and gypsum and even a little copper and iron. Deposits of lead, manganese, zinc, sulfur, gold, silver, tungsten, platinum, emery, graphite, tin, and magnetite have been discovered, but none in workable quantities. Crete is still further limited in its energy resources, for it has to import all its fuels.

Commerce and industry. Industry is largely confined to food processing (olive and grape presses), building materials (stone quarries and building blocks), and a few ceramics, textiles, soap, leather, and steel-tool enterprises. Most concerns are still run by their owners, employ only a few people, and are located along the northern coast. Some traditional handicrafts, such as weaving or embroidery, are done at home. The harbours, construction trades, transport, public services, and tourism also provide employment.

Crete has to import all but the most basic items, even building materials, fertilizers, and food. Increasingly consumer outlooks need higher incomes, but inflowing foreign and mainland capital must fight conservative ways, family-run enterprises, consolidation rather than expansion, and age-old distribution patterns. Tourism, however, has brought changes, with more large employers and some organization of labour.

Transport. Crete has no railways and no navigable rivers, but its road network is good. Private vehicles and commercial trucks are multiplying, but most Cretans still travel by bus. Olympic Airways flights link Iráklion with Athens and with Rhodes, and Khaniá with Athens; occasional charter flights also serve Iráklion. Small cargo ships and caïques (light skiffs) ply between Crete and other islands or ports, and there are almost daily ferries between Khaniá and Iráklion and Athens-Piraiévs (and one ship weekly to Rhodes and to Thíra); large merchantmen and liners frequently call, but mainly Crete remains tied to Athens.

Administration. Crete, itself an administrative region of Greece, consists of four prefectures (*nomoi*)—Khaniá, Réthimnon, Iráklion, and Lasíthi. Each has a somewhat powerless prefect (*nomarch*), appointed via Athens and responsible to the minister of the interior. Khaniá, the administrative capital, housing various government offices and the island's highest court of appeal, exercises little power. Crete is subdivided into 570 communities, administered by elected mayors or presidents and small councils with little authority. This want of power stems

Nature of Cretan politics

more from the general trend of state centralization than from any particular Athens government. Cretans—when given the choice—have taken the liberal, republican, antimonarchist side in Greek political life. In any case, physical remoteness from Athens and traditional skepti-

cism about governments have always rendered local problems more pressing than mainland matters.

Notwithstanding, Crete has long provided disproportionately large numbers of government personnel, a result of the job shortage on the island. Military installations include a Greek army training school outside Iráklion and a large NATO naval base and a NATO anti-ballistic-missile training school around Soúdha Bay, the large harbour near Khaniá. Educationally, Crete has nothing above lycée (gymnasium) or vocational school level, though every community provides some schooling. Similarly, health and welfare services deteriorate in remote districts, but doctors with some foreign study behind them are scattered throughout the island. With the only government representative in the villages a policeman, in most crises Cretans must rely on their own meagre family resources.

Cultural life. *The arts.* Perhaps the island's greatest resource is its popular culture. Crete also has attained respectable artistic "heights"—as during something of a renaissance that flowered from 1560 to 1660, when poems, plays, and paintings with a peculiarly Cretan tang were produced by Greek settlers from Constantinople or by Venetian-Cretans. Vitzéntzos Kornáros, born near Sitia, is credited with *Erotókritos*, a romantic epic that has become a national poem for some Greeks, and *The Sacrifice of Abraham*, a sturdy drama. Painting drew on the Byzantine tradition of Crete's churches. Some of the best artists went to paint in the monasteries of Mt. Áthos and the Meteora in Greece, although one late-16th-century artist who chose to work on Crete was Michael Damaskinos. The tradition's finest product was Doménikos Theotokópoulos, born on Crete in 1541. He studied and painted there until perhaps 1566, going on to Venice, Rome, Toledo, and universal fame as El Greco.

Notable writers

After a pause of some two centuries, Crete produced three notable writers, John Kondylakis, Pandelis Prevelakis, and the internationally known Nikos Kazantzakis. Two of Kazantzakis' finest works are specifically Cretan: *Freedom or Death*, a novel about one of the island's 19th-century revolts, and *Zorba the Greek*, whose hero has become the epitome of the Cretan spirit.

The folk element. Cretans are less attached to their ancient past than to Byzantine–Orthodox culture and their popular traditions: icons and sacred ground, name days and feast days, lore about spirits and vampires, legends about the centuries of struggle against foreign occupiers, *mantinades* (rhymed couplets in song), and *palikhari* (brigand-chieftains in folklore). Occasionally, high and popular culture join, as with the *Erotókritos*, which is still recited, or with *The Song of Daskaloyiannis*, a late 18th-century poem about another revolt. But on Crete, as elsewhere, this popular culture is slowly disappearing as the elderly die off. Some men still dress in boots, baggy pantaloons, sash, and embroidered vest and have a proverb or superstition for everything. Old motifs are embroidered into rugs, shoulder bags, or bridal linens, and there have been conscious attempts at folklore revivals. But the true Cretan spirit must survive naturally, in village festivals or in cafés, where people sing and dance to old tunes spontaneously.

The outlook. Crete today confronts the dilemmas of many underdeveloped, insular societies. Freed from foreign threats, it now must deal with itself, though economic advancement may seem indistinct from uniformity. Its immediate prospects rest in its archaeological treasures and its beaches and in their imaginative and sensitive exploitation. Alongside these assets, wiser land use and economic development rooted in native materials, skills, and patterns could resolve the problem. With accompanying restraint, development's ill effects may be avoided and Crete may well emerge a robust yet modern society.

BIBLIOGRAPHY. The most accessible general surveys of Crete are JOHN S. BOWMAN, *Crete*, rev. and enl. ed. (1969); R.W. HUTCHINSON, *Prehistoric Crete* (1962); and RAYMOND MATTON, *La Crète au cours des siècles* (1957). For the Minoan world, see ARTHUR J. EVANS, *The Palace of Minos*, 6 vol. (1921–36), is still the seminal (but unwieldy) work. More accessible are R.W. HUTCHINSON, *Prehistoric Crete* (1962); J.D.S. PENDLEBURY, *The Archaeology of Crete* (1965); NICHOLAS PLATON, *Crete* (Eng. trans. 1966); and for fine photographs: S. MARINATOS and M. HIRMER, *Krētē kai Mykēnaikē Hellas* (1959; Eng. trans., *Crete and Mycenae*, 1960); and LEONARD VON MATT *et al.*, *Das antike Kreta* (1967; Eng. trans., *Ancient Crete*, 1968). For post-Minoan history, the Dorian world is discussed in R.F. WILLETTS, *Ancient Crete: A Social History from Early Times Until the Roman Occupation* (1965); the Venetian period in WILLIAM MILLER, *Essays on the Latin Orient* (1921); GIUSEPPE GEROLA, *Monumenti veneti nell'isola di Creta*, 4 vol. (1905–32); and DENO J. GEANAKOPLOS, *Greek Scholars in Venice* (1962). For the struggle for independence and union, see EDWARD S. FORSTER, *A Short History of Modern Greece, 1821–1956*, 3rd ed. rev. by DOUGLAS DAKIN (1958); PRINCE GEORGE OF GREECE, *The Cretan Drama*, ed. by A.A. PALLIS (1959). For World War II, see ALAN CLARK, *The Fall of Crete* (1962); and GEORGE PSYCHOUNDAKIS, *The Cretan Runner* (Eng. trans. 1955). The only readily available general survey of the geography, economy, and sociology of contemporary Crete is LELAND G. ALLBAUGH, *Crete: A Case Study of an Underdeveloped Area* (1953), although dating from 1948, many generalizations still hold true. For the popular culture of modern Crete, the best introduction is MICHAEL L. SMITH, *The Great Island* (1965). Translations of Crete's "renaissance" plays are in F.H. MARSHALL and JOHN MAVROGORDATO, *Three Cretan Plays* (1929). Travels on Crete include the classic, ROBERT PASHLEY, *Travels in Crete* (1837); and a modern experience, XAN FIELDING, *The Stronghold* (1953). JOHN S. BOWMAN, *op. cit.*; and STERGHIOS SPANAKIS, *Crete*, 2 vol. (1968), are guides to Crete.

(J.S.Bo.)

Cribbage

Cribbage is a card game in which the object is to form counting combinations that traditionally are scored by moving pegs on a special Cribbage board, the dealer scoring an extra hand, the crib, formed of discards. The appeal of the game, usually played by two but with a popular variant played by four or, occasionally, by three, is evident from two facts: few changes have been made in the original rules, and it remains one of the most popular of all card games. In Great Britain, during the four years prior to 1970, the "Card Corner" in *News of the World* had more requests for information on Cribbage than for any other game. A 1970 appraisal indicated little overlap between devotees of Bridge and those of Cribbage. In the United States, Cribbage is played by more than 10,000,000 people, principally across the northern states, from New England to the Pacific, and the game has remained popular in Canada as well.

The game of Cribbage (earlier spelled Cribbidge) was invented by the English poet Sir John Suckling (1609–42). Although Cribbage quite clearly developed from Noddy, an older game for which a special scoring board also was used, it appears to be the only existing game in its family. Cribbage would quite likely have become the most popular of all two-hand card games if so many descriptions had not called the Cribbage board indispensable, which it is not.

Almost the only big change from the original rules is that in modern two-hand Cribbage each player is dealt six cards instead of five, as played originally.

Scoring. Scoring is traditionally called pegging, because it usually is done by moving pegs on a scoring device, the Cribbage board. This Cribbage board is essentially a tablet with 60 counting holes (in two rows of 30) for each player, plus one game hole for each, and often extra holes for holding pegs when not in play and for keeping track of games won. Game is 121 (twice around the board plus 1 for the game hole) or 61 in the less frequently played game of Once Around. Each player has two pegs, and each scoring point is marked by jumping the rearmost peg ahead of the other (thus showing at a glance the number of points scored on a move as well as the total). Scores must be pegged in order (see below *The showing*), because the first player to reach 121 (or 61) or, in some games, to pass it is the winner. Emphasis on the board as a scoring device created the idea that the game could not be played without it, but the score can be kept with pencil and paper or with chips or other counters; indeed, keeping score by discarding counters

Cribbage board

(each player starting with 121 or 61) is so efficient and simple a method that the enduring primacy of the board is difficult to understand.

The cut and deal. The customary deck of 52 cards is used, the cards ranking from king (high) to ace (low). Face cards and tens count 10 each; other cards count their index value (number of pips). The player cutting low card deals first, the deal alternating with each hand. The dealer deals six cards, alternately, to the nondealer and to himself. Each player then discards two cards, facedown, to form the crib. In discarding to the crib, since it scores for the dealer, the nondealer tries to lay away "balking" cards, those least likely to create scoring combinations. After the discard, the undealt remainder of the pack is cut by the nondealer; the top card of the lower packet is turned faceup on top of the reunited deck and becomes the starter. If the starter is a jack, dealer immediately pegs (scores) 2, called "2 for his heels." If the starter is any other card, the jack of that suit—formerly called "knave noddy," an unmistakable link with the earlier game—is worth 1 point to the holder for "his nobs" but is not scored until later (see below *The showing*). This is followed by the two stages of scoring, the play and the showing.

The play. The nondealer begins the play by laying faceup before him any card from his hand, announcing its counting value. Dealer then plays a card (each adds cards to his own pile, so that his original hand may be counted later in the showing) and announces the total of the two cards. Play continues alternately, each player announcing the new total, until the total reaches 31, or until one player cannot play without increasing the total beyond 31. If either player cannot add a card without exceeding 31, his opponent must play any card(s) in his hand that may be added without exceeding 31. The last to play in each sequence scores a "go"—2 points if he reaches exactly 31, or 1 for any lesser total. After a go, count begins again at zero.

Playing for go

In addition to go, the object is to peg for certain combinations of cards played consecutively. These combinations score whether the cards are played in strict alternation or in succession by one player when his opponent cannot play. The score in every case is pegged by the player whose card completes the combination. Any player who can add to a combination, providing there has been no intervening card, can score the value of the new combination. Combinations are scored for playing a card that makes the count exactly 15 (score 2); for playing cards of the same rank to make a pair (2), three of a kind (6), or four of a kind (12); and for playing a third or later card to form a run, or sequence, regardless of suits and regardless of the order in which the cards are played (1 for each card in the run).

The showing. The next stage of scoring is the showing. After all four cards are played, the values in each hand are counted—the nondealer's hand first, then the dealer's hand, then the crib. The starter counts as a fifth card in each of the three hands. Every combination of two or more cards totalling 15 scores 2; each pair, 2; every sequence of three or more cards, 1 for each card in the sequence; four cards of the same suit, 4, or 5 if of the same suit as the starter (but only a five-card flush matching the starter counts in the crib); and his nobs (jack of the same suit as the starter), 1. Every possible different grouping of cards in the hand, plus starter, counts separately, except that a sequence of four or five cards may be counted only once, and not as two or more separate sequences of three.

As indicated above, the order of scoring on each hand is important and is as follows: (1) scoring of starter, if it is a jack, (2) scoring in play for various combinations, (3) scoring in play for go, (4) scoring of nondealer's hand, (5) scoring of dealer's hand, and (6) scoring of crib. When either or both players approach a score of 121 (or 61), whose turn it is to score becomes important. The game ends immediately if either player is able to count out in the play or the showing. If nondealer is able to count out in the showing, it does not matter if the dealer, with or without counting his crib, could have scored a higher total. The loser scores only what he has already pegged before his opponent counts out, and if he has not already counted at least 61 (or 31), he is "lurched" ("left in the lurch") and, if the play is for stakes, loses doubly. (As sometimes played, the winner must be able to count out to exactly 121, just as, in playing for a go, he tries to reach 31 exactly. Thus, for example, if a player's score is 120, he can count out only if he can score exactly 1 point, as for his nobs or for go.) Some play that, if a player fails to claim his full score on any turn, his opponent may call out "muggins" and score for himself any points overlooked.

Counting out

After each player has played all four of his cards, and the showing has been completed, the cards are put back in the deck and shuffled and dealt as before.

Variants. *Five-card Cribbage.* This was the original game. Each player discards two cards into the crib, remaining with only three, plus starter. At the beginning of the initial hand nondealer pegs 3 to offset dealer's advantage. Game is 61.

Four-hand. Play is in partnerships of two on a side, partners seated across the table from each other. The dealer gives each player five cards; each discards only one into the crib. The score is usually slightly less in the showing, but the average per side is about 9 points in the play. Game is always 121.

Three-hand. Each player is dealt five cards and discards one into the crib, and a single card is dealt blind to complete the crib, which belongs to the dealer. Each player scores for himself. Eldest hand (the one to the left of the dealer) shows first.

BIBLIOGRAPHY. CHARLES COTTON, *The Compleat Gamester* (1674); R.L. FREY (ed.), *According to Hoyle* (1965); CAVENDISH, *Pocket Guide to Cribbage* (1925).

(R.L.Fr.)

Cricket

Cricket, generally considered to be England's national summer sport, is a game of skill played with bat and ball between two teams of 11 players each on a large field. It has devotees throughout the world, particularly in the British Isles and the Commonwealth.

For a list of major all-time batting, bowling, and fielding records in cricket see SPORTING RECORD in the *Ready Reference and Index.*

THE GAME

Grounds and equipment. *The ground.* Cricket grounds vary in size from great arenas such as Lord's in London (5½ acres) and one in Melbourne that is even larger (9¼ acres) to village greens and small meadows. Level turf of fine texture is the ideal surface. The limits of the circular or oval playing area are usually marked by a boundary line or fence.

Wicket and creases. A wicket consists of three stumps or stakes, each 28 inches (71 centimetres) high and of equal thickness (about 1¼ inches [3.2 centimetres] in diameter), stuck into the ground and so spaced that the ball cannot pass between them. Two pieces of wood called bails, each 4⅜ inches (11 centimetres) long, lie in grooves on the tops of the stumps. The bails do not extend beyond the stumps and do not project more than half an inch above them. The whole wicket is nine inches (23 centimetres) in width. There are two of these wickets, which a batsman defends and a bowler attacks, and they are approximately in the centre of the ground, facing one another at each end of the pitch (also sometimes called the wicket), which is an area 22 yards long and 10 feet wide (20 by three metres) between the wickets.

Lines of whitewash, known as creases, are marked on the ground at each wicket: the bowling crease is a line drawn through the stumps and extending four feet four inches (1.3 metres) on either side of the centre stump; the return crease is a line at each end of and at right angles to the bowling crease, extending a short distance behind the wicket; and the popping crease is a line parallel with the bowling crease and four feet in front of it. The bowling and return creases mark the area within which the bowler's rear foot must be grounded in delivering the

Bowling, return, and popping creases

ball; the popping crease, which is 62 feet (19 metres) from the opposing bowling crease, demarks the batsman's ground.

Bat and ball. The blade of the paddle-shaped bat is made of willow and must not be broader than 4¼ inches (11 centimetres). The length of the bat, including the handle, must not exceed 38 inches (97 centimetres). The ball which has a core of cork built up with string, is encased in polished red leather, the halves of which are sewn together with a raised seam. Slightly smaller, harder, and heavier than a baseball, it must weigh between 5½ and 5¾ ounces (156 and 163 grams) and be between 8 13/16 and nine inches (22 and 23 centimetres) in circumference.

Dress. Players usually wear white flannel trousers and shirt, white boots (shoes) of buck or canvas, a white woollen sweater, often trimmed with club colours, and club caps of kaleidoscopic variety. The batsman wears white pads, or leg guards, a body protector, and articulated batting gloves, usually of thick tubular rubber or of leather stuffed with hair. The wicketkeeper also wears pads and reinforced gauntlets (cricketers, unlike baseball players, do not wear gloves when fielding).

Language of cricket. Cricket has evolved a vocabulary of its own. Many of the more important terms are explained in other sections of this article (see below *Methods of dismissal*). Some of the variant or multiple meanings of certain terms and some additional terms and phrases follow.

Bowled for a duck. Bowled for a duck is to be dismissed (put out) without scoring.

Capped. A player is said to be capped when he is selected to play for a representative team.

Century. A century is 100 runs scored by one batsman in a single innings.

Follow on. To follow on is to have a team bat second innings immediately after first innings (*i.e.*, out of turn) after scoring less than their opponents in first innings by a specific number of runs (75 in a one-day match; 100 in a two-day match; 150 in a match of three days; 200 in one of five or six days).

Hit a six. To hit a six is to hit a ball over the boundary without its first touching the ground to score six runs.

Innings. Innings is either (1) a turn of a batsman to bat, (2) a turn of a team to bat, or (3) in recording a result, when one team still has a turn to bat but has scored more runs than the opposition, which already has completed its two innings, the result is recorded as an innings (to go) and so many runs (thus, for example, Hambledon beat "a representative England eleven by an innings and 168 runs").

Pitch. A pitch is either (1) an area extending five feet (1.5 metres) on either side of the centre line between the wickets, (2) the distance between wicket and wicket, or (3) the impact of the bowled ball on the ground (pitch).

Representative team. A representative team is a team of players selected, on the basis of ability, to represent a county, a country, etc. (as distinguished from a club team or a minor team).

Sticky wicket. A sticky wicket is a pitch drying after rain, when soft turf makes it difficult for a batsman.

Wicket. A wicket is either (1) the goal, consisting of three stumps (stakes) on top of which lie two sticks (bails), which the bowler attempts to hit and the batsman attempts to defend, (2) the area between two sets of stumps, synonymous with pitch, (3) a turn to bat (*i.e.*, the partnership between two batsmen for the first wicket, second wicket, etc.), or (4) in scoring, when a side is batting last, the number of batsmen still to be dismissed (put out) when the opponent's score is passed.

Conduct of the game. One player on each team acts as captain. There are two umpires, one standing at the bowler's wicket, the other at square leg near the batsman's wicket (see Figure 1), to control the game according to the laws and two scorers to record its progress. The object of the game is for one side to score more runs than the other.

At the start of a match, the captain who wins the toss of a coin decides whether his own or the other side shall take first innings (always plural); *i.e.*, proceed successively as batsmen—the first two as a pair together—to the wicket and try to make as many runs as possible against the bowling and fielding of their opponents. An innings is completed when ten batsmen have been dismissed (the remaining batsman, having no partner, is declared "not out"). The captain of the batting side may, however, declare his innings closed before all ten men are out if his side is far ahead on runs and he wants to bowl out the opponents before the time limit expires and the game is called a draw. Results are recorded by the margin of runs, or, if the side batting last passes the other side's total before all their batsmen have been dismissed, by the number of their wickets (*i.e.*, batsmen still to be dismissed) outstanding.

Scoring innings and matches

Matches are decided either by the number of runs scored in one innings each (usually one-day matches), or on the aggregate of runs made by each side in two innings. International (Test) matches last five or six days (30 playing hours), other first-class matches from two to four days, and the bulk of club, school, and village matches one day.

How the game is played. The nonbatting side takes up positions in the field. One man is the bowler, another is the wicketkeeper, and the remaining nine are positioned as the captain or the bowler directs (see Figure 1). The first batsman (the striker) guards his wicket by standing with at least one foot behind the popping crease. His partner stands behind the popping crease at the bowler's end. The bowler tries to hit the batsman's wicket or to dismiss him in other ways (see below *Methods of dismissal*).

Runs. The batsman tries to keep the bowler from hitting the wicket, while also trying to hit the ball sufficiently hard to score a run; *i.e.*, enable him to run to the other end of the pitch before any fieldsman can pick up the ball and throw it to either wicket to knock off the bails. If the wicket is broken, either by a thrown ball or by the wicketkeeper or bowler with ball in hand, before either batsman is in his ground, the batsman is dismissed. The striker does not have to run after he has hit the ball nor does it count in any way if he misses the ball. But if he gets a good hit and thinks he can score a run, he races for the opposite wicket and his partner (the nonstriker) runs toward him. When each has made good his ground at the opposite end, one run is recorded to the striker; if there is time, they will run back for a second or more runs, crossing again. If an even number of runs is scored, the striker will receive the next ball; if an odd number, then the nonstriker will be at the wicket opposite the bowler and will face the next ball. Any runs thus made count to the batsman, otherwise they are extras. When a ball from a hit or any of the extras mentioned below goes as far as the boundary, the runners stop and four runs are added to the score. If the batsman hits the ball full pitch over the boundary (on the fly) he scores six runs.

Extras. Only runs scored from the bat count to the batsman, but to the side's score may be added the following extras: (1) byes (when a ball from the bowler passes the wicket without being touched by the bat and the batsmen are able to make good a run); (2) leg byes (when in similar circumstances the ball has touched any part of the batsman's body except his hand); (3) wides (when a ball passes out of reach of the striker); (4) no balls (improperly bowled balls; for a fair delivery the ball must be bowled, not thrown, the arm neither bent nor jerked, and in the delivery stride the bowler's front foot must land behind or covering the popping crease), off which a batsman cannot be out (except as noted under *Methods of dismissal* below) and which, apprized in time by the umpire's cry of "no ball," he may try to hit.

Byes, leg byes, wides, and no balls

Overs. When a bowler has bowled six balls (eight in Australia and South Africa), not counting wides and no balls, he has completed an over and a new over is begun by a different bowler at the opposite wicket, with a corresponding adjustment of the field. If a bowler delivers a complete over without a run being scored from the bat (even though the opponents may have scored extras by means of byes, etc.), he has achieved a maiden over, a feat of some accuracy.

Methods of dismissal. The nine ways in which a batsman or striker can be dismissed (put out) are as follows.

Bowled. The batsman is out bowled if the bowler breaks the wicket; *i.e.*, dislodges a bail with the ball.

Caught. He is out caught if a ball hit by the batsman is caught before it touches the ground.

Stumped. He is out stumped if, in playing a stroke, he is outside the popping crease (out of his ground) and the wicket is broken by the wicketkeeper with ball in hand.

Leg before wicket. Essentially, leg before wicket (lbw) is illegal interference, whether accidental or intentional, with a ball that otherwise, in the opinion of the umpire, would hit the wicket. The batsman is out lbw if he intercepts with any part of his person, except his hand, that is in line between wicket and wicket, a ball that has not first touched his bat or his hand and that has or would have pitched (hit the ground) in a straight line between the wickets or on the off side provided the ball would have hit the wicket.

Run out. Either batsman is out run out if, while the ball is in play, his wicket is broken while he is out of his ground. If the batsmen have passed each other, the one running for the wicket that is broken is out; if they have not crossed, the one running from that wicket is out.

Hit wicket. The batsman is out hit wicket if he breaks his own wicket with his bat or any part of his person while playing the ball.

Handling the ball. Either batsman is out if he touches the ball with his hands while it is in play.

Hitting the ball twice. Batsman is out if he hits the ball, except in defense of his wicket, after it has been struck or stopped by any part of his person.

Obstructing the field. Either batsman is out if he willfully obstructs the opposite side.

Strategy and technique. The disposition of the field will vary widely according to the technique of the bowler or of the batsman, the condition of the pitch, the state of the game, and the tactics determined by the captain. He may place his fieldsmen as he thinks best, and he may alter their positions, if he wishes, after each ball.

There are no foul lines in cricket so a hit in any direction is a fair ball. The objectives of the captain of the fielding side are: (1) to place his men in positions where the batsman may give a catch; *i.e.*, hit a straight drive or a fly ball to a fielder and (2) to save runs; *i.e.*, to block the path of the ball from the batsman's scoring strokes (intercept or trap grounders). The tactical possibilities for a thoughtful and ingenious captain in directing the battle of wits between his bowlers and fieldsmen and the batsmen are manifold and constitute one of the attractions of the game for player and spectator alike.

Fielding positions

The names of the generally accepted positions are shown in the drawing. As there are only 11 players on a team and two of them must be the bowler and wicketkeeper, only nine other positions can be occupied at any one time. The field is spoken of as being divided lengthwise into off and on, or leg, sides in relation to the batsmen's stance, depending upon whether he bats right or left-handed; the off side is the side facing the batsman, and the on, or leg, side is the side behind him as he stands to receive the ball.

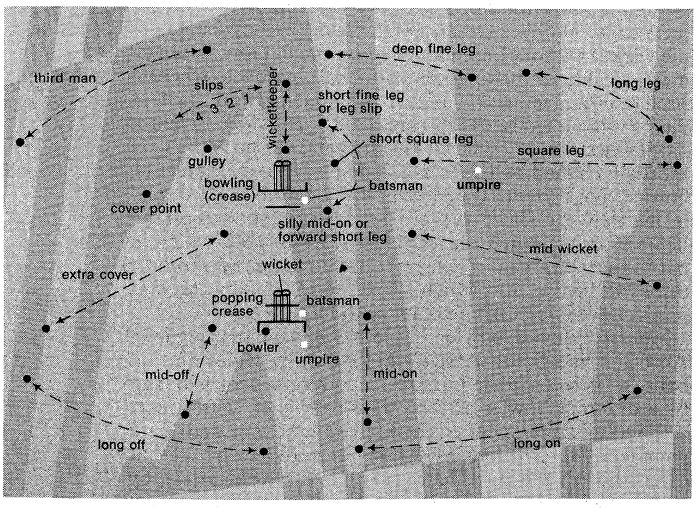


Figure 1: Location of wickets and principal playing positions on cricket field.

To sum up, the objective of the bowler is primarily to get the batsman out and only secondarily to prevent him from getting runs. The objective of the batsman is to protect his wicket and to make runs, for only runs can win a match, but to make runs he must stay in. The objective of each fielder (and of the general distribution of the field) is first, to dismiss the batsmen, and, secondly, to prevent the striker making runs. The arts of batting, bowling, and fielding are therefore a fusion of attack and defense, but ideally attack dominates.

Bowling. Bowling can be right- or left-arm. For a fair delivery the ball must be propelled, usually overhand, without bending the elbow. The bowler may run any desired number of paces as a part of his delivery, as in bowling (with the restriction, of course, that he not cross the bowling crease). The ball generally hits the ground (the pitch) before reaching the batsman, although it need not. The first requisite of a good bowler is command of length; *i.e.*, the ability to pitch (bounce) the ball on a desired spot, usually at or slightly in front of the batsman's feet, and varying with the pace of the bowler, the state of the pitch, and the reach and technique of the batsman. The second requisite is command of direction. On this foundation a bowler may elaborate with variations—finger spin, swerve, alteration of pace and flight, the path of the ball, and the manner in which it is propelled—that lend deceptiveness and uncertainty as to exactly where and how it will pitch. A good-length ball is one that causes the batsman to be uncertain whether to move forward to play his stroke or to move back. A half volley is a ball pitched so far up to the batsman that he can drive it fractionally after it has hit the ground without having to move forward. A yorker is a ball pitched on or inside the popping crease. A full pitch is a ball that the batsmen can reach before it hits the ground. A long hop is a ball short of good length or "short of a length."

Length and spin in bowling

The primary purpose of the spin in bowling is to bring the ball up from the pitch at an angle or in a direction that is difficult for the batsman to anticipate.

The two swerves are the inswinger, which moves in the air from off to leg, and the away swinger, or outswinger, which swerves from leg to off.

Batting. A batsman may hit right-handed or left-handed. Good batting is based on a straight (*i.e.*, vertical) bat with its full face presented to the ball.

Batting strokes

The chief strokes are: forward stroke, in which the batsman advances his front leg to the pitch (direction) of the ball and plays it in front of the wicket (if played with aggressive intent, this stroke becomes the drive); back stroke, in which the batsman moves his rear leg back before playing the ball; leg glance (or glide), in which the ball, when pitched in a line with or outside the batsman's body, is deflected behind the wicket on the leg side; cut, in which the batsman hits a ball on the uprise (after it has hit the ground on the off side), over and down behind the wicket or through the slips with a vigorous whiplike action of the wrists.

Fielding. The ideal fieldsman is a fast runner, with keen eyesight, quick reactions, and the ability to throw straight and far. He should be able to anticipate the batsman's strokes, to move quickly to cut off the ball in its path, and to pick it up and throw it to the stumps in one movement. He must be able to judge the flight of the ball in the air to make a safe catch. By his alertness and skillful play he can save runs and contribute to a batsman's downfall.

Wicketkeeping. The wicketkeeper is a specialist position requiring abnormally quick reactions, sharp eyesight, and courage. He is the hardest worked member of the team and must concentrate on every ball, whether standing 12 to 15 yards behind the stumps for the fast bowlers or crouching close to them for those of slow or medium pace.



Figure 2: (Left) F.E. Woolley, a left-handed batsman, at the finish of a "hook," showing perfect balance and follow-through. (Right) The perfect stance of J.B. Hobbs, a right-handed batsman. Both players set county and England records during the early decades of the 20th century.
(Left) Sport and General Press Agency, (right) Central Press Photos

CRICKET IN ENGLAND

Early history. Cricket has been played under recognized rules at least since the beginning of the 18th century. The first definite match of which there is record was played in Sussex in 1697, 11 a side, and for a stake of 50 guineas. In 1719 the "Londoners" met the "Kentish men" in what was virtually the first match between two county sides, London actually being synonymous with Middlesex.

At first the greatest enthusiasm and the most expert skill were concentrated in the southern counties near London. There, on the short turf of the open downs, cricket was discovered by "society" and transplanted to London and to the home grounds of its noble patrons. In London by far the most famous cricket centre was the Artillery Ground, Finsbury.

A feature of the play in those days was the heavy stake money and side bets that more often than not depended on big matches. The crowds were often disorderly and violently partisan.

The next stage in the game's development was marked by the rise of the Hambledon Club. On Broad-Halfpenny Down (a historic site later acquired by Winchester College) this little Hampshire village for 30 years challenged and was a match for all comers. Indeed, in June 1777 they beat a representative England eleven by an innings and 168 runs. This phenomenal ascendancy resulted in the main from the coincidence in the area of about a dozen men of extraordinary cricket genius who had a profound influence on the evolution of the game's technique. The club's historian John Nyren wrote the first prose classics in cricket's literature.

Lord's cricket ground and the MCC

Hambledon played its last recorded match in 1793, appropriately enough at Lord's in London. This, the acknowledged Mecca of all cricketers, was first opened in Dorset Square as a private ground for certain members of the White Conduit Club by Thomas Lord, a Yorkshireman who was ground superintendent and bowler to the club. In 1809, to avoid a rise in rent, Lord removed to part of the St. John's Wood estate, and four years later moved again to the present location. In each case he relaid the original Dorset Square turf. He enclosed his last (the present) ground with a high fence and built a pavilion and a tavern. The Marylebone Cricket Club (MCC), with its home at Lord's, was founded in 1787. In its first year it revised the cricket laws, proving from its inception that its authority was paramount. Today it is accepted throughout the world as the authoritative source of all cricket legislation.

For half a century at least, the MCC was the great match-making agency in cricket, inviting subscriptions from its members to meet match expenses and advertising in advance at the chief London social clubs. Most of the big matches were played for money and were a field for heavy wagering by professional backers.

In 1836 the first North counties versus South counties match was played, clear evidence of the spread of cricket. The missionary efforts of the MCC begun in 1846 were enhanced and extended by the touring cricket of the All-England eleven, which played all over the country and was the focus of attraction wherever it went. It brought knowledge and appreciation of the game into whole districts where hitherto it had been unappreciated or undeveloped, and its success led in 1852 to the secession of some of the leading professionals and their formation into the United All-England eleven. These two teams monopolized the best cricket talent in the country until, eventually, they both gave way before the rising tide of county cricket (see below *History of county cricket*).

Sport and General Press Agency

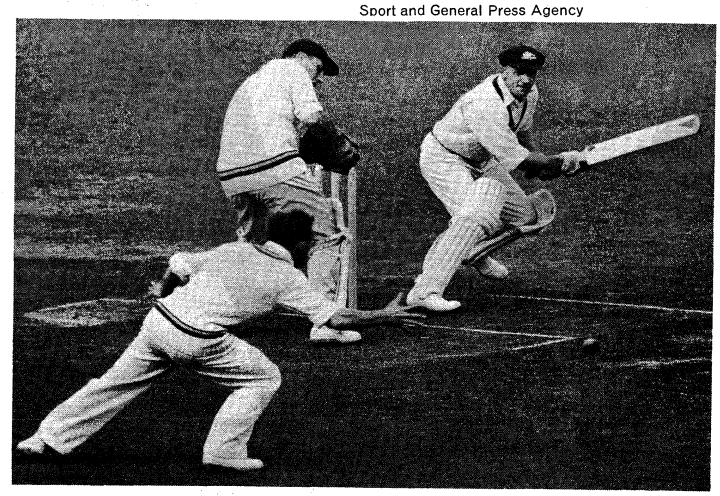


Figure 3: Sir Donald Bradman (Australia) at the finish of a cut. S.C. Griffith (England) is wicketkeeper and W.J. Edrich is at slip. Bradman established Australian and Test match records between 1928–48.

Technical development. Cricket was originally a game in which country boys bowled at a tree stump or at the hurdle gate into a sheep pen. This gate consisted of two uprights and a crossbar resting on the slotted tops, the crossbar called a bail and the whole gate a wicket. The fact that the bail could be dislodged when the wicket was struck made this preferable to the stump, which name was later applied to the hurdle uprights. Early manuscripts differ about the size of the wicket, but by 1706 the pitch was 22 yards (20 metres) long.

The ball was probably much the same in the 17th century as it is today. Originally it weighed between five and six ounces (140 and 170 grams). Its modern weight was laid down in 1774; its circumference, which had been standardized in 1838, was slightly reduced in 1927.

The primitive bat was no doubt a shaped branch of a tree, resembling a modern hockey stick but considerably longer and very much heavier. The change to a straight bat was made to meet the cult of length bowling (see above *Bowling*) which had been evolved by the Hambledon cricketers. The bat was shortened in the handle and straightened and broadened in the blade, which led to forward play, driving, and cutting (see above *Batting*). As few bowlers were able to combine length with break, swerve, and flight, batting continued to dominate bowling in the 18th century. About the same time the first leg before wicket law was promulgated requiring the umpire to perform the impossible task of deciding whether the obstruction had been deliberate. All bowling was underhanded early in the 19th century and most bowlers favoured the high tossed lob.

Bowling methods

The next bowling development was "the round arm revolution" with which many bowlers deliberately threw.

Controversy raged furiously, and in 1835 the MCC rephrased the law to allow the hand to be raised as high as the shoulder. The new style led to a great increase in pace. Gradually bowlers raised the hand higher and higher in defiance of the law, until it became more honoured in the breach than in the observance. Matters were brought to a head when all nine professionals of an England eleven playing against Surrey left the field at the Oval in protest against one of their number's being no balled for throwing, and, as a result, in 1864 the bowler was officially accorded full liberty to bowl overhand. From time to time umpires have had trouble with bowlers who threw.

Most bowling continued to be fast throughout the middle of the 19th century, and, though the standard of wicketkeeping was high, the role of long stop, a fieldsman positioned to back up the wicketkeeper, became most important (the position was abolished after wicketkeepers learned to contain the fast bowlers). The batsmen now learned to protect themselves with additional armour. Pads and tubular rubber batting gloves were developed, and a cane handle greatly increased the resilience of the bat. Thus fortified, batsmen developed their strokes. Only the best professional batsmen, however, could cope with fast bowling because most pitches were bad. Until 1849, when permission was first given to sweep and roll the pitch at the beginning of each innings, it was unlawful to touch it from the beginning to the end of a match. Gradually the grounds improved because of the advent of the heavy roller and the perfection of the lawn mower, and in the 1870s fast bowling declined. The batsmen went over to the offensive, and the bowlers fell back on accuracy of length. They also developed "off theory," keeping the ball some distance wide of the off stump and concentrating the fieldsmen on the off side of the wicket. The batsmen countered by a policy of masterly inactivity, and the pace of some county cricket became so funereal that interest flagged and attendance fell off. The game was saved from a lingering death only by a transfusion of new and vigorous blood.

Profiting by the coaching of English professionals, Australians developed their cricket, especially their bowling, and, when in 1882 they travelled to England and defeated its best players, a further evolution in technique had taken place. Their great bowlers combined pace with spin and variety of flight. They bowled much straighter than the English and were much more adaptable in their field placing. With the wicket-taking offspin, many catches were made in the then unprecedented position of silly mid on or forward shortleg (see Figure 1).

Developments in the 20th century. The opening years of the 20th century produced such an orgy of run scoring that a reform of the leg before wicket law was debated, and the Marylebone Cricket Club (MCC) denounced the overpreparation of pitches. But the heavy scores were due primarily to the arrival of batsmen who triumphed over much formidable and varied bowling by methods both versatile and individual. This was cricket's golden age.

The googly and new batting styles

There now appeared in cricketing vocabulary the googly, a word coined in Australia when B.J.T. Bosanquet, on the 1903–04 MCC tour, first exploited his ability to bowl an offbreak with a leg-break action. This freak ball was brought to something like perfection in South Africa, where the matting used to form the surface of the pitch intensified its spin.

Contemporarily, bowlers discovered a new weapon for the discomfiture of the batsmen—the swerve. Fast left-handers exploited inswerve, and right-handers learned to swing the ball either way, especially into a head wind or in a heavy atmosphere. A final development was the ball from the right-hander that swung into the batsman and dipped at the end of its flight. To supplement this inswinger, the field was largely reoriented, with the on side reinforced at the expense of the off.

To meet these new problems batsmen had to adjust their technique. At first players led the way by abandoning the long-striding forward stroke and making back play, together with a mastery of all on-side strikes (hits), the foundation of their batting. With the less gifted majority of batsmen, a two-shouldered and right-handed defense off the back foot began to predominate, and batting lost its aggression and attraction. Such methods played into the hands of bowlers of real pace and of those who could flight as well as spin the ball, but the concentration on defense and the increasingly elaborate preparation of the pitch often thwarted all but the greatest bowlers. The most drastic and dramatic tactic was the so-called body-line attack in the 1930s (see below *Test matches*).

Revision of rules. The reaction of authority to the evolution of cricket in the 20th century was visible in a series of attempts to help the bowler and to quicken the tempo of the game. As early as 1902 the bowling crease had been extended in length; in 1907 the use of a new ball after every 200 runs had been legalized; during the MCC tour of Australia in 1924–25 the eight-ball over was first used (this became standard practice in Australia and in South Africa); in 1927 a smaller ball was authorized; in 1937 the leg before wicket law was extended to cover balls pitching outside the off stump; and in 1947 the size of the wicket, increased in 1931, was again increased to its present size (see above).

In 1947 the MCC issued a revised code of laws aimed at the clarification and better arrangement of the previous laws. The extension of the leg before wicket law encouraged the inswinging bowler and the offspinner, thus further restricting the batsman's more interesting off-side strokes. During the 1950s the rate of run getting deteriorated, as too many bowlers bowled short of a length, and the MCC and other governing bodies were much concerned about slow play. As attendance continued to drop there were widespread appeals for more positive cricket, and groundsmen were urged to prepare truer and faster wickets. The MCC's attempts by legislation to instill more enterprising tactics into county cricket failed through inadequate response by players, whose offensive spirit was being dulled by playing six days a week before diminishing crowds.

Efforts to enliven the game. The 1960s were a period of great change, with continued efforts to enliven the game. In 1966 the MCC and the counties introduced Sunday play, allowing 12 three-day matches to be played on weekends. This was sufficiently successful for a continuation in 1967 and 1968, with more weekend matches, and in 1969 a one-day, Sunday-afternoon league was established among the first-class counties. The prototype one-day competition had been launched in 1963 by the inauguration of the Gillette Cup Knockout Competition, limits being placed on the number of overs, 60 per side and 12 per bowler. This proved a valuable means of interesting spectators and increasing the tempo of the game and developed into a highly popular competition, with a cup final at Lord's. A similar model was adopted for the Player's Sunday League, each county playing all the others on Sunday afternoons throughout the season, 40 overs per side, and eight overs per bowler. The league was an equal success and good crowds attended many of the matches.

History of county cricket. The greater part of first-class cricket, aside from the Oxford versus Cambridge and Test matches, is played in the county championship between sides representing counties.

As the 1960s progressed, county cricket continued to be of a low standard, and the same reasons were given for diminishing public support—the dull defensive attitude of the players and the frequent bad pitches. Consequently, in 1968 three steps were taken to meet objections: (1) a change in the regulations for registration of overseas players; (2) a new system of bonus points in the county championship; and (3) the appointment of an inspector of pitches.

Use of overseas players

1. To inject new blood into the championship, the ban on residential qualification of overseas players was partially lifted and immediate registration of one per county was permitted, with the proviso that a second could not be engaged within three years and that at no time could more than two per county be registered, though under a

previous arrangement an overseas player ranked as homeborn after five years. To safeguard his interests, an overseas player could at any time, without breaking his qualification, play for the country of his birth. Although Yorkshire continued to rely on homeborn players, most counties accepted the change with alacrity, so that famous Test cricketers such as G.S. Sobers, L.R. Gibbs, R. Kanhai, and C.H. Lloyd of the West Indies; B.A. Richards and M.J. Procter of South Africa; and G.R. McKenzie and A.N. Connolly of Australia were seen regularly in county cricket. All played a notable part, especially Sobers, whose all-around gifts as player and captain lifted his adopted county, Nottinghamshire, from bottom of the championship table to fourth in his first year. The experiment was a resounding success, though Australia, South Africa, and the West Indies all sounded a warning note as to this depredation of their best players to bolster what one Australian critic described as "England's archaic system of cricket."

2. As a reward for more positive cricket, it was agreed that during the first 85 overs of the first innings of each side, the batting side would be awarded one point for every 25 runs obtained over 150 and the bowling side one point for every two wickets taken. No other points were to be awarded on the first innings, which would not necessarily terminate after 85 overs. After two years' experiment there was a consensus that the players were learning to adjust to the new formula, thus providing better entertainment, and that spectators were coming to watch in increasing numbers.

Supervision of pitches

3. A county pitches committee was set up and an inspector of pitches was appointed. Complaints to the committee from captains and umpires were investigated by him, and some grounds were banned as being below standard.

No doubt these innovations helped the counties to lift themselves from a depressing rut. The overseas stars all played a notable part, and the abolition of first innings points produced keener cricket in the early stages of a three-day match. It is noteworthy that players of other countries, with less cricket, were less affected by a defensive complex, which was basically an English problem.

Early county champions. The origin of county cricket may be found in local antagonisms of the so-called home counties (those counties in the London area). Successive county supremacies were enjoyed by Kent (about 1750), Hampshire (1780–90), Surrey (1790–1810), Sussex (about 1825), and the great Kent eleven of the 1830s and 1840s. Yorkshire first took the field in 1833 (club formed in 1863) and Nottinghamshire in 1835 (club formed in 1859). Surrey (club formed in 1845) enjoyed a renewed ascendancy in the late 1850s and early 1860s. Between 1867 and 1870 Yorkshire, with a wholly professional eleven, were three-time champions.

The modern county championship is generally reckoned to date from 1873, when the MCC first laid down rules governing qualification. In the next five years, Gloucestershire, owing almost everything to the legendary Victorian cricket-playing brothers E.M., W.G., and G.F. Grace, were champions three times, but it was another ten years before the honour went south again. From the 1880s up to World War I, Nottinghamshire, Surrey, and Yorkshire (which with Lancashire, Kent, and Middlesex constituted the Big Six) predominated, but in 1911 Warwickshire became the first county outside the Big Six to win the honour in 34 years.

The early years of the 20th century produced J.B. (later Sir Jack) Hobbs of Surrey and F.E. Woolley of Kent. Hobbs was the complete batsman for 30 years, during which he amassed records for Surrey and England (61,237 runs, 197 centuries) that into the 1970s had not been beaten. Woolley, whose aggregate was second only to that of Hobbs, was cricket's most graceful left-handed batsman.

Reign of the northern counties. After World War I, Middlesex had a brief period of triumph, but between 1922 and 1939 the championship was monopolized by the northern counties, especially Yorkshire and Lancashire, which fielded largely professional teams that remained virtually unchanged throughout the season. They exhibited a toughness of fibre, an economical competence of technique, and a concentration on the game compared with which most of their rivals tended to appear, as in fact some of them continued to be, relatively unskilled. Above all they excelled in the accuracy and hostility of their bowling, supported by aggressive fielding and handled by captains who knew their jobs. During this era Yorkshire, with the help of W. Rhodes, greatest of all slow left-arm bowlers and also a good batsman, won the title 11 times, Lancashire five, and Nottinghamshire and Derbyshire one each, Derbyshire's victory in 1936 being the first outside the Big Six since Warwickshire's in 1911.

The nucleus of the Yorkshire side, which had won the championship in the last three seasons before World War II, was sufficiently strong to retain it in 1946, but thereafter it was gradually a waning force before the onslaught of Middlesex and Surrey, although Glamorgan, the youngest of the counties, surprised everybody by winning the championship for the first time in 1948.

First professional captain

In the meantime, amateurs good enough to be worth their place in first-class cricket and able to spare the time to play regularly were becoming increasingly difficult to find. Warwickshire in 1949 took the unprecedented step of appointing one of its senior professionals, H.E. Dollery, as full-time captain.

Two years later his team won the 1951 championship, 40 years after its only other success. Within a few years, five other counties had followed suit by appointing professional captains.

The 1950s belonged to Surrey; by 1958 it had been champion for seven consecutive years. Then, after a lapse of 13 years, Yorkshire came into its own again and won the championship seven times between 1959 and 1968. Enlivening this period were strong bids by other teams and the first-ever victories of Hampshire (1961), Worcestershire (1964 and 1965), and, in 1969, the return of 1948 surprise winner Glamorgan to the championship. Kent won in 1970 for the first time since 1913, and Surrey was back on top in 1971.

Other first-class cricket. For almost 150 years the annual Gentlemen versus Players match between two teams made up of the best amateur and the best professional players was the leading representative domestic fixture. Invariably the appearance of 22 of the leading cricketers in the same game produced high-class cricket without the tensions of intercounty competition and championship points. The series that began in 1806 and became annual in 1819 was ended in 1962 when the Marylebone Cricket Club (MCC) and the counties abandoned the distinction between amateurs and professionals.

The Oxford versus Cambridge match has been played at Lord's since 1827. These two universities rank with the counties as first-class and are the chief nurseries of amateur cricket. Until the outbreak of World War II, the match was one of the social occasions of the summer season in London.

Reorganization: the Cricket Council. A big reorganization of the overall administration of English cricket took place in 1969, resulting in the end of the MCC's long reign as the controlling body of the game, though it still retained responsibility for the laws. With the setting up of the Sports Council (a government agency charged with the rational promotion and control of sports in Great Britain) and with the possibility of obtaining government grant aid for cricket, the MCC was asked to create a governing body of the game on the lines generally accepted by other sports in Great Britain. Thus was inaugurated the Cricket Council, comprising the National Cricket Association, the Test and County Cricket Board (TCCB), and the MCC. The council consists of representatives from clubs, schools, Royal Army, Navy, and Air Force cricket, umpires, and the Women's Cricket Association. The Test and County Cricket Board amalgamated the Advisory County Cricket Committee and the Board of Control of Test matches at home. The board is responsible for all first-class cricket in England and for overseas tours which had previously been the respon-

sibility of the MCC, which continued, however, to play an important part.

Apartheid and test cricket

Unhappily, the problem of apartheid arose when the Cape Coloured South African, B.L. D'Oliveira, who had qualified by residence to play for England, was not included on the English team chosen to tour South Africa in 1968–69. There was a great public outcry, and, when one of the players had to withdraw because of injuries, D'Oliveira was chosen but was declared unacceptable by the South African government, and, as a result, the tour was cancelled. An even more publicized crisis arose over South Africa's projected tour in England in 1970; the threat of widespread anti-apartheid demonstrations if the tour were to take place caused elaborate precautions to be taken on the cricket grounds where matches with the visitors had been arranged, and, eventually, the MCC was persuaded by the British government to cancel the tour. A similar situation arose in 1971 with the cancellation of a South African tour of Australia.

Women's cricket. Women first played cricket in England in the 18th century, and spasmodic references appear during the 19th century. In 1887 the first club, White Heather, was formed, and in 1888 two professional teams known as the Original English Lady Cricketers were in action. About the same time, women's cricket was first played in Australia and New Zealand.

In 1926 the Women's Cricket Association was founded after a successful cricket week at Malvern, and these weeks became an annual event. Representative meetings soon followed, and in 1933–34 the first team toured Australia and New Zealand. Australia paid a return visit in 1937. By the second half of the 20th century the Women's Cricket Association was a flourishing organization, responsible for the administration of the game in England and for organizing Test matches at home and away. Test matches take place at regular intervals against Australia and New Zealand, and new ground was broken in 1960–61 with a first tour of South Africa, where the women's game had begun only in 1951.

INTERNATIONAL CRICKET

The English introduced their national game wherever a pitch could be found and two teams collected. The firmest roots are in those countries that, with England, were members of the Imperial Cricket Conference, founded in 1909 and including Australia, New Zealand, and West Indies; India and Pakistan (after 1947); and South Africa until the latter's withdrawal from the Commonwealth in 1961. In 1965 the conference was renamed the International Cricket Conference to permit the election, as full or associate members, of countries outside the Commonwealth, and the United States, Ceylon, and Fiji were elected associate members followed in 1969 by Bermuda, Canada, Denmark, East Africa, Gibraltar, Hong Kong, Malaysia, and The Netherlands. Cricket also flourishes in West Africa, Brazil, and Argentina. The following account is concerned with cricket in the main countries that are members of the conference.

Australia. Cricket began in Australia in the early 19th century and was competitive between clubs from the start. The first interstate match took place in 1856 between Victoria and New South Wales. By the end of the century South Australia was playing with them in a tri-

National cups and trophies

angular competition for the Sheffield Shield, presented by Lord Sheffield who took an English team to Australia in 1891–92. Queensland was the next to join, and Western Australia was accepted to membership after World War II. Despite the huge distances, all five states play one another home and away each season. New South Wales and Victoria are the dominant teams.

The real foundation of Australian excellence is club cricket. Each state has an association of affiliated clubs that controls players and grounds. Clubs are graded and competitions are arranged for each grade. First-grade teams, in which novices play with great cricketers, are the recruiting ground for the state team.

New Zealand. The game was first played in New Zealand in 1841, and the first representative game was between Auckland and Wellington in 1860. In South Island, the discovery of gold in Otago brought many Australians, who helped to spread the game, and the first of the annual meetings between Otago and Canterbury took place in 1864. Canterbury and Wellington met for the first time in 1885.

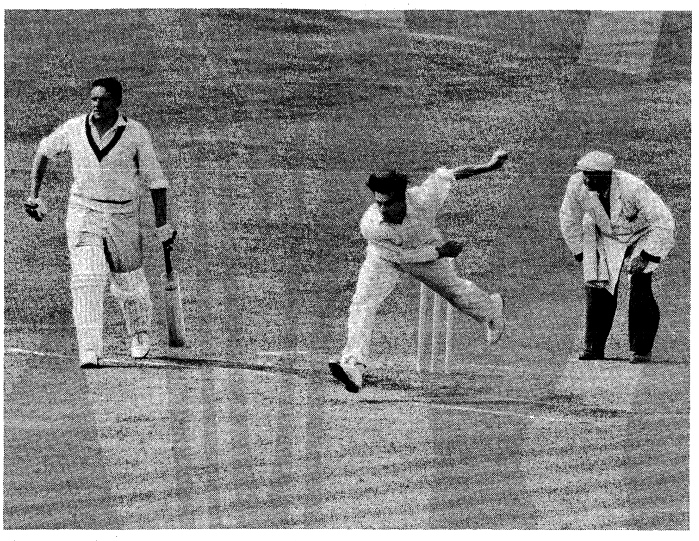


Figure 4: F.S. Trueman bowling for England in a Test match against Australia. Batsman A.K. Davidson and umpire J. Langridge watch the delivery. Trueman took a world-record number of wickets in Test matches in the period 1952–65.
Sport and General Press Agency

The New Zealand Cricket Council was formed in 1894. Each province has its own association, made up of delegates from affiliated clubs. The Plunket Shield (named after a former governor general) is played for annually between Auckland, Wellington, Canterbury, Otago, Northern Districts, and Central Districts. The first English team visited the country in 1864, and others have gone at frequent intervals.

West Indies. Cricket has probably been played in the West Indies since mid-19th century, but the first intercolonial match took place in 1865 between Demerara (British Guiana, now Guyana) and Barbados. By the end of the century, a triangular tournament was being held between Demerara, Barbados, and Trinidad. In 1926 a board representing all the islands was set up to control inter-island competitions and to organize foreign tours. One of its problems has been to agree upon a national captain, so intense is the rivalry between the islands.

After World War II air travel enabled Jamaica to play the other islands, and in 1956 Barbados, British Guiana, Trinidad, and Jamaica played a knockout tournament at Georgetown, which British Guiana won. In the 1960s these four were joined by the Windward Islands and the Leeward Islands in an annual competition for the Shell Shield, presented by the Shell Oil Company.

India. English settlers introduced cricket in India and the army helped to popularize it. The first all-Indian club was the Oriental Cricket Club, for Parsis (an Indian religious sect), formed in 1848. English professionals improved the standard of play, and there was a gradual sinking of jealousies of the various races and creeds in the cause of cricket. In 1877 the first match between the Parsis and the Europeans in Bombay was played, and the Parsis showed such improvement that Parsis teams went to England in 1886 and 1888.

In 1906 the Hindus competed with the Parsis and the presidency (the Europeans) in a triangular tournament, and by 1912 this had become quadrangular with the advent of a Muslim team. Pitches were uneven and gear crude, but enthusiasm was tremendous. In 1926 the MCC sent a team to India, and the princes began to pick their own sides—a sign that cricket was spreading. Since 1934 India has staged a national championship for the Ranji Trophy, named after famous Indian cricketer K.S. Ranjitsinhji, with "the Rest" joining the quadrangular in 1937.

Pakistan. Since partition in 1947, cricketers in Pakistan have had a difficult task of organization. An embryo test team was soon formed and during the first years of its existence it has had a good record. In an attempt to build up a cricket tradition and find reserves for the national eleven, the country has been divided into four groups for a championship for the Qā'id-e A'zam Trophy (meaning "the great leader" in honour of Mohammed Ali Jinnah, one of the creators of Pakistan).

South Africa. Cricket was taken to South Africa in the middle of the 19th century by British troops and was played in military and police centres of the Cape and Natal till 1889. Despite a delightful climate, interstate meets were difficult owing to wide distances and sparse population.

In 1889 Sir Donald Currie offered a cup to the team that put up the best performance against the first English touring team. Kimberley won, and next year Transvaal challenged and beat them. Thus began the Currie Cup competition, which is South Africa's provincial championship. Before the turn of the century, Cape Colony (now Cape Province) had entered two teams (Western Province and Eastern Province), and Natal had also joined. After World War I, a board of control became the ruling body for South African cricket. The three most powerful provinces are Transvaal, Western Province, and Natal.

TEST MATCHES

History. The first international, or Test, match between sides of players chosen on the basis of their ability to represent the best cricket each country had to offer was played in Melbourne in 1877, when a team made up of Australia's best players beat a team of England's finest by 45 runs. Australia's success was repeated in 1882 at the Oval, in London, in the game celebrated in the *Sporting Times* by an obituary notice announcing cricket would be cremated and the ashes sent to Australia, giving birth to the legend of "playing for the Ashes," which became the popular objective in games between England and Australia. The ashes, kept in an urn at Lord's, are those of a stump burned on the England tour of Australia in 1883. For the rest of the century, the two countries met almost yearly. With W.G. Grace in his heyday, England was generally too strong, though Australia had the greatest bowler of his era in F.R. Spofforth and the first of the great wicketkeepers in J.McC. Blackham.

The "Ashes"

In the meantime new opponents for England were gathering strength. In 1894 the South Africans first went to England, and their cricketers gained invaluable experience during two subsequent tours in South Africa, one by England and one by Australia. In 1905 the first MCC team to South Africa lost the rubber primarily because of googly bowling, which was too much for English batsmen on the fast matting wickets. The triumph won for South Africa the right to Test matches in England, and in 1907 they proved their ability to play England's best.

After World War I, England, Australia, and South Africa continued to play one another at regular intervals, and in 1928 the West Indies entered the Test match arena. In 1931 New Zealand joined them, followed by India in 1932. The result was a serious crowding of the schedule. England, with its county cricket for six days a week, was hardest hit, and life for the top players became a round of cricket and travel. Interest in Test matches was stimulated in the 1930s by the introduction of radio broadcasts of games and by the advent of the correspondent whose objective was news stories and off-the-field activities at the expense of straightforward cricket reporting. In 1930 Test matches between England and Australia were increased to four days' duration.

Growth in number of test matches and in interest

England was slow to recover from World War I, and Australia dominated the Test-match scene for six years, until England recovered the Ashes in 1926 and held them for four years. Then, in 1930, English crowds for the first time saw the new, young Australian phenomenon D.G. (later Sir Donald) Bradman. Between 1928 and 1948 he scored a record 29 centuries, 19 of them against England, and his average in all Tests was 99.94 runs. On his first English tour he made innings of 254, 334, and 232 in the Tests, and these, with the slow bowling of C.V. Grimmett, were sufficient to win the rubber.

D.R. Jardine, captain of the next MCC team to Australia (1932–33), evolved the plan known as body-line, with H. Larwood as chief protagonist: bowling very fast and accurately at the leg stump (the one nearest the batsman) with a majority of fieldsmen on the leg side, he effectively curbed Bradman and his colleagues and was largely instrumental in winning the rubber for England by four games to one. The new technique, however, savoured so much of direct attack on the batsman that it caused a storm of protest and was subsequently outlawed by the MCC as being contrary to the spirit of the game.

Body-line bowling

In the remaining Anglo-Australian rubbers before World War II, Australia won in 1934 in England and in 1936–37 in Australia; but in 1938 a new English batting star, Leonard (later Sir Leonard) Hutton, made history with an innings of 364 at the Kennington Oval, London, which ensured a record victory by an innings and 579 runs. Meanwhile, South Africa won for the first time at Lord's in 1935 by 157 runs.

Between these Tests against ancient rivals, England sandwiched three at-home matches against the West Indies and two each against India and New Zealand, plus return visits to New Zealand after each of the Australian tours and one to India, in 1933. In 1934–35 an MCC team toured the West Indies, which had been to Australia for the first time in 1930–31. In 1931–32 the South Africans toured Australia and New Zealand for the first time, and in 1935–36 they received Australia. Hence the pattern of continuous cricket was already being evolved when war broke out in 1939.

After World War II Test cricket began again, and in every summer since 1946 a touring side has engaged in Test matches against England. Before the late 1960s Australia made the trip twice in every eight years, South Africa every five years, and the other countries at longer intervals. Pakistan was admitted to the international association in 1952 and first toured England in 1954. In 1966 a four-year cycle of tours was announced, in which twin tours to England (by India and Pakistan, or by the West Indies and New Zealand) would alternate with single tours (by Australia or South Africa). England has continued to travel to Australia every four years, and visits to other countries have been a feature in two out of every three winters. The other members of the International Cricket Conference have gradually extended their schedules.

Increase of tours

Test matches last 30 hours (six days in Australia and the West Indies and five days elsewhere), except by local agreement.

During the 1960s, violence among spectators gave increasing concern to cricketing authorities. Test matches in Pakistan, India, and the West Indies were drawn or abandoned owing to riots, part political, part nationalistic; and, as noted above, in the 1970s South African tours of England were canceled because of the threat of antiapartheid demonstrations.

The test countries. *Australia.* As in the 1920s, the first few years after World War II were dominated by Australia. Bradman was still a tremendous force, and the team he led to England in 1948 was as strong as any in history, with K.R. Miller and R.R. Lindwall as one of the great fast-bowling combinations of all time. This team held its own in world cricket until 1953 when, at the Oval, Australia at last relinquished the Ashes held since 1934. Australian recovery began with a successful tour of South Africa in 1957–58 and continued with victory over England in Australia in 1958–59. The next four series produced only eight results with 12 draws in 20 matches, Australia retaining the Ashes by winning two rubbers in England and drawing two at home. Australia was beaten for the first time by the West Indies in the Caribbean in 1965. However, at home in 1968–69, they had a successful series against the West Indies. Australia was beaten by South Africa in 1969–70, and lost the Ashes to England in 1970–71.

England. England took some time to recover after World War II, but returned to top form in the 1950s

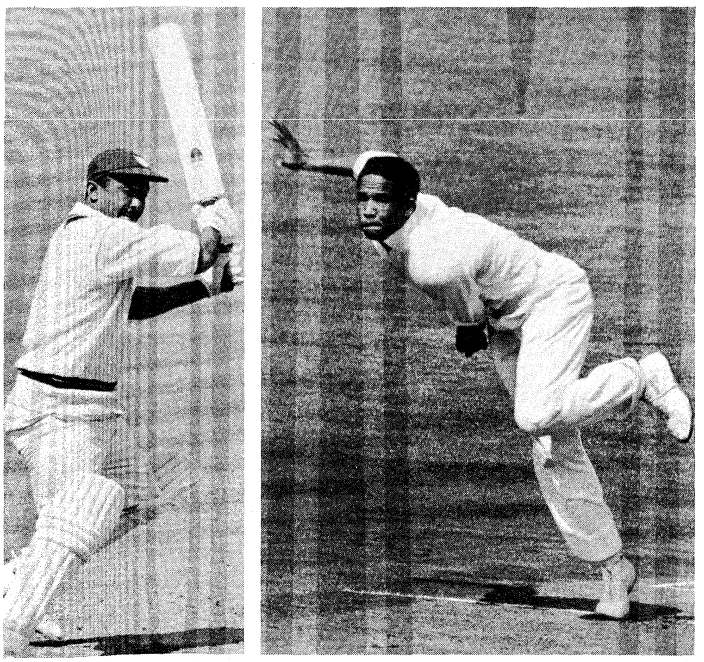


Figure 5: F.M. Worrell batting and G. Sobers bowling for the West Indies. Both exceptional all-round players, they captained the West Indies to world supremacy during the 1960s.
(Left) Sport and General Press Agency, (right) Central Press Photos

and did not lose a Test rubber between 1952 and 1958 (in 1953 it recaptured the Ashes under Hutton, its first professional captain). After 1958 English teams were usually too strong for New Zealand, India, and Pakistan and had equality with South Africa but in the 1960s lost four rubbers at home, two each to Australia and the West Indies. In the period 1952–65, fast bowler F.S. Trueman took a world-record number of wickets (307) in Test matches. Having regained the Ashes, England lost for the first time at home to India in 1971.

South Africa. After World War II South Africa contested eight rubbers against England and achieved its first win in a three-match series in England in 1965. Australia had always been invincible in South Africa, until 1966–67, when South Africa for the first time won the Test series and overwhelmed Australia again in South Africa in 1969–70. But South Africa has done better in Australia, drawing series in 1952–53 and 1963–64. South Africa had an easy success in New Zealand in 1953–54 and played two drawn rubbers, home and away, in 1961–62 and 1963–64. South Africa's participation in Test cricket remained clouded at the beginning of the 1970s because of its adherence to the principle of apartheid and the resentment which this aroused abroad.

West Indies. The West Indies reached world class after World War II, and in the mid-1960s they were undoubted world champions. They owed their early success primarily to C.L. Walcott, wicketkeeper-batsman, E.D. Weekes, batsman, and F.M. (later Sir Frank) Worrell, all-rounder—and to two young spin bowlers, S. Ramadhin and A.L. Valentine. After beating England in England for the first time in 1950 and holding them to a draw in the Caribbean in 1953–54, they suffered defeats against England, both home and away, but in the early 1960s Worrell's team, though narrowly losing a rubber, won the hearts of Australian crowds with its joyous cricket and, at Brisbane, took part in the only tied match in Test history. Worrell's final triumph was the defeat of England in England in 1963. He was knighted for his services to cricket in 1964 and died in 1967. His successor G. Sobers, holder of the record Test score of 365 "not out" against Pakistan, led the West Indies to victory against Australia in the Caribbean and against England in England in successive years, by which time he was acclaimed the most gifted all-rounder of all time.

The West Indies were beaten by India for the first time in 1971. Pakistan won a Test in 1958 at Trinidad and a three-match rubber at home in 1958–59. Ironically New Zealand's first Test win was against the West Indies in 1956 at Auckland, and they won again at Wellington in 1969.

New Zealand. New Zealand won its first Test match in 1956, against the West Indies, and won two against South Africa in 1961–62 when it drew a rubber in South Africa. It has also toured England, India, and Pakistan without success and has entertained at home all the Test-playing countries except Australia. In 1969 at home it inflicted a second defeat on the West Indies in a drawn three-match rubber and then lost 0–2 in England. Leading players since World War II have included left-handed opening batsman B. Sutcliffe and fast bowler R.C. Motz, the only New Zealander to have taken 100 wickets in Test cricket.

India. In the years after World War II, India played rubbers against all countries except South Africa, and, while being no match for England and Australia away from home, proved hard to beat on its own pitches, where England was twice beaten in 1961–62 and Australia once in 1959–60 and once in 1964. Other wins have been against New Zealand (three) and Pakistan (two). It has yet to win a match against the West Indies. The pattern away from home was continued when it lost a short series in 1967 in England (3–0) and all four matches in Australia (1967–68) but it triumphed in 1971 over both the West Indies in the Caribbean and England in England, in each case for the first time.

Pakistan. Pakistan made an auspicious entry into world cricket by beating England at the Oval in 1954, Australia at Karāchi in 1956, and the West Indies at Port-of-Spain in 1958. In Pakistan it beat the West Indies in 1958–59, and other victories have been against New Zealand (two) and India (one). Two consecutive drawn series with India were played in 1954–55 and 1960–61, neither side winning a match. It was beaten in England in 1967, and an English visit in 1969 had to be abandoned (0–0) due to political riots.

BIBLIOGRAPHY

History: JOHN NYREN, *The Young Cricketer's Tutor* (1833), *Cricketers of My Time* (1836); JOHN ARLOTT (ed.), *From Hambledon to Lord's* (1948), *The Middle Ages of Cricket* (1949), and *Rothman's Jubilee History of Cricket, 1890–1965* (1965); E.V. LUCAS (ed.), *The Hambledon Men* (1954); H.S. ALTHAM and E.W. SWANTON, *A History of Cricket*, 4th ed. (1948), with extensive bibliography; E.W. SWANTON (ed.), *The World of Cricket* (1966); ROY WEBBER, *The County Cricket Championship* (1957); R.S. RAIT KERR, *The Laws of Cricket: Their History and Growth* (1950); MARYLEBONE CRICKET CLUB (ed.), *Laws of Cricket* (various years and editions). (*Australia*): GEORGE GIFFEN, *With Bat and Ball* (1898); F.A. IREDALE, *Thirty-Three Years' Cricket* (1923); A.G. MOYES, *A Century of Cricketers* (1950), *Australian Bowlers from Spofforth to Lindwall* (1953), *Australian Batsmen from Charles Bannerman to Neil Harvey* (1954), *Australian Cricket* (1959). (*South Africa*): M.W. LUCKIN, *History of South African Cricket*, 2 vol. (1915–27); LOUIS DUFFUS, *South African Cricket, 1927–1947* (1948). (*New Zealand*): T.W. REESE, *New Zealand Cricket, 1841–1914* (1927), . . . *1914–1933* (1936). (*West Indies*): CHRISTOPHER NICOLE, *West Indian Cricket* (1957). (India): D.B. DEODHAR, *March of Indian Cricket* (1948); "Arbi," *Indian Cricket Cavalcade* (1956).

Records, scores, and statistics: JOHN WISDEN, *Cricketers' Almanack* (annual, 1864–); ROY WEBBER, *Book of Cricket Records* (1961); ARTHUR WRIGLEY, *The Book of Test Cricket, 1876–1964* (1965); BILL FRINDALL, *The Kaye Book of Cricket Records* (1968). In addition to the British *Wisden*, almanacs, annuals, or yearbooks are published in Australia, India, New Zealand, Pakistan, and South Africa; and in the U.K. there are two monthly magazines and one quarterly.

Technique: SIR DONALD G. BRADMAN, *The Art of Cricket* (1958); MARYLEBONE CRICKET CLUB, *Cricket Coaching Book* (1952), *How to Play Cricket* (1957).

Anthologies: GERALD BRODRIBB (ed.), *The English Game* (1948), *The Book of Cricket Verse* (1953); ALAN ROSS (ed.), *The Cricketer's Companion* (1960); L. FREWIN (ed.), *The Best of Cricket's Fiction*, 2 vol. (1966–68).

(R.Al.)

Crime and Delinquency

There is no generally accepted definition of crime and delinquency that is of universal application. Within each culture the line between normal and criminal, or merely deviant, behaviour is drawn differently by varying codes or bodies of criminal law. In general, however, the term crime is used to refer to adult behaviour; the term delinquency is used to refer to criminal or, in some countries, precriminal behaviour by a juvenile. The age limit between juveniles and adults varies greatly from country to country, ranging, in general, from a lower limit of 15 years to a maximum of 18.

Crime, law, and morality

The definition of crime within a given culture depends largely on two sets of related norms: the legal and the moral codes prevailing within that culture. In legal terms, a crime is an act of human behaviour banned by criminal law. Even this deceptively simple definition is currently challenged as being too broad; not all acts that are penalized under the law are really criminal. Minor traffic offenses are an obvious example, and the range of acts to which penalties attach, but which are not legally or socially thought of as crimes, is widening. This is particularly true in industrialized countries, in which the emergence of huge urban areas and the increased involvement of government in all aspects of life has resulted in a new type of offense: the frequently unintentional violation of administrative rules, sometimes referred to as a public welfare offense. In addition, there are certain types of behaviour that, because of their increased frequency or because of increased public tolerance, have lost their former status as criminal. Some minor drug addictions, many abortions, and certain kinds of sexual behaviour may fall into this group.

Though the term crime should be reserved only for behaviour legally defined as criminal, criminologists must be concerned as well with all potentially criminal behaviour, regardless of legal definitions.

Because a broad social definition of crime is more widely used than a narrow legal one, moral and cultural elements, including value systems, mores, and religious attitudes operative in a given culture, are of central concern to the criminologist. Crime and morality may be likened to two circles overlapping with a wide common area. A large part of proscribed human behaviour is both criminal and immoral, but a considerable part of it is only one or the other. Whereas criminal law is primarily concerned with the external aspects of human behaviour, morality focusses on the inner aspect. The criminal law is negative—it forbids a certain act—whereas moral considerations may dictate that certain acts be performed.

Society's response to crime

The fate of the criminal offender is contingent on the uneven, and sometimes contradictory, functioning of the typical criminal-justice system, and the reaction of that system to crime is far from being uniform or certain. The response of society to crime reflects different ideologies concerning the treatment of offenders. Punishment and deterrence have prominent places within those ideologies. The punitive societal reaction has two major characteristics: it is inflicted by the group on a member of the group as an instrument of public justice, and it presupposes pain and suffering—justified by its retributive value.

Attitudes and reactions to the criminal offender have broadened, so that punishment alone is no longer the only motive informing society's response toward criminal behaviour. Crime prevention is largely viewed as involving much more than the prevention of further offenses and the deterrence of potential offenders by means of exemplary sentences. Crime prevention is now approached from two different perspectives—individual and general. Society's response to the individual offender is determined by a variety of cultural and social factors mentioned above (see also PUNISHMENT; and PRISONS AND PENOLOGY). General crime prevention is an aspect of social policy; mental- and physical-health policies and social-welfare programs, for example, contribute to general prevention, and the dividing line between criminal prevention and social welfare is sometimes only a matter of definition. Delinquency in some countries includes conduct that is considered dangerous, inappropriate, or symptomatic of future criminality. There is, however, a general tendency to limit delinquency to conduct that, if carried out by an adult, would be identified as crime.

Most of the following considerations are taken from crime and delinquency studies in Western countries, particularly the United States. Few data are available from either Eastern countries or the developing nations. The article is divided into the following sections:

I. Causes and patterns of crime
- The measurement of crime
  - The administrative crime index
  - The research crime index
  - Crime indexes in Europe
  - Trends and comparisons
- The analysis of crime
  - The variety of approaches
  - Density and size of the community
  - The criminal population
  - Class structure
  - Criminal subcultures
  - Influence of the family
  - Educational factors
  - Victimology
- Criminal classifications
  - Normal and abnormal offenders
  - Offender careers
  - Occupational offenders and white-collar criminals
  - Organized crime
  - Homicide and crimes of violence
  - Drug addiction
  - Intoxication and alcoholism
- The control of crime
  - Police and the public
  - Corrections

II. Patterns of juvenile delinquency
- The analysis of delinquency
  - Sex and age distribution
  - Lower and middle class delinquency
  - The delinquent gang
  - Home conditions
  - Prediction and prevention
- Responses to delinquency
  - Probation
  - Juvenile aftercare
  - Foster care

I. Causes and patterns of crime

THE MEASUREMENT OF CRIME

Economists are able to construct a cost-of-living index and various other indexes that serve social administrators and policy makers as useful indicators of how an economy is functioning. Though without equally clear policy aims, criminologists have long urged the systematic collection of crime statistics, in the hope that an index could be constructed to provide a useful measure of the amount of crime in society and of its trends.

Unreliability of crime statistics

The difficulties in constructing a reliable crime index are immense. Criminal statistics are well-known for their unreliability, which results from a number of factors prominently including, first, the wide diversity in the definition of various crimes under different criminal codes and, second, the diversity of statistical sources, which may be local police departments, state or federal police departments, courts, or correctional institutions, and the fact that not all reportable crime is either properly or uniformly recorded by police. It has, in fact, been argued that crime statistics are merely an indication of the activities of the law enforcement agencies. Attempts, however, have been made to construct an index that would be of value from either an administrative or a research point of view.

The administrative crime index. In 1930 the Federal Bureau of Investigation (FBI) of the United States assumed responsibility for collecting and publishing reports of crimes "known to the police." These were voluntarily submitted by various police departments throughout the country. A standard form is now used, including standardized instructions for defining and classifying offenses. The offenses used to make up the overall crime index are murder, forcible rape, robbery, aggravated as-

sault, burglary, larceny over $50, and auto theft; and the *Uniform Crime Reports*, as published, provides a breakdown of the index into violent crimes and crimes against property. Much research into crime trends usually devotes attention to the more serious crimes of violence, especially homicide, because this crime has a high reporting rate.

The research crime index. In the collection of criminal statistics there has long been a controversy about whether to count the number of offenses or the number of offenders. The favoured solution, so far, appears to be to concentrate on the delinquent event, which may be reconstructed from records of offenses known to the police and analyzed according to three main criteria—the presence of injury, of damage, or of theft. Because the purpose of measuring the amount of crime is to examine its relationship to the rest of society—that is, to determine the "social harm" caused by criminal conduct—it is necessary to make a further assessment of the "seriousness" of the crime involved, taking societal attitudes into account. The inclusion of auto theft in crime indexes has been criticized because the amount of social harm or seriousness of the conduct has been left unassessed. Research has shown that when people are asked to assess the seriousness of various criminal acts in comparison to one another, some crimes are more consistently rated as more serious or less serious than others. Forcible rape, for example, has been found to be rated about six times more serious than auto theft. The final research crime index is constructed by assessing each offense according to the criteria of injury, theft, or damage or to the seriousness estimated in public opinion surveys.

Crime indexes in Europe. It is difficult to speak of crime statistics in Europe without discussing the problem of collecting international crime statistics. Each country has its own tradition, or established way, of crime reporting. None have attempted to present an overall administrative crime index as has the FBI in the U.S. With the exception of Britain, crime statistics in Europe have in the past concentrated mainly upon court or prison statistics, not on police reporting statistics as have the *Uniform Crime Reports* and the research index.

Britain probably produces the most extensive set of crime statistics, mostly of the administrative variety. *Criminal Statistics in England and Wales* is published six months after the close of the year and presents a very detailed classification of offenses. It publishes both police and court statistics, crimes cleared by the police, and tables that allow comparison of police and court figures for individual offenses per million of population since the 1940s. In another publication, *Report of the Commissioners of Prisons*, extensive statistics on the prisoner population are presented.

The French volume entitled *Comte Général de l'administration de la justice civile et commerciale et de la justice criminelle* has been published since 1825 and contains criminal statistics based upon court, not police, statistics. These provide a distinction between repeaters and first offenders; detailed information of the "milieu social"—the occupation and other factors related to the offense; and a classification according to age, sex, offense, and sentence. In Italy the *Annuario di statistiche Giudiziarie* presents every year police and judicial statistics. Unfortunately there is usually a two-to-three-year delay in publication, and the criteria and method of reporting are changed frequently.

Denmark's annual publication, *Kriminalstatistik* (with English translations of tables), has for the past 20 years provided very comprehensive statistics, containing detailed breakdowns of offenses, police and judicial statistics, and such factors as residence at time of committal, the presence of alcohol, previous criminality, types of sentences and disposition, occupation of the offender, and many others.

Other countries in Europe have collected criminal statistics according to their own traditions. Because of the divergent ways in which the law is administered in the various countries in Europe, it has proved difficult to establish any standard procedure of crime reporting.

The Council of Europe has recently turned attention away from the insurmountable difficulties in obtaining a direct administrative solution to the problem of uniform crime reporting, to a close study of the research index, in the hope that this might provide the basis for an international agreement on the construction of an index of crime that is comparable from country to country, and from region to region. The suggestion has been made that it may be possible to apply the "seriousness" weights that the research index provides to already existing administrative crime statistics. Much work has yet to be done on this, but there is hope that in the future the research index may provide truly comparable international crime statistics.

Trends and comparisons. Although it is probably fair to say that, since the end of World War II, crime defined in the broadest of terms has increased substantially in most Western countries and in a number of developing countries as well, any interpretation of this trend must be made with great care. Apart from taking into account the sources of unreliability already described, it is necessary to consider such factors as changes in the structure of the population with regard to age, sex, and race. The *Uniform Crime Reports* of the United States for 1969 shows that, while the population of the United States rose 13 percent between 1960 and 1970, the overall crime rate indicated by the index of offenses described above rose 148 percent; violent crimes rose 130 percent, and property crimes 151 percent. In England and Wales from 1938 to 1960, crimes of violence increased by 130 percent, with substantial increases occurring over the latter ten years of the period. Substantial increases also have been reported in a number of European countries and Japan.

THE ANALYSIS OF CRIME

The variety of approaches. Since the beginning of the scientific study of crime, attempts have been made to isolate those factors that, alone or, more frequently, in complex interactions, appear to be associated with criminal behaviour. Many hypotheses have been presented, from those centred around the existence, in the offender, of biological, constitutional, or acquired anomalies to those that search for the "cause" of crime in economic and social factors or that identify the "causes" in psychological deviations.

Today, primarily as a consequence of the development of a "clinical," individualized approach to the study of offenders, the existence of several mechanisms in which many factors enter simultaneously into play or in different periods of the life of the individual is conceded. This has not, however, reduced interest in general theories about criminal behaviour. But the search for a single cause of crime has been largely abandoned as fruitless. Although, at an individual level, an interdisciplinary clinical study may establish cause–effect relationships and give useful indications for modifying or controlling criminal behaviour, no such attempt is successful at a general level.

Many theories that have been presented in the past are today of only historical interest because they have either been disproved or have been found to account only for a small percentage of criminal events. Early statements on the influence of geographical and climatic factors on crime rates, for example, have been negated by subsequent researches. Seasonal variations in crime have been repeatedly presented as pointing to a direct relationship, but again, these are now considered to be of little or no usefulness in explaining antisocial or criminal behaviour.

Biological and genetic factors

A search for biological and genetic factors in crime causation characterized the beginning of scientific criminology. Several factors, including such things as supposedly "criminal constitutions," physical type, organic brain injuries or diseases, endocrinal abnormalities, hereditary traits, and, more recently, chromosomal aberrations, have been postulated as causes of criminality. None of these attempts has been successful in presenting a valid general theory of crime causation. At an individual level, all the factors presented may help in understanding the

specific makeup of an offender or may explain a single offence, particularly if it is one of a violent, unpredictable, and seemingly motiveless nature. But, even when such factors provide an adequate explanation of a particular criminal personality or criminal act, it is invariably the case that the individual involved is in some way abnormal. Biological factors, therefore, at best appear to account for a minor group of offenders and offenses and leave unexplained the majority group of "normal" offenders and offenses.

Similar considerations are relevant to psychological, psychoanalytical, and psychopathological perspectives on crime. Criminal behaviour may occur as a symptom of several types of mental illness. Motiveless murder, for example, may indicate schizophrenia, and kleptomania may be a symptom of a compulsive neurosis. But such explanations account only for those cases in which the mental illness or the psychological abnormality has reached a causative level in the specific offender, and not for homicide or theft in general. Descriptive analyses of offenders in general, or of specific types of offenders, have been attempted by psychologists using psychodiagnostic tests or by psychoanalysts employing the theoretical approaches of dynamic psychology. Few general causative facts have emerged. At a descriptive level, psychological studies have been of some usefulness in presenting standard descriptions of offenders as types. Most typologies of offenders, however, fail to account for all the varieties of criminal behaviour. Recent typological efforts have tended to include sociological factors, leading sociologists to shift their perspective from the individual offender to the study of group societal processes affecting the individual and contributing to his criminal behaviour. The relationship between the offender, or offenders, and the social environment becomes the focus of the search for causative factors. Sociologists face the problem, however, that, regardless of the explanatory level of their theories, sociological factors can never account for individually different responses to supposed crime-causing factors. Faced with similar social factors—such as slums or social discrimination—only a minority of the individuals so exposed become criminals.

This difficulty has shifted modern research efforts away from causes toward processes and models. It has, at the same time, made clear that a causative explanation of crime must of necessity be "integrated," fusing sociological and psychological concepts and occasionally including biological factors in an individualized explanation of each criminal and each criminal offense. General theories, if they emerge, will follow this balanced interdisciplinary trend.

Dichotomy between rural and urban areas

Density and size of the community. Statistics for the United States show a significantly high rate of crime in urban areas as against rural areas, and this is generally true for most industrial and urbanized countries, including, as far as information is available, eastern Europe. The reasons frequently put forth to explain this are that rural communities tend to be more homogeneous in moral behavioral patterns; that rural families are more closely knit because of tradition and because of the work they perform together; and that the rural community is less complex in organization, enabling the church and the school to exert the necessary external control. In addition, police control in rural areas may be exerted more informally and less strictly, and the police often have been considered as part of the community and not as an external agency. Within the city itself, some interesting patterns emerge. Studies in the 1920s and 1930s in America found that delinquency areas could be characterized by their proximity to industry and commerce, physical deterioration, decreasing neighbourhood population, nonownership of homes, the presence of ethnic minorities, and the lack of community facilities. These observations gave rise to the theory that where populations are constantly shifting, people lose interest in the appearance and moral reputation of their neighbourhood. This implied an alienation from one's surroundings, so that the small neighbourhood ceased to act as a controller or regulator of individual behaviour the way it normally does in the village or rural community. Thus, some important controls were removed and crime and delinquency erupted. This theory, however, is now criticized by those who assert that modern suburbs, with their emphasis upon privacy and anonymity, can hardly be said to promote neighbourliness, and although it is often argued that alienation in these areas is even more extensive, their delinquency and crime rates are generally lower than those of inner urban areas. One reason sometimes put forth to explain this anomaly is that the police may be more likely to let middle class suburban offenders off with a warning, whereas those offenders in the slum areas will more likely be arrested. Research has suggested that there are differences in the attitudes of the police and in enforcement policies between slum and middle class areas.

The criminal population. *Sex.* Throughout the world, most crimes are committed by men. Very roughly, the ratio of arrests for murder is six men to every one woman; for assault seven to one; robbery 22 to one; and burglary 30 to one. Although many interpretations of these ratios are possible, current opinion is that they attest to the importance of social rather than personality factors in crime. Another difference between male and female criminality is that women tend to commit their crimes in isolation rather than in the company of a group, as is the tendency for men. In the case of violent crimes, women tend to commit most of them upon members of their own family.

Age. Offenders committing serious crimes are most frequently under the age of 25. An important observation is that the age group 11 to 17 accounts for one-half of those arrested, and the group under 25 accounts for three-quarters of the arrests. On the other hand, two-thirds of those who commit homicide are 25 and over.

Race. Most crime statistics in the United States indicate a higher rate for blacks than for whites over a wide range of offences. Crimes that show the least difference in incidence as between blacks and whites are those that blacks have less opportunity to commit because of their minority status—embezzling and counterfeiting, for example. Two opposite interpretations of this difference are sometimes offered: first, that it represents the discriminatory attitudes of law enforcement officers, who are more prone to arrest blacks than whites; and second, that it is because the black is somehow predisposed to crime. Both of these explanations are simplistic, and, in light of the findings of geneticists, and because of the semantic confusion that always surrounds the interpretation of the findings of geneticists and environmentalists, both views must be regarded as definitely unproven.

Crime and discrimination

Other alleged explanations of the high crime rate among blacks in the United States relate to the vicious circle of discrimination; the black in America, and other minority groups in America and elsewhere, generally suffer from lower wages and more unemployment than others and are forced into less desirable jobs, especially as a result of being kept out of white-dominated labour unions. Because of segregation, the quality of black education is generally inferior, which leads directly to lack of job opportunity. It should also be recognized that because blacks have been forced to live in cultural ghettos they have built up patterns of child rearing and moral teaching of their own. The use of physical punishment is extensive in teaching children right from wrong, so that certain violent ways of behaving are learned at an early age. Further, the vast majority of crimes committed by blacks are committed against other blacks, not against whites. In a study of homicide in Philadelphia, only 6 percent of homicides were found to be cross-racial.

Class structure. Traditional Marxist theory attributes crime to the frustration of the working class resulting from constant exploitation by the upper class. As a general statement it may be said that economic conditions and poverty have *not* been found to be the critical factor in the cause of criminal and deviant behaviour, with the possible exception of suicide. A number of studies have, however, shown that such crimes as robbery and other property offenses do increase when there is a large

amount of unemployment. Other studies have shown that crime increases during periods of economic prosperity.

Statistics on apprehended and punished offenders probably overrate the relationship between economic conditions and crime because the poor are more likely to be identified and arrested, are less able to raise bail, and may be assisted by inexperienced lawyers. But even when these factors have been allowed for, it is clear that crime most certainly occurs more frequently in the slums.

Among the specific variables that have been found to have some relation to delinquency are a higher proportion of working mothers, irregular employment, the existence of child labour, and lack of adequate child supervision. Distribution according to social class can occasionally be discerned for certain offenses. Studies in England have shown that homosexuality is more common in higher or professional classes. Some crimes, such as embezzlement and art fraud, require certain levels of educational and social status in order for them to be committed.

Criminal subcultures. A concept currently gaining wide acceptance and for which empirical evidence may be growing is that of the criminal subculture, in which a different set of rules or norms for conduct arise, along with different child-rearing patterns, family relationships, and traditions. Thus the subculture of violence may be one in which the quick resort to physical aggression in the face of frustrating circumstances has come to be the pattern of reaction acquired by the member of the subculture at a very early age through social learning and identification with criminal models. Certain other types of crime, such as safecracking, require that offenders spend a good deal of time in a hardened criminal's company in order to learn the skills of the trade. A subculture may thus arise to foster the development of a pattern of norms of conduct quite at variance with those of the prevailing culture.

Crime among migratory groups

Early investigations of the mobility of the criminal population concentrated upon migratory groups—either within a country (rural to city migration) or internationally—and claimed to demonstrate a close relationship between crime and migration. It is now clear, however, that migration itself does not have a significant statistical effect upon crime. This has been demonstrated in studies of black migrant populations in the United States and of the migration from south to north in Europe. One of the complicating factors in such studies is that those who do migrate tend to be young, adult males, the group most prone to high rates of criminality anyway. In analyses of the higher crime rates for migrant groups, the difference is often entirely accounted for by the selective age factor.

Crimes committed in the company of others also have received special study. In Europe, studies appear to have identified a common type of relationship in which two persons of opposite physical and psychological character collaborate, not uncommonly as lovers or partners, in perpetrating crimes against third persons. Such crimes as infanticide, parricide, and abortion are common in this category. Other research has suggested that the influence of the group in causing a person to partake in crime may be substantial, so that a crime that an individual on his own would not have committed may be encouraged by group support. Two examples of this type of crime are gang rape and robberies that require extensive cooperation and planning.

Influence of the family. It is a well-established canon among many social scientists that the family is the primary agent for the social control of conduct. Family structure and values may influence an individual's tendency toward criminality in a variety of ways. Family disorganization, resulting from desertion, the death of a parent or other important family figure, and extreme tension preceding divorce, is often a significant factor. A higher proportion of delinquents as compared to nondelinquents is found to be illegitimate, and studies indicate that the backgrounds of habitual criminals frequently include an abnormal amount of parental marital discord. Multiple-problem families, which are usually very large and headed by parents of low intelligence, poor health, and low and irregular incomes, are also frequently held to generate criminal behaviours. There is considerable disagreement as to whether these families display a higher rate of crime because of their internal structure, tensions, and shortcomings or because they are invariably situated in the higher crime areas. Both aspects are clearly important, but what is of interest is that the worst types of problem families do not often produce the most serious types of criminals.

The family is usually one of the agents that transmits conflicts of cultures. American studies conducted soon after World War II suggested that crime and delinquency were extensively a problem of poor immigrants, and particularly of their American-born children. Broadly speaking, criminality tends to diminish with each successive indigenously born generation.

Some families may train their children to commit crimes, and there is evidence that this may especially be so among professional thieves. Studies in many countries have shown that a very large proportion of delinquents and criminals have delinquents or criminals as members of their household. But there are other conflicting findings. The question that so far remains unanswered is this: why do some members of a family become criminal while others do not?

The family and the process of socialization

Family disciplinary policies may be either too lax, too severe, or too inconsistent. American research has suggested that unsound discipline may be related to about 70 percent of criminal men. But it must be remembered that the process by which the child learns to abide by the rules and to control his inner impulses (socialization) is still little understood, and factors that are hypothesized to contribute to a breakdown of socialization must be considered as only tentative. Reservations aside, it would seem that in Western societies consistency of discipline is of utmost importance. The process of socialization appears to be an extremely delicate one, so much so that its success may depend quite heavily on factors difficult to define—such as the mother–child relationship, the father–child relationship, and associated psychological ramifications of early childhood experience. Accidental or chance events that may occur at critical stages of development may in fact play a major part. In addition, this complex of factors must be seen against a background of variables related to country, social class, and fashion, which strongly influence child rearing. This makes the location of the cause of crime in the breakdown of any one aspect of the socialization process an extremely difficult, if not impossible task, particularly if general "laws" or principles are sought.

The family may also set the scene for later emotional disturbances. The potential offender may be revengeful because he feels that his parents were unjust to him; he may have felt deprived in many different ways (money, gifts, love); he may develop a poor self-image because of constant parental rejection. These are only a few possibilities, and psychoanalytically oriented criminologists propose many more explanations of criminal behaviour in terms of early family relationships.

Some modern theorists now argue that the role played by the family in transmitting the rules and values of society is becoming less and less meaningful as society grows more complex, as people become more alienated from one another, and as formal agencies outside the family increasingly begin to take over the responsibilities that were traditionally entrusted to the family. An example of this may be in the realm of sex education, in which the school in some countries is beginning to assume responsibility for imparting knowledge and values concerning sex behaviour. Furthermore, the enormous effect of the mass media in conveying values to both children and adults, often in a conflicting and indiscriminate pattern, may tend to slant responsibility away from the parents as the prime transmitters of values.

Educational factors. Does education help to prevent crime? The most pessimistic view is that a better education simply enables a criminal to be a more effective

criminal. It was once believed that education for all would help to eliminate crime. Despite the fact that this hope has not been realized, there does appear to be a clear relationship between educational level and crime. Studies have generally suggested that prisoners as a group have poorer education and higher illiteracy rates than nonprisoners. Explanations of this relationship are many. Earlier theories suggested that the criminal achieved poorer educational levels simply because he was dull and lacking in intelligence. It has since been recognized that this explanation is an oversimplification, for while it is true that prisoners generally score lower on intelligence tests, the built-in social class bias of such tests is now recognized. Furthermore, inadequate schooling, commonly a factor related to performance on intelligence tests, is also closely related to criminality.

Relation of education and crime

There is considerable evidence that delinquency and criminal behaviour first become visible to society in the school setting. Large numbers of repeating offenders have been found to have had unhappy school careers and truancy records. Those who consider the school more important than the family in teaching children self-control and obedience would presumably charge the school with failing to fulfill its functions as far as these particular children are concerned. It seems at least arguable that, while schools have been encumbered with more and more responsibilities of this kind, the extent to which their organization, curricula, and teacher attitudes have changed to meet the challenge is minimal, particularly in urban settings. The difficulty may be that many teachers consider the central function of the school to be to impart knowledge, and the transmission of anticriminal attitudes to be only a by-product of secondary importance. Furthermore, the way in which the transmission of social values may be most effectively achieved in the school setting awaits further study. This area of criminology has received little attention from researchers to date.

Victimology. Any significant study of crime in society must include an analysis of the victims of crime, as well as of the perpetrators. In certain specific crimes the distinction in law between criminal and victim, guilty party and innocent, is difficult to maintain. This applies especially to some crimes of violence which commonly occur between a victim and offender who are closely related to each other. In England, for example, two-thirds of female victims of violent crime were attacked by relatives. An extensive study of homicides in the United States revealed that half the victims had had previous arrests, and a very large number were either close friends or relatives of their assailant. A quarter of the homicides were found to have been precipitated by the victim. Homosexuals and other sexual deviants appear to be especially prone to violent attacks by others, quite often their partners. In the field of economic crime, a similar relationship between victim and offender is apparent. Large department stores that display goods on open shelves clearly invite their theft; art forgeries flourish because of the speculative habits of the victims.

In surveys conducted in the 1960s in the United States, 20 percent of the sample questioned reported that they had been victims of crimes. This represents a crime rate twice that reported by the FBI. Blacks were found to be twice as likely to be victims of both violent and property crimes. More striking was the finding that only 60 percent of the serious crimes were reported to the police. The main reason that people gave for this was that they thought that the police could do nothing.

England, Canada, and some states of the United States have introduced new legislation to compensate victims of crimes, especially violent crimes. Other countries have enacted, or are enacting, similar legislation. In the light of the information above, it can be seen that the assessment of the victim's role in the criminal episode is extremely complex and may often make it difficult to apportion responsibility and decide on compensation.

Compensation for victims of crime

CRIMINAL CLASSIFICATIONS

Normal and abnormal offenders. The concept of normality is a very elusive one, and some recent schools of thought have suggested that whether a person is normal or abnormal is merely a question of definition. Only recently has it come to be accepted that the diagnosis and labelling system used by mental health specialists is heavily laden with the social class and cultural values of the definers. This includes the moral judgment, most often implicit, of right and wrong, so that the apparently simple labelling of a person as "sick" may also imply that he is "bad." The proponents of this perspective tend to view antisocial behaviour not as a sickness but simply as behaviour that has come to be defined that way. The implication, to some extent substantiated empirically, is that criminals are not all that different from noncriminals, except for the fact that they have been processed through a system of criminal justice. The evidence is, in any case, conflicting. A number of studies have indicated no significant difference between the personality traits of prisoners and nonprisoners, and no definable personality makeup of delinquents compared to nondelinquents. Some accumulated evidence does suggest that traits of aggressiveness and feelings of inadequacy are common features in delinquent personality. But beyond this, there is really no agreement on what psychological factors differentiate criminals from noncriminals, and the reason for this could well be that there is no great difference between the two groups.

In advanced Western societies, more and more behaviour that was once defined as criminal is now often seen as behaviour requiring treatment, usually of a psychological type. Various legal codes recognize a number of forms of such mental abnormality as epilepsy, schizophrenia, mental defectiveness, and paranoia.

Offender careers. It is becoming increasingly apparent that such terms as delinquent or criminal have little real meaning. But, by taking into account the career of the offender, the support that his behaviour receives from his group, the correspondence between criminal behaviour patterns and legitimate behaviour patterns, and the societal reaction to criminal behaviour, one can construct a number of useful types of behaviour systems. The criminal career is made up of the social roles that the offender plays, his self-identification with crime, his self-concept, his relationship with others, and the degree to which criminal behaviour has become a part of his everyday life. Some offenders may have association almost entirely with other criminal types, whereas others may not. Some typical offender career patterns are those of (1) *the occasional offender*, who does not have a conception of himself as a criminal, does not play a criminal role, and does not have much group support for his activities; (2) *the habitual petty offender*, who has a long criminal career, but is generally not very sophisticated in his activities and tends to be more easily caught; (3) *the conventional criminal offender*, who displays a steady progression from juvenile gang delinquency to adult criminal behaviour of a more serious type. Such career offenders continually acquire new techniques of committing crime along with new self-justifications for committing them. Studies have shown that these career offenders usually have intimate associations with others of a similar type from whom they learn and to whom they impart the skills and justifications for their acts. Typically, these crimes reach their peak during the 20s, and terminate in the early 30s of the life history of the offender; (4) *the professional offender*, who has the most highly developed career and level of sophistication. Professional criminals are generally looked up to by conventional criminals. They engage in a variety of highly specialized crimes, such as pickpocketing, sneak-thieving from stores, banks, and offices, stealing from jewelry stores and hotel rooms, passing illegal checks, and extorting money from others who are engaged in illegal activity. These criminals are not usually connected with any "heavy rackets" or organized crime, although some, such as safecrackers, may be. In comparison to other criminals, an extremely high degree of consensus exists among professional criminals, who develop common attitudes toward themselves, toward their crimes, and toward the police, who are seen as the "common enemy." There is quite definitely a code

Typical criminal career patterns

of honour among them and a highly specialized argot, developed and transmitted from one generation to another.

Occupational offenders and white-collar criminals. Offenders in this category include businessmen, politicians, government employees, doctors, and others who commit crimes that are closely related to their work. For occupations of relatively high status, the term white-collar crime is often used. Violations of the law by businessmen include those related to receiverships and bankruptcies, restraint of trade such as monopoly, illegal rebates, infringement of patents, trademarks, and copyrights, and misrepresentation in advertising. Norms concerning food and drugs sale and antipollution regulations also may be violated. Employers have been found to violate laws regarding wages, hours, and public contracts. Politicians and government employees may obtain illegal financial gains by furnishing favours or confidential information to business firms and obtaining illegal commissions. In the medical profession, doctors may give illegal prescriptions for narcotics and give false testimony in accident cases. Lawyers may misappropriate funds and secure perjured testimony from witnesses. Embezzlement is a common form of occupational crime.

The major difference between occupational offenders and other offenders lies in the offender's conception of himself. The occupational offender rarely sees himself as a criminal. He may, in fact, see himself as a respectable citizen. In addition, the usual middle or high social status of these offenders is such that it also makes it difficult for the public to conceive of them as criminals.

Significance of white-collar crime

The consideration of white-collar crime in criminology has introduced an important balance to an otherwise distorted picture that is presented by the exclusive study of conventional crime in society. Perpetrators of white-collar crime, who are usually better educated, are less often sentenced to imprisonment and are often not prosecuted formally by the system of criminal justice. Instead, they may be disbarred from their profession or receive fines that they are usually able to pay more easily than are conventional criminals. In purely monetary terms, it has been estimated that the cost of conventional crime is trivial compared to the millions involved in the crimes of fraud and embezzlement. Studies of the activities of large corporations have suggested that their illegal activities may be highly organized and persistent and, as a whole, show clear disregard for the government, the law, and the people who administer it. But the concept of white-collar crime has been strongly criticized from some quarters because it has expanded the concept of crime beyond what has conventionally been considered the proper area of study for criminology—the study of the overt acts of convicted offenders. These critics also insist that there is a basic incongruity involved in the proposition that implies that a community's political and business leaders may also be criminals. Proponents of the concept of white-collar crime are accused of being unrealistic and of having diverted attention away from the serious nature of conventional crime. This controversy, in fact, relates to the overlapping between criminal law and civil and administrative law, and to the question of which kinds of conduct are best controlled by the application of the criminal law, and which kinds by other means. There has been a tendency in most countries to delimit the range of conduct with which the system of criminal justice must be concerned. The trend in research, on the other hand, has been to broaden the concept of criminal behaviour, or, more accurately, to include it and many other types of behaviour under the more general rubric of "deviant behaviour."

Organized crime. The study of organized crime has been conspicuously neglected in British and continental criminology. A few studies in Germany and Scotland have suggested that the activities of well-organized professional gangs are extensive throughout Britain and continental Europe. Such gangs engage in well-executed robberies (such as the great train robbery near Aylesbury, England in 1963), in organized smuggling, art frauds, the drugging of racehorses, and so forth. Occasionally, it has been noted that gangs of adult criminals are often active in times of political and social upheaval. The robber-knights of the German Middle Ages, the gangs of coiners in the 17th century, and the secret societies that terrorized the population of Germany and other continental countries during the 18th century are examples. Adult gangs flourished in France during the Revolution.

The Mafia

But organized crime is most prevalent in Western countries. Though it is widely believed that organized crime was imported into the United States via the immigration of Mafia members from Sicily, a number of writers have questioned this. Undoubtedly, many members of the Mafia did migrate to America, and many notorious criminals have been Sicilian. But it also should be noted that there are in America many other ethnic groups involved in organized crime. The proportions of Jews and Irish appear to be quite large.

The Sicilian Mafia is described as not so much a secret society as a "habit" of tackling all enterprises in an indirect and underhanded manner, by mediation, by recommendation, and by other dubious forms of "influence." It provides redress against evildoers but also protects criminals from the law and encourages their criminal behaviour. In Sicily it has been shown that, for a time, the Mafia was in fact able to safeguard the interests of the common people more effectively than a geographically remote and uninterested government. The Mafia enjoyed a revival just after World War II, but at present an organized, large-scale war is being waged against it by law enforcement agencies.

There are certain similarities between the Mafia and its American counterpart, but these may be common features of all organized criminal groups. The conspiracy of silence, the separate and highly effective system of group "justice," and the protection racket are typical not only of the Sicilian and American Mafia but have been present in the past in organized crime in India and China.

In the United States the term organized crime has come to mean something quite unique and is used to refer to large-scale rackets, extensively organized, carried on over long periods of time, and relying on political corruption for protection from the law. Crime syndicates have abundant capital, may have control of factories and warehouses, and control their own managerial personnel and employees. They operate rackets in drug traffic, liquor sales, protection, gambling, and prostitution.

Non-American students of organized crime have tended to be dubious that crime on such a scale exists in the United States. Although they usually concede that organized crime exists on a local basis, they are often skeptical about the claimed immense countrywide organization. On the other hand, "conventions" of notorious criminals in America do occur, and are reported by the press.

Homicide and crimes of violence. The majority of homicides are committed within the family or between close friends. About half occur as the result of trivial altercation. In the United States, the method is equally distributed among gunning, stabbing, and beatings. In contrast, guns are rarely used in England: assailants often use blunt objects that happen to be on hand at the time. The age of homicide offenders in America is predominantly 20 to 24, but in England it is a little higher, though both countries report an increasing use of violence by youthful offenders. Many early studies have reported that participants in homicide in the United States do not generally have previous criminal records, but one thorough study reported that nearly two-thirds of the offenders had prior arrest records. In England, more than one-half of convicted violent offenders have had previous records of arrest for nonviolent offenses; and 20 percent have had previous convictions for crimes of violence.

Of all crimes, homicide has the highest "clearance rate" (that is, more murders are solved than any other crime); the rate is around 90 percent in most countries for which statistics are available. The conviction rate, however, appears to be somewhat lower. About 80 percent of American blacks who are tried are convicted, as compared to a 60 percent conviction rate for whites. A low conviction rate for crimes of violence is also reported for London.

Drug addiction. The extent to which the taking of drugs is unlawful varies enormously from one country to the next, but severer sentences seem to be becoming more common in Western nations.

Certain types of crime are committed more often by narcotic addicts, mainly because they need money to support a habit made impossibly expensive by illegal trafficking and blackmarketing. But it is also possible, at least in some cases, that taking these drugs may deter a person from committing certain types of crime, especially those of violence. In the light of a very large amount of research evidence, criminologists have advocated that some drugs, marijuana in particular, should not be outlawed, that penalties should not be too severe, and that the legalization or controlled distribution of drugs to stamp out illegal trafficking should be introduced.

Intoxication and alcoholism. In many countries, drunkenness itself is not an offense, but there are usually many other kinds of conduct that are closely related to it, such as disorderly conduct or disturbing the peace. In England and Wales, there has been a decline in convictions for these offenses, but in the United States the situation is different. It has been estimated that of about 70,000,000 Americans who drink, 4,700,000 are alcoholics. A substantial proportion of these alcoholics is made up of skid row drunks, who, in an endless procession, are brought before the courts, given a few days jail, and released, only to be picked up again for vagrancy, disorderly conduct, or some other public nuisance offense. It is generally agreed that these people are not helped by incarceration, but there appear to be no ready solutions to the problem.

Alcohol and driving offenses

The relationship between drunkenness and motoring offenses is also complex, but it is apparent that many serious accidents and injuries on the road involve persons driving under the influence of liquor. Recent trends are toward a definite increase in the severity of punishment for drunk driving offenses. In cases of criminal homicide, American research suggests that in almost half the cases studied excessive drinking prior to the offense was evident in either the victim or the offender or both. Research in England suggests that the majority of alcoholics do not get into serious conflicts with the law, though research in Scandinavia suggests the opposite.

THE CONTROL OF CRIME

Police and the public. The attitudes and policies of law enforcement agencies vary widely between countries. The behaviour of police depends in large part upon their perception of how the public sees them. Who will be arrested depends on the policeman's attitudes, on the enforcement policy of his department, and on the area in which he is working. Observation of policemen on duty affirms that their job is extremely difficult and sometimes dangerous and, if done conscientiously, requires a great deal of skill and quick thinking. The decision as to whether or not to make an arrest depends upon an enormous number of factors that the policeman must take into account, often very quickly and in the heat of the moment. It is understandable that many police departments, in the face of rising defiance of police and questioning of their competence, have become organized along quasi-military lines. This provides the policeman with a definite, though possibly unwritten, set of rules that he may, with his superiors' approval, rigidly abide by when dealing with offenders. The result is that the policeman may remain completely detached from the public with whom he deals, interpret all situations according to the rules, and often appear callous or rigid to the public. The opposite type of police organization is one in which the policeman is allowed wide freedom of discretion and thus may enforce the law to whatever degree he thinks is suitable to the situation. Although this may sound like a more reasonable approach, the freedom may easily lead to the law enforcement officer exercising his own personal prejudice. In addition, the situation is complicated by the fact that the police departments that are organized along military lines are more formally structured, offer more professional training, and offer greater chances for promotion. Thus, it happens that those police with the most training, and the best chance to make law enforcement a career, are allowed less discretionary powers.

Public attitudes toward police

Though there exists no formal research evidence, it is probable that more police corruption exists in a country such as the United States than in a country such as Great Britain. It would seem that the American policeman is less socially esteemed than his British counterpart and is likely to use violence more often, especially when his status is affronted by the offender. It appears to be a common factor with police in many countries that they remain socially apart from the rest of society. Whether or not this is by choice is still a matter of speculation. Studies have shown that the informal social and recreational life of policemen is almost exclusively carried on among themselves and that they feel themselves set apart from the rest of society. Despite this, public opinion polls in America suggest that the majority of people think that the police do a good job and hold them in higher esteem than the police themselves realize.

Another common feeling that police have is one of ineffectiveness. They commonly adopt the view that they are fighting a war against crime, yet find that of the many offenders they arrest, only a very small proportion are convicted and found guilty. In the United States, the President's Commission on Law Enforcement and Administration of Justice found in 1965 that for 2,780,140 crimes reported, only 727,000 persons were arrested and that of these, 290,000 were never charged. The courts in many countries are hopelessly clogged with cases, and it is common for an offender to wait months, and sometimes years, for his trial. A practical suggestion has been to set up a special court to deal solely with offenders who have been assessed as dangerous and hardened criminals, so that their cases may be dealt with speedily, to prevent their being free in the community on bail while they await their trial. There is some evidence that a disproportionate number of crimes of violence are committed by hardened criminals awaiting trial on bail.

Preventive detention—which is the denial of bail or the imprisonment of a person accused of an offense on the grounds that he is too dangerous to be free, even though he has not been found guilty—raises very difficult issues of individual rights. The only solution to the problem would seem to be to provide an efficient and swift system of bringing people to trial. Underlying this whole problem are two basic contradictory attitudes toward crime control. One school of thought, assuming that persons having contact with the police are more likely to be guilty than innocent, advocates an efficient control of crime by establishing wide freedom to arrest on suspicion, fewer restrictions on searching and entering people's residences, more freedom to use such methods as wiretapping for investigation and subsequent evidence in court, and more freedom to obtain confessions. In short, this view is oriented to seeking out the wrongdoer at almost any cost. Although it is recognized that mistakes may occur, it is obvious that with greater freedom to investigate, the police would catch more criminals and be likely to make less mistakes because they would have more information. In the long run, the argument goes, this method is fairer, especially because it would eliminate such doubtful procedures as negotiated pleas, or "bargain justice," in which the offender is encouraged to plead guilty to a lesser offense if the prosecutor agrees not to press a more serious charge.

Proposals for extending the power of police

But the contrary view argues that individual rights must be preserved at all costs, and insists that the presumption of innocence must be maintained right through to the end of the trial. The individual is conceived of as needing protection from the police, who are seen as representing the potential tyranny of the state. In general, this view tries to make obtaining a conviction more complicated, usually on the rationale that by doing so there will be less chance of miscarriages of justice. This approach sounds highly humanitarian and liberal and more protective of individual rights, but it may be that the disorganized and complicated machinery of police, courts,

and corrections is closely related to it. It has been argued that the insistance upon individual rights may lead to such injustices as long imprisonment while awaiting trial, or the increased risk of crimes committed by those on bail awaiting trial. The solution to this paradox appears to be issuing from the increasing agreement among criminologists and administrators that there exists an identifiable group of hardened criminals, who have long records, are often violent, and are the only types of criminal whom society needs to imprison because of their dangerousness. Fortunately, it is believed to be likely that they comprise only a very small proportion of the present criminal population. The future trend, therefore, may be to infringe less upon the individual's freedom by imprisoning only those offenders assessed as dangerous, but, on the other hand, to increase the efficiency with which these people are sought out.

Corrections. Though there is much controversy over the effectiveness of prisons and other forms of treatment-oriented sanctions, the resigned opinion of leading authorities on corrections now is that most forms of punishment and treatment have little or no discernible effect upon the rate of crime. The only possible exception may be fines for traffic offenses.

Re-evaluation of penological techniques

There is considerable evidence that prison may play a role in extending criminal behaviour rather than limiting it. The study of total institutions (that is, any institution that creates its own little society and system of values completely isolated from the rest of society, including monasteries, boarding schools, and mental asylums, as well as prisons) has shown that prisons sometimes place insurmountable obstacles in the way of rehabilitating the offender. The growing belief today is, therefore, that an offender cannot be taught to live in society by separating him from it. There is now a strong move in some countries to create "halfway houses" and other similar institutions, in which the offender may receive supervision but still work and live in the community.

The assumption at the bottom of all these studies of effectiveness of treatment is that in some way or other the offender is abnormal or significantly different from other people and that this "maladjustment" must be changed. But research now suggests that offenders may not be all that different from nonoffenders, so that what exactly these treatment programs seek to change is questionable. Certainly, some offenders can readily be identified as abnormal by many established psychological theories. Most, however, cannot (see also PRISONS AND PENOLOGY).

II. Patterns of juvenile delinquency

THE ANALYSIS OF DELINQUENCY

Sex and age distribution. In Western society, the age during which delinquent acts are most frequent is around 14 and 15 years of age, and this appears to hold true for both boys and girls. The peak may vary according to the class of offense. It has been found that the peak of 14 years is mostly comprised of juveniles involved in stealing or minor theft; however, a peak at 16 to 17 is found for such offenses as malicious damage, vandalism, illegal use of a motorcar, carrying an offensive weapon, and assaults against police. Although there is some evidence for an increasing amount of delinquency in late adolescence, it seems fair to say that delinquency for the majority is mostly an episode or phase through which some young people pass. Delinquent acts begin at about age 11, continue up to the peak of 15, then decline quickly at about 20 to 22 years of age. The majority of delinquents do not commit new offenses after this age. It seems evident that a younger child is likely to be warned rather than prosecuted by police, which may explain the rise to the peak of 15 as the age more likely to invite prosecution. Theories to explain the termination after adolescence are many and conjectural. It may be explained by maturation, marriage, obtaining a steady job, or cessation of the emotional turmoil that sometimes accompanies adolescence. Though the evidence is conflicting, it is likely that those who begin committing offenses at an early age are more likely to continue them into adulthood.

Delinquency is also predominantly a male activity. In the United States, approximately 80 percent of delinquents are boys, and similar high ratios of boys to girls are apparent throughout Europe and Japan. Studies also suggest that there is a slightly higher female rate for blacks in America, and that the ratio of girls to boys for both races increases by about 15 percent from age 10 to 16.

Delinquency and social class

Lower and middle class delinquency. The relationship between social class and delinquency is difficult to assess. When school children are asked confidential questions about offenses that they have committed, one usually finds a very large number of offenses distributed evenly across all social classes. When one considers the statistics of convicted delinquents, however, the picture is fairly clearly one in which the lower classes are overly represented.

There are many theories that seek to explain delinquency in social class terms. One of these suggests that the delinquency of working class boys is a response to disappointing experiences in high school, when the delinquents were unable to meet the qualifications for social achievement in an essentially middle class school. Thus, working class boys coalesce into social groups and compensate for one another's failure by disavowing the values of the middle class. The result is academic indifference, rudeness, truancy, and property destruction. Alternatively, it is suggested that working class boys must adopt certain modes of behaviour and live up to the cultural values of their group, such as living dangerously, displaying masculine toughness, and showing contempt for external authority. Another theory argues that working class boys will become delinquent because they are required by modern society to pursue certain dominant goals, such as money and other status symbols, but are not given the opportunity to acquire them because of poor educational facilities, poor housing, poor job opportunities, or other discriminatory factors.

A basic disagreement among theorists is whether the lower class delinquent is mainly oriented to the goals outside his class (that is, those of the middle class) or to the values built up over generations within his own social class. The current prevailing trend is to postulate a degree of ambivalence about middle class and lower class values and to suggest that the delinquent has difficulty understanding what values his peers adhere to. It is apparent, in any event, that there are a large number of values that delinquents hold in common with many other people in society.

The theory of masculine protest

Middle class delinquency does occur but has been comparatively little studied. The most popular theory to date has been that of "masculine protest." Briefly, the idea is that children very early in life adopt as an overall model for their own behaviour the parent with whom they have the closest and most continuous relationship. Because of the isolation of the middle class family in suburbia, because of its smallness, and because of the daily separation of the working father from the home, the mother becomes the dominant figure. This system works well with girls because they are expected to behave in a feminine way. Boys, however, are thrust into conflict because society expects them to be "manly" and yet the model they are forced to adopt early in life is that of the female. The boy must, therefore, assert his masculinity, and, because his mother has represented from early in life what is "good," his delinquent behaviour acquires the function of denying femininity and asserting his masculinity. This theory might also explain why there is a much larger number of boy delinquents than girls.

The delinquent gang. Juvenile gangs are common among both delinquent and nondelinquent boys, though delinquents tend to be members of gangs more often than nondelinquents. It was earlier thought that delinquent gangs were highly disorganized groups whose members behaved impulsively, and in which there was no particular goal or organized direction for conduct. Thus, it was popularly thought that organizations, such as the

Boy Scouts, could channel this impulsive behaviour into constructive activity, thus preventing delinquency. Students of gangs now recognize that there are many different types of delinquent gangs, and many of these are highly organized and structured. Many urban gangs have been found to have a long history, with many behavioral patterns handed down from one group of gang members to the next. Types of gangs that have been proposed are the following: (1) "Criminal" gangs existing in slum areas in which there is a strong tradition of crime, so that adult criminal persons serve as models for the gang members. The activity of these gangs is clearly organized and directed toward such predatory crimes as extortion, theft, and burglary. (2) "Conflict" gangs, commonly existing in new urban housing developments, where a criminal tradition has not yet arisen. Members of these gangs, oriented, it is assumed, to the values of the middle class culture and not to their own, use other adolescents as their models. Because their situation is a conflicting one, they are more likely to "act out" their frustration in acts of violence and vandalism. (3) "Retreatist" gangs contain members who, again living in slums, use drugs and other "escape" experiences because they are unable to find any path, either legitimate or illegitimate, to obtain success. It has been noted recently, however, that retreatist groups are in fact predominantly a middle class phenomenon. (This typology is not applicable, however, to the study of gangs in all modern countries.)

A typology of gangs

Some evidence suggests that the explanation of gang delinquency does not lie in forces outside the group (such as middle class values), but rather within it, in the typical striving for status by adolescent gang members within the group. Delinquent boys have been found to value their peer-group status very highly. In other words, membership in the gang provides much needed satisfaction that cannot be obtained elsewhere in society's established institutions such as school and church. The membership not only offers rewards of status recognition but also excitement. Also, the development of an adolescent "subculture" with values and interests more and more removed from those of the rest of society, and often in conflict with them, accentuates the separation of groups that pronounced urbanization brings about.

An often neglected factor in delinquent behaviour is the role of the school in the delinquent's life. The majority of research points to the very unhappy school histories of a large number of delinquents. They have high records of truancy and poor school records and generally perform less well than nondelinquents. They display more problem behaviour at school, including social maladjustment, usually quite early in their school career. A relationship between deprived social and cultural conditions and school adjustment is well established, and what adds to the problem is that there are fewer manual jobs for lower class boys who are unable to succeed at school. They thus have no prospect of satisfactory adult adjustment through work. Furthermore, it seems that a great number of delinquent acts are committed by early school leavers, or dropouts, who have no job to go to. The compulsory age of school attendance continues to increase, though special curricula to deal with such young people remain only poorly developed.

The juvenile drug culture

Though it is popular to characterize the juvenile drug culture as retreatist, it is clear that many who smoke marijuana are certainly not retreaters but protesters and demonstrators. Whether one should apply the term juvenile delinquent to such persons who are subsequently prosecuted and convicted as a result of their drug taking or protesting is questionable. There is a danger that the important issues with which young people are concerned will be overshadowed by the application of such stigmatizing labels as "delinquent" or "criminal."

In conclusion, it may be fair to say that when criminologists are speaking of juvenile delinquents, they are speaking more of young adolescents, considerably less mature, both socially and emotionally, than today's usually sophisticated and well-informed youthful protester.

Home conditions. Some research suggests that delinquents more often have parents characterized by drunkenness or criminality, ineffective household management, economic difficulties, and low self-respect. Delinquents often indicate that their families are not concerned about their welfare. Disciplinary practices were more erratic and physical punishments more frequent for delinquents. Since the methodology used in these studies has been criticized from many quarters, the confirmation of these findings awaits further research.

In Great Britain, delinquents have been found to come from families characterized by difficult social relationships. The delinquent frequently expresses the wish to run away from home and often does so; parents often complain of difficulty in early relationships.

Some physical and biological differences between delinquents and nondelinquents have been found, but these studies are highly controversial. Delinquents are found to have sturdier bodies than nondelinquents. They are somewhat more "masculine," and they tend to be mesomorphic (a muscular, well-knit body type). No significant difference in physical health has been found. Some evidence has accumulated to suggest that delinquents as a group have a different personality structure than nondelinquents, but because of added difficulties in defining and measuring personality traits the findings remain difficult to apply to treatment. Delinquents are possibly more extroverted, more narcissistic, destructive, and sadistic, more impulsive, and less able to delay gratification. Current opinion, however, prefers to view delinquents as not substantially different from the normal population. It is, of course, also recognized that a small proportion of them may display psychological abnormality.

Prediction and prevention. Within very narrow limits, it now appears that delinquency can be predicted, probably at quite an early age. Although some criminologists claim to be able to do this at the age of three or four, it seems more reasonable that the delinquency of certain types of children can be predicted at least by the time the child has been at school for two or three years. Teachers have informally made such predictions for many years, and over the past decade some researchers have developed instruments to refine these intuitive judgments. Basically, three types of delinquency prediction have been developed: (1) that which uses certain "background" variables, such as quality of the neighbourhood, absence of the father from home, and alcoholism in the family; (2) that which uses variables concerning the parental attitudes toward the child, such as the varieties of discipline or the quality of the parent–child relationship; and (3) that which relies upon close observation of the child's behaviour in school and in the classroom.

Methods of delinquency prediction

Much dissatisfaction has arisen concerning research on the prediction of delinquency. The ethics of the procedure may be highly questionable. Current research suggests that the mere act of labelling another person with a stigmatizing label, such as "delinquent," may cause people in authority to treat the child as though he were indeed a delinquent, when in fact he may not be. But by this process, he may come to see himself as one.

There is a popular belief that if a delinquent could be treated at a very early age, his delinquency might be prevented. The proposition has never been conclusively tested. To date, no treatment program has been shown conclusively to have prevented delinquency, so that the value of early identification of the delinquency prone itself seems questionable. It is only very slowly coming to be realized that, as long as society has laws and has to impose them, delinquency can never be prevented. Rather, delinquency is an integral part of society. It is pointless to speak of the prevention of delinquency, when, by definition, the act of preventing it by law enforcement agencies is also the act of defining it. In addition, the idea that persons may be "treated" during childhood, so that they will be good, law abiding citizens, envisages the doubtful Utopia of a stagnant society in which everyone meekly conforms to the rules.

RESPONSES TO DELINQUENCY

Probation. Most Western countries today have a formally instituted probation system, usually supervised by

a local authority or by the state. Many countries, however, have retained a volunteer probation system alongside the formal system. Probation may be defined as a form of disposition whereby the court suspends the sentence for selected offenders and releases them conditionally upon good behaviour, subject to prescribed rules, and subject to supervision by officers of the court (probation officers).

Probation is most often granted to first offenders and for less serious offenses. Eligibility for probation is sometimes clearly prescribed by the law; at other times, it is left to the discretion of the court. Children's courts tend often to institute informal probation following investigation by a probation officer, who avoids bringing the case before the court. Recent cases have suggested that miscarriages of justice, either by way of uncontrolled, arbitrary decisions or by sentimental leniency, may occur under this procedure. The court also may call for a presentence report that requires the probation officer to investigate the offender's situation and to provide a recommendation to the court as to what disposition may best be applied to the offender.

The conditions of probation usually emphasize hard work, wholesome companions, moderate hours, and concern for financial obligations. These conditions make clear what is not to be done, under the threat of possible commitment to an institution. The relation between the probation officer and the probationer is, therefore, often an extremely difficult one, especially when probation officers are trained as social workers who have conceived themselves as helpers rather than as authoritarian figures. It is apparent that the typical probation officer has little chance of establishing any sort of close relationship with all but a few of his cases—partly, also, because of heavy case loads.

The effectiveness of probation

In light of this, it may be surprising that probation is a comparatively successful way of dealing with offenders, perhaps the most successful yet. Studies on probation show a success rate of 70 to 90 percent, though, these figures also may be questioned because of inadequately designed studies.

Types of treatment that have been tried under the probationary system and within institutions include psychotherapy, or intensive individual sessions between the delinquent and a therapist who seeks to establish a close, warm relationship with him, and group therapy, which involves guided discussion in small groups of delinquents, from which those seriously disturbed psychologically are excluded. It is believed that in group therapy the discussion of common problems assists the delinquent to accept more readily the restrictions of society, and enables him to understand his own behaviour more easily in relation to others. This method of treatment, which has been extremely popular since the late 1950s, is conducted mostly within institutions, often in conjunction with psychotherapy.

Not all juvenile institutions, however, use such methods. Many are oriented toward obedience and conformity. Such institutions maintain high levels of staff domination and negative sanctions and an undifferentiated view of their inmates. The Borstal System in England is typical of this type.

Common in many countries is the institution oriented toward re-education and individual development. The English Approved School is a good example. Attempts are made to change the inmates through training; and good staff–inmate relationships are fostered. There is an emphasis upon education, job training, and "team spirit." In England, many institutions have duplicated the "house" system of traditional upper class private schools, in an effort to change values and attitudes and to promote the acquisition of new skills and personal resources.

At the other end of the continuum, there is the institution that focusses upon the psychological reconstruction of the individual. It emphasizes gratifications and varied activities, with punishments few and seldom severe. Emphasis is upon the individual and his self-insight.

Juvenile aftercare. Juvenile aftercare takes place after the release of a youth from an institution, and at a time when he can most benefit from life in the community under the guidance of a counsellor. It is the juvenile equivalent to parole. Aftercare first began in the United States in the early 19th century but has become an integral part of the correction and rehabilitation of young offenders only in the past 15 years. The rationale recognizes the fact that institutions may have effects upon persons that are not conducive to adjustment in society. Some offenders may learn simply to be more delinquent by associating with other delinquents. Others may become highly dependent upon the institution and find it difficult to adjust to life outside. Most must face the inevitable stigma that attaches to confinement in a correctional institution. The majority of released juveniles receive very little aftercare supervision, mainly because of inadequate staffing and the great difficulty in coordinating community agencies.

Foster care. Foster care has become increasingly important in recent years. It has very often been used in a negative way in an effort to prevent a child from being sent to an institution. In the past, the ideal family for foster placement has usually fitted the stereotype of the well-adjusted, stable middle class family. There is some question as to whether or not it is realistic to expect delinquents, who very often come from lower social classes, to fully adjust to the quite rigid requirements of middle class conduct. It is usually advocated that the local community also must be well informed in order to cater specifically to children through the schools, churches, and youth groups. Without an enlightened attitude on the part of the community, the delinquent will be rejected and be unable to adjust.

Very little scientific research exists on the effectiveness of foster care. In fact, the whole area of the treatment and disposition of juveniles is in its infancy as far as scientific study is concerned. Generally, research on the treatment of delinquents has been largely anecdotal, relating the success of various treatments; and, where formal studies have been conducted, they have been poorly designed and thus have produced inconclusive results.

BIBLIOGRAPHY

General texts: H. MANNHEIM, *Comparative Criminology* (1965), the most comprehensive review of both European and American criminology; M.B. CLINARD, *The Sociology of Deviant Behaviour*, 3rd ed. (1968); a sociological account of all types of deviant behaviour, including crime; M.E. WOLFGANG, N. JOHNSTON, and L. SAVITZ (eds.), *The Sociology of Crime and Delinquency*, 2nd ed. (1970), a collection of key articles on theoretical criminology, crime statistics, scientific methods in criminology, patterns of criminal activity, and sociological theories of crime and delinquency; E.H. SUTHERLAND and D.R. CRESSEY, *Principles of Criminology*, 8th ed. (1970), a standard textbook on criminology, which develops the theory of crime causation known as "differential association," with a large section on white collar crime; S. SCHAFER, *Theories in Criminology* (1969), a detailed analysis of the development and status of criminological theories; L. RADZINOWICZ, *Ideology and Crime* (1966), a succinct analysis of the historical development of criminological schools; P.H. SOLOMON, JR., "A Selected Bibliography of Soviet Criminology," *Journal of Criminal Law, Criminology and Police Science*, 61:393–428 (1970), a briefly annotated bibliography on Russian criminology.

Crime index: Crime in the United States: Uniform Crime Reports, issued by the U.S. FEDERAL BUREAU OF INVESTIGATION (annual), a comprehensive statistical account of crime in America; *International Crime Statistics*, issued by the INTERNATIONAL CRIMINAL POLICE ORGANIZATION (INTERPOL), an annual publication of crime statistics of approximately 100 different countries; W.A. LUNDEN, *Statistics on Delinquents and Delinquency* (1964), a competent analysis of delinquency statistics, mainly of the U.S., but also from a number of European countries; COUNCIL OF EUROPE, *The Index of Crime* (1970), a collection of three papers examining the theoretical and practical problems involved in the application of a research index to the collection of criminal statistics by administrative agencies.

Social variables affecting crime: W.A. BONGER, *Criminalité et conditions économiques* (1905; Eng. trans., *Criminality and Economic Conditions*, 1967), the classic Marxist statement on the relationships between crime and poverty; H. VON HENTIG, *The Criminal and his Victim: Studies in the Sociology of*

Crime (1948), an analysis of the relationship between the criminal and his victim; M.E. WOLFGANG and B. COHEN, *Crime and Race* (1970), a balanced account of the meaning of the relationship between crime and race in America; S. SCHAFER, *The Victim and his Criminal* (1968), an updated discussion of victimology; A. AICHORN, *Verwahrloste Jugend*, 2nd ed. (1951; Eng. trans., *Wayward Youth*, 1951), a classic study applying psychoanalysis to the treatment of the delinquent; A.V.S. DE REUCK and R. PORTER (eds.), *The Mentally Abnormal Offender* (1968), a symposium report containing the most up-to-date statement on the mentally abnormal offender, including bibliography: H.J. EYSENCK, *Crime and Personality* (1964), and J.M. MCDONALD, *Psychiatry and the Criminal: A Guide to Psychiatric Examinations for the Criminal Courts* (1958), two books covering the field of psychological and psychopathological aspects of criminal behaviour and of the legal implications of mental abnormality in the disposition of offenders.

Homicide and violence: F.H. MCCLINTOCK *et al.*, *Crimes of Violence* (1963), an analysis of statistics on crimes of violence in England and Wales; Reports of the National Commission on the Causes and Prevention of Violence, detailed studies of violence in America covering its history, political violence, and criminal violence, giving an overview of sociological and psychological theories; M.E. WOLFGANG and F. FERRACUTI, *The Subculture of Violence* (1967), an attempt to integrate many different perspectives in criminology, within the frame of a subcultural hypothesis.

Criminal careers: H.S. BECKER, *Outsiders: Studies in the Sociology of Deviance* (1966), the basic statement of modern deviance theory, with an account of the lives of dance band musicians and how severe drug laws came about in America; CHIC CONWELL, *The Professional Thief* (1937), a classic of the field; E.H. SUTHERLAND, *White Collar Crime* (1949, reissued 1960), another classic text, which has widely influenced current criminological theory; D.R. CRESSEY, *Theft of the Nation* (1969), a recent account of Mafia operations.

Control of crime: J.Q. WILSON, *Varieties of Police Behavior* (1968), an appraisal of three American police departments, each representing distinctly different kinds of orientation to their jobs, from the "watchman" style to the military model; M.P. BANTON, *The Policeman in the Community* (1964), the authoritative statement upon the relationship between the policeman and the community in Britain.

Delinquency: D.H. STOTT, *Delinquency and Human Nature* (1950), a proposed typology of maladjusted families and of types of social maladjustment displayed by children at school, based on the study of many delinquents and their families; H. MANNHEIM and L.T. WILKINS, *Prediction Methods in Relation to Borstal Training* (1955), the most sophisticated study on the prediction of delinquency; S. and E. GLUECK, *Delinquents and Nondelinquents in Perspective* (1968), a summary of early work on juvenile delinquency, plus the report of a follow-up study of 500 delinquents who are now adults; D.J. WEST, *The Young Offender* (1967), a comprehensive account of theories, incidence, and treatment of the young offender.

Delinquent gangs and subcultures: F.M. THRASHER, *The Gang*, 2nd ed. (1936, reissued 1963), the early classic study of gangs, which emphasized their disorganization and the impulsiveness of their members; D.M. DOWNES, *The Delinquent Solution* (1966), the most critical appraisal of the theories of delinquent subcultures to date; J.B. MAYS, *Crime and Social Structure* (1963), a presentation of the thesis that delinquency results from the lack of opportunity and deprivation as experienced by lower class boys in English industrial cities.

(F.F./G.Ne.)

Criminal Law

Criminal law in a broad sense is the body of law that defines criminal offenses, regulates the apprehension, charging, and trial of suspected persons, and fixes penalties and modes of treatment applicable to convicted offenders. Criminal law is only one of the devices by which organized societies protect the security of individual interests and assure the survival of the group. There are, in addition, the standards of conduct instilled by family, school, and religion; the rules of the office and factory; the regulations of civil life enforced by ordinary police powers; and the sanctions available through tort actions brought by private persons. The distinction between criminal law and tort law is difficult to draw with any real precision, but in general one may say that a tort is a private injury while a crime is conceived as an offense against the state.

The scope of criminal law

The traditional approach to criminal law has been that crime is an act that is morally wrong. The purpose of criminal sanctions was to make the offender give retribution for harm done and expiate his moral guilt; punishment was to be meted out in proportion to the guilt of the accused. In modern times more rationalistic and pragmatic views have predominated. Writers of the Enlightenment such as Cesare Beccaria in Italy, Montesquieu and Voltaire in France, and P.J.A. von Feuerbach in Germany considered the main purpose of criminal law to be the prevention of crime. With the development of the social sciences, there arose new concepts such as those of the protection of the public and the reform of the offender. Such a purpose can be seen in the West German criminal code of 1969, which provides that the court "has to consider the consequences of the sentence upon the future life of the offender in society." In the United States, a model penal code proposed by the American Law Institute in 1962 states that the objective of criminal law should be "to give fair warning of the nature of the conduct declared to constitute a crime" and "to promote the correction and rehabilitation of offenders."

Comparisons of Anglo-American and continental European criminal law

Important differences exist between the criminal law of most English-speaking countries and that of other countries. The criminal law of England and the United States derives from the traditional English common law of crimes and has its origins in the judicial decisions embodied in reports of decided cases. Although legislation also played an important part in the development of the English law of crimes, England has consistently rejected all efforts toward comprehensive legislative codification of its criminal law. There was no statutory definition of murder in English law as late as the 1960s. Some Commonwealth countries, notably India, have enacted criminal codes that are based on the English common law of crimes.

The criminal law of the United States, derived from the English common law, has been adapted in some respects to American conditions. In about one-third of the states of the United States the common law of crimes has been repealed by legislation. The effect of such statutes is that no person may be tried for any offense not specified in the statutory law of the state. But even in these states the common-law principles retain great influence, for the criminal statutes often are simply codifications of the common law, and their terms and provisions are ordinarily interpreted by reference to the common law. In the remaining states, prosecutions for common-law offenses not specified in statutes are possible and do sometimes occur; such offenses include conspiracy, criminal solicitation, and breach of peace. In many states the so-called penal, or criminal, codes are simply collections of individual provisions enacted at different times in response to particular problems then current, with little effort to relate the parts to the whole or to define or implement any general theory of social control by penal measures. As a result, U.S. criminal legislation characteristically is poorly drafted and inconsistent.

In western Europe the criminal law of modern times has emerged from various codifications. By far the most important were the two Napoleonic codes, the *Code d'Instruction Criminelle* of 1808 and the *Code Pénal* of 1810. The latter constituted the leading model for European criminal legislation throughout the first half of the 19th century, after which, although its influence in Europe waned, it continued to play an important role in the legislation of certain Latin American and Middle Eastern countries. The German codes of 1871 (penal code) and 1879 (procedure) provided the models for a number of European countries and have had significant influence in Japan and South Korea, although after World War II the United States laws of criminal procedure were the predominant influence in the latter countries. The Italian codes of 1930 represent one of the most interesting legislative efforts in the modern period. English criminal law has strongly influenced the law of Israel and that of the English-speaking African states. French criminal law has predominated in the French-speaking African states. Ethiopia's criminal law is based on Swiss law. Italian criminal law and theory have been strongly influential in Latin America.

Movements for codification and reform

In the last few decades the movement for codification and law reform has made considerable progress everywhere. In the United States the National Commission on Reform of Federal Criminal Laws established by the Congress in 1966 published a study draft in 1970. The American Law Institute's model penal code (1962) stimulated a thorough re-examination of both federal and state criminal law. New codes were enacted in Louisiana (1942), Wisconsin (1955), Illinois (1961), Minnesota and New Mexico (1963), New York (1965), and Michigan (1969). England has enacted several important reform laws (including those on theft, sexual offenses, and homicide), and its Law Commission is at work on the codification of criminal law. Sweden enacted a new strongly progressive penal code in 1965. In West Germany a reform of the general part of the Criminal Code was completed in 1969. France enacted important reform laws in 1958 and 1970. Other reforms have been under way in Italy, Austria, Switzerland, Brazil, and Japan. The Soviet Union's constituent republics began enacting revised criminal codes in 1960, as did Czechoslovakia and Hungary (1961), the German Democratic Republic, Bulgaria, and Romania (1968), and Poland (1969).

Comparisons among the systems of penal law developed in the western European countries and those having their historical origins in the English common law must be stated cautiously. Substantial variations exist even among the nations that adhere generally to the Anglo-American system or to the law derived from the French, Italian, and German codes. In many respects, however, the similarities of the criminal law in all states are more important than the differences. Certain forms of behaviour are everywhere condemned by law. In matters of mitigation and justification the continental law tends to be more explicit and articulate than the Anglo-American law, although modern legislation in countries adhering to the latter is reducing these differences. Contrasts can be drawn between the procedures of the two systems, yet even here there is a common effort to provide fair proceedings for the accused and protection for basic social interests.

PRINCIPLES AND DOCTRINES OF CRIMINAL LAW

This section deals with substantive criminal law. The elements of substantive criminal law are: the definitions of the types of offenses that are held to be punishable; the classification of crimes (as, for example, felonies and misdemeanours in the United States, or *crime*, *délit*, and *contravention* in continental law); the principles and doctrines applied to the judgment of crime that qualify the provisions of criminal legislation (such as self-defense, necessity, insanity, and so forth); and principles determining national jurisdiction over crimes with an international aspect (crimes committed by foreigners, nationals abroad, or on ships and aircraft).

The definition of criminal conduct. *Legality.* The principle of legality is recognized in almost all civilized countries as the keystone of the criminal law. It is employed in four senses. The first is that there can be no crime without a rule of law; thus immoral or antisocial conduct not forbidden by law is not criminal. The law may be customary, as in common-law countries; in most countries, however, the only source of criminal law is a statute (*nullum crimen sine lege*, "no crime without a law").

Defining the application of the law

Second, the principle of legality directs that criminal statutes be interpreted strictly and that they not be applied by analogical extension. If a criminal statute is ambiguous in its meaning or application, it is often given a narrow interpretation favourable to the accused. This does not mean that the law must be interpreted literally, if to do so would defeat the clear purpose of the statute. The model penal code of the American Law Institute incorporates a provision that has been enacted in some U.S. state laws: abandoning the principle of "strict construction," it recommends that its provisions be construed "according to the fair import of their terms," which comes closer to the European practice.

Third, the principle of legality forbids the application of the law retroactively. In order that a person may be convicted, a law must have been in effect at the time the act was committed. This aspect of the principle is embodied in the ex post facto provisions of the U.S. Constitution and such international treaties as the European Convention for the Protection of Human Rights and Fundamental Freedoms (1950).

Fourth, the language of criminal statutes must be as clear and unambiguous as possible in order to provide fair warning to the potential lawbreaker. In some countries statutes may be considered inapplicable if they are vague. In *Lanzetta* v. *New Jersey*, 306 U.S. 451 (1939), the U.S. Supreme Court declared a statute that penalized membership in a "gang" to be unconstitutional.

Protection against double jeopardy. Legal systems generally include some restriction against prosecuting a person more than once for the same offense. In Anglo-American law the most difficult problems of double jeopardy involve the question whether the second prosecution is for the "same" or a "different" offense. It is held that acquittal or conviction of an offense prohibits subsequent prosecution of a lesser offense that was included in the first (*e.g.*, manslaughter is included in murder and cannot be prosecuted separately). It is sometimes possible, however, to prosecute a more serious crime that was included in a lesser one for which the offender has already been tried. In European law, on the other hand, the question is whether or not the second prosecution concerns the same "material fact" or "historical event," and the state cannot subject a person to a second trial for any offense arising out of the same factual situation.

A problem under the federal system of the United States is whether or not an offender may be prosecuted under both state and federal law for the same conduct (the specific offenses being different). By the beginning of the 1960s an increasing number of state laws prohibited state prosecutions after acquittals or convictions in federal courts for offenses involving the same conduct. The European Convention on the International Validity of Criminal Judgments (1970) extended the protection against double jeopardy even to judgments made in more than one country.

Statutes of limitation. All systems of law have statutes restricting the time within which legal proceedings may be brought. These are commonly called statutes of repose, enacted to protect against stale claims after evidence has been lost, memories have faded, or witnesses have disappeared. The periods prescribed may vary according to the seriousness of the offense. In German law, for example, the periods range from three months for petty misdemeanours to 30 years for crimes involving a life sentence. General statutes limiting the times within which prosecutions for crimes must be begun are common in Europe and the United States. In England there is no general statute of limitation applicable to criminal actions, although statutes for specific crimes frequently have included time limits.

In many countries there are no statutes of limitation for particularly heinous offenses, including certain felonies in the United States, genocide in West Germany, and—in the Soviet Union and other Communist-governed countries—crimes against the peace, war crimes, and crimes against humanity. In 1968 the United Nations General Assembly adopted a Convention on the Non-applicability of Statutes of Limitation on War Crimes and Crimes against Humanity, despite strong opposition among the majority of Western members on the ground that it was retroactive.

Requirements of jurisdiction. The jurisdiction of a court refers to its capacity to take valid legal action. Most governments claim jurisdiction over the acts of their own nationals, even when these acts have occurred abroad. Accordingly, most states decline any obligation to surrender their nationals to the jurisdiction of other countries; the constitutions of Brazil, West Germany, and The Netherlands prohibit extradition of their nationals; and in other states extradition is prohibited by law, as in Belgium, France, and Switzerland. The Italian constitution permits extradition of nationals only if specifically agreed upon in international conventions.

In Anglo-American practice, on the other hand, the jurisdiction of the courts is limited to acts occurring in whole or part within the geographical boundaries of the state. Nationals who commit crimes in foreign countries may be extradited, but only if this is required or authorized by treaty with the country concerned. Within the United States, jurisdiction over criminal conduct is primarily limited to acts occurring within the territorial limits of a particular state. Thus, if a person fires a bullet across a state line and kills someone in another state, sometimes only the latter state is considered to have jurisdiction. Both state and federal statutes have modified this principle in matters directly affecting state or federal interests. Federal statutes confer jurisdiction on U.S. courts in cases involving treason, forgery of ship's papers, enticing to desertion from military service, bribery of a U.S. official, and other acts, even though the conduct occurred outside the national boundaries. The United States also claims jurisdiction over crimes committed on U.S. vessels and aircraft on or over the high seas. The Tokyo Convention on Offenses and Certain Other Acts Committed on Board Aircraft (1963) recognizes that states have the right and even the duty of jurisdiction with respect to any crime committed upon a public or private aircraft that has its national character.

The elements of crime. It is generally agreed that the essential ingredients of any crime are (1) a voluntary act or omission (*actus reus*) accompanied by (2) a certain state of mind (*mens rea*). An act may be any kind of voluntary human behaviour. Movements made in an epileptic seizure are not acts. So also, movements made by a somnambulist before awakening, even if resulting in the death of another, are not acts and do not render the sleepwalker criminally liable. Criminal liability for the result also requires that the harm done must have been caused by the accused. The test of causal relationship between conduct and result is that the event would not have happened the same way without participation of the offender.

The nature of a criminal act

Criminal liability may also be predicated on a failure to act when the accused was under a legal duty to act and was reasonably capable of doing so. The legal duty to act may be imposed directly by statute, such as the requirement to file an income tax return, or it may arise out of the relationship between the parties, as the obligation of parents to provide their child with food.

The mental element. Although most legal systems recognize the importance of the guilty mind, or *mens rea*, the statutes have not always spelled out exactly what is meant by this concept. The American Law Institute's model penal code has attempted to clarify the concept by reducing the variety of mental states to four. Guilt is attributed to a person who acts "purposely," "knowingly," "recklessly," or "negligently." These terms correspond roughly to those used in European legal theory: purpose to *intention* in French and *Absicht* in German; knowledge to *dol éventuel* in French and *bedingter Vorsatz* in German (although in German law an additional element of consent is required); recklessness to *imprévoyance consciente* and *faute lourde* in French, the German equivalents being *bewusste Fahrlässigkeit* and *grobe Fahrlässigkeit;* negligence to *imprudence inconsciente* and *unbewusste Fahrlässigkeit*. These terms, singly or in combination, appear adequate to deal with most of the common *mens rea* problems. Their general adoption would greatly clarify and rationalize the substantive law of crimes.

Liability without mens rea. Some penal offenses do not require the demonstration of culpable mind on the part of the accused. These include statutory rape, in which knowledge that the girl is below the age of consent is not necessary to liability, and bigamy, which may be committed even though the parties believe in complete good faith that they are free to marry. There is also a large class of "public welfare offenses," involving such things as economic regulations or laws concerning public health and safety. The rationale for eliminating the *mens rea* requirement in such offenses is that to require the prosecution to establish the defendant's intent, or even recklessness, would render such regulatory legislation largely ineffective and unenforceable. Such cases are known in French law as *infractions purement matérielles.* In German law they are excluded because the requirement of *mens rea* is considered a constitutional principle.

There has been much criticism of statutes that create liability without moral fault. To expose citizens to the condemnation of a criminal conviction without a showing of moral culpability raises issues of justice. In many instances, moreover, the regulatory objectives of such legislation can more effectively be achieved by civil sanctions such as suits for damages, injunctions, and the revocation of licenses.

Ignorance and mistake. In most countries the law recognizes that a person who acts in ignorance of the *facts* of his action should not be held criminally responsible. Thus, one who takes and carries away the goods of another person, believing them to be his own, does not commit larceny for he lacks the intent to steal. Ignorance of the *law*, on the other hand, is generally held not to excuse the actor; it is no defense that he was unaware that his conduct was criminal. This doctrine is supported by the proposition that criminal acts may be recognized as harmful and immoral by any reasonable adult. The matter is not so clear, however, when the conduct is not obviously dangerous or immoral; a substantial body of opinion would permit mistakes of law to be asserted in defense of criminal charges in such cases, particularly when the defendant has in good faith made reasonable efforts to discover what the law is. In West Germany the Federal Court of Justice in 1952 adopted the proposition that, if a person engages in criminal conduct but is unaware of its criminality, he cannot be charged with a criminal offense. Law and practice in Switzerland are quite similar, although the Swiss courts are rather reluctant to recognize mistake of law as a good defense. In Austria the argument is accepted as a defense in cases which involve regulatory legislation when the conduct is not obviously dangerous or harmful.

Responsibility. It is universally recognized that, in appropriate cases, persons suffering from serious mental disorders should be relieved of the consequences of their criminal conduct. Much controversy has arisen, however, as to the appropriate legal tests of responsibility. Most legal definitions of mental disorder are not based on modern concepts of medical science, and the psychiatrist accordingly finds it difficult to make his knowledge relevant to the requirements of the court. Various attempts have been made to formulate a new legal test of responsibility. The Durham rule, set forth by the court of appeals for the U.S. District of Columbia in 1954, held that "an accused is not criminally responsible if his unlawful act was the product of mental disease or mental defect." The American Law Institute's model penal code endeavours to meet the manifold difficulties of this problem by requiring that the defendant be deprived of "substantial capacity either to appreciate the criminality of his conduct or to conform his conduct to the requirements of the law" as a result of mental disease or defect. This resembles the Soviet formulation of 1958, which requires a mental disease as the medical condition and incapacity to appreciate or control as the psychological condition resulting from it. The same may be said of the West German law, although the latter includes in mental illness such disorders as psychopathy and neurosis in addition to psychoses and provides for various gradations of diminished responsibility. The English Homicide Act of 1957 also recognizes diminished responsibility, though to less effect. It provides that a person who kills another shall not be guilty of murder "if he was suffering from such abnormality of mind . . . as substantially impaired his mental responsibility for his acts or omissions in doing or being a party to the killing." The primary effect of this provision is to reduce an offense of murder to one of manslaughter.

Intoxication is usually not treated as a mental incapacity. Soviet law is especially harsh, holding that the mental-disease defense is not applicable to persons who commit a crime while drunk and that drunkenness may even be

an aggravating circumstance. In West German law, on the other hand, intoxication like any other mental defect is acceptable as a defense in criminal cases.

Mitigating circumstances and other defenses. The law generally recognizes a number of particular situations in which the use of force, even deadly force, is excused or justified. The most important body of law in this area is that which relates to self-defense. In general, in Anglo-American law, one may kill an assailant when the killer reasonably believes that he is in imminent peril of losing his life or of suffering serious bodily injury and that killing the assailant is necessary to avoid imminent peril. Some jurisdictions require that the party claiming exculpation must avoid danger by retreat when this can be made without increasing his peril. Under many European laws, however, the defendant may stand his ground unless he has provoked his assailant purposely or by gross negligence, or unless the assailant is acting under some incapacity such as infancy, inebriation, mistake, or mental disease. Other situations in which the use of force is generally held to be justifiable, both in Anglo-American law and in European law, include the use of force in defense of others, law enforcement, and protection of property.

When the use of force is not considered justifiable, it may nevertheless be excused if the defendant believed himself to be acting under necessity. The doctrine of necessity in Anglo-American law relates to situations in which a person, confronted by the overwhelming pressure of natural forces, must make a choice between evils and chooses to engage in conduct that would otherwise be considered criminal. In the oft-cited case of *U.S.* v. *Holmes*, in 1842, a longboat containing passengers and members of the crew of a sunken American vessel was cast adrift in the stormy sea. To prevent the boat from being swamped, members of the crew threw some of the passengers overboard. In the trial of one of the crew members, the court recognized that such circumstances of necessity may constitute a defense to a charge of criminal homicide, provided that those sacrificed be fairly selected, as by lot. Because this had not been done, a conviction for manslaughter was returned. The leading English case, *Regina* v. *Dudley and Stephens*, 14 Q.B.D. 273 (1884), appears to reject the necessity defense in homicide cases. In German or French courts, however, the defendants would probably have been acquitted on the excuse either of *l'état de nécessité* or of *contrainte morale irrésistible*.

In general the use of force may also be excused if the defendant believed himself to be acting under duress or coercion, or to be carrying out military orders believed by the defendant to be lawful.

Some particular offenses. All advanced legal systems condemn as criminal the sorts of conduct described in the Anglo-American law as treason, murder, aggravated assault, theft, robbery, burglary, arson, and rape. With respect to minor police regulations, however, substantial differences in the definition of criminal behaviour occur even among jurisdictions of the Anglo-American system. Comparisons of the continental European criminal law with that based on the English common law of crimes also reveal significant differences in the definition of certain aspects of more serious crimes. Continental law, for example, frequently articulates grounds for mitigation involving considerations that are taken into account in the Anglo-American countries only in the exercise of discretion by the sentencing authority or in the behaviour of lay juries. This may be illustrated with respect to so-called mercy killings. The Anglo-American law of murder recognizes no formal grounds of defense or mitigation in the fact that the accused killed to relieve the agonies of one suffering from an apparently incurable disease. Many continental and Latin American codes, however, provide for mitigation of offenses prompted by such motives, and in a few instances even recognize in such motives a defense to the criminal charge.

Varying definitions of criminal behaviour

Degrees of participation. The common-law tradition distinguishes four degrees of participation in crime. One who commits the act "with his own hand" is a principal in the first degree. His counterpart in French law is the *auteur* (literally, "author"), or *co-auteur* when two or more persons are directly engaged. A principal in the second degree is one who intentionally aids or abets the principal in the first degree, being present when the crime occurs; this is comparable to the French concept of *complicité par aide et assistance*, although in some countries such as West Germany and the Soviet Union that have adopted a wider (more subjective) interpretation of the concept it includes the activity of *co-auteurs*. In Anglo-American law one who instigates, encourages, or counsels the principal without being present at the time the crime is committed is called an accessory before the fact; in continental law this degree of participation is covered partly by the concept of *instigation* and partly by the above-mentioned *aide et assistance*. The fourth and last degree of participation is that of accessory after the fact, who is punishable because he receives, conceals, or comforts one known by him to have committed a crime so as to obstruct his apprehension or to impede his punishment. In continental legal systems this has become a separate offense, known as *Begünstigung* in German and as *recel des malfaiteurs* in French. Italian and Austrian law treat all participants in a crime as principals in the first degree, with the exception of accessories after the fact. The American Law Institute's model penal code proposes the same simplification.

Conspiracy. In Anglo-American law conspiracy is usually described as an agreement between two or more persons to commit an unlawful act or to accomplish a lawful end by unlawful means. This definition is delusively simple, however, for each of its terms has been the object of extended judicial exposition. Criminal conspiracy is perhaps the most amorphous area in the Anglo-American law of crimes. In most jurisdictions, for example, the "unlawful" end of the conspiracy need not be one that would be criminal if accomplished by a single individual; but courts have not always agreed as to what constitutes an "unlawful" objective for these purposes. The European codes have no conception of conspiracy as broad as that found in the Anglo-American legal system. In some of the European countries, such as France or Germany, punishment of crimes may be enhanced when the offense was committed by more than one person acting in concert. In most countries the punishment of agreements to commit offenses irrespective of whether the criminal purpose was attempted or executed is largely confined to political offenses against the state. Some extension of the conspiracy idea to other areas has occurred, however. Thus in the Italian code of 1930 association for the purpose of committing more than one crime was made criminal. The Yugoslav code of 1951 makes criminal the organization of a group for the purpose of committing any offense punishable by more than five years' imprisonment. None of these provisions, however, has the generality of the Anglo-American concept. None, for example, condemns agreements to commit acts not otherwise criminal. The American Law Institute's model penal code restricts conspiracy to agreements to commit a crime or to aid other persons in the planning or commission of a crime.

Attempt. In Anglo-American law there is a class of offenses known as inchoate, or preliminary, crimes because guilt attaches even though the criminal purpose of the parties may not have been achieved. Thus the offense of incitement or solicitation consists of urging or requesting another to commit a crime. Certain specified types of solicitation may be criminal, such as solicitation of a bribe or solicitation for immoral purposes, or inciting members of the armed forces to mutiny.

Inchoate offenses

The most important category of inchoate offenses is attempt, which consists of any conduct intended to accomplish a criminal result that fails of consummation but goes beyond acts of preparation to a point dangerously close to completion of the intended harm. The line between acts of mere preparation and attempt is difficult to draw in many cases. In Soviet law acts of preparation are punishable if they are socially dangerous. Attempt is a further step and is generally defined in Soviet legal

theory as an act that contains all the elements of an offense except the result. In both Anglo-American and European legal systems, attempt may also consist of conduct that would be criminal if the attendant circumstances were as the actor believed them to be; in such cases the "defense of impossibility" is allowed only if the mistake can be shown to be absolutely unreasonable. Unlike the law of some European countries, including the Soviet Union, no defense is granted in Anglo-American law to an offender who voluntarily desists from committing the intended harm after his conduct has reached a point beyond mere preparation. The American Law Institute's model penal code, however, provides for an affirmative defense if it can be shown that the actor "abandoned his effort to commit the crime or otherwise prevented its commission, under circumstances manifesting a complete and voluntary renunciation of his criminal purpose."

CRIMINAL PROCEDURE

The law of criminal procedure regulates the modes of apprehending, charging, and trying suspected offenders; the imposition of penalties on convicted offenders; and the methods of challenging the legality of conviction after judgment is entered. The law in this area is called upon to advance and reconcile interests of the greatest importance. It must, on the one hand, contribute effectively to the attainment of public peace and good order; at the same time it must afford realistic protection to the rights of persons proceeded against by the system of criminal justice.

The investigatory phase. Criminal procedure begins with investigation. In this phase the material elements of the crime (the corpus delicti) are examined by the competent authority (the police or officers of the court or the prosecuting attorney), and all available evidence is collected for the use of the prosecution.

Pretrial procedures

The role of the police. In Anglo-American procedural law the police play a primary role in the pretrial stage, being responsible for the arrest of suspects, the execution of warrants, the questioning of witnesses, and the carrying out of searches and seizures. In Europe, however, the police generally act as the agents of the prosecuting authorities (*magistrature debout*) or examining judges (*magistrat instructeur*).

The role of the prosecutor in the pretrial stage. In Anglo-American law the prosecutor plays only a limited role in the preliminary investigation. In the United States the prosecuting attorney has to decide whether or not he has a case that will stand up in court. In England the director of public prosecutions undertakes the more difficult and important cases, but most charges are preferred and most prosecutions are undertaken by the police. In West Germany the police have acquired a considerable degree of independence in pretrial investigations, although they are nominally under the guidance of the prosecutor. In the Communist-governed countries of Europe, the procurator is both the prosecuting authority in criminal cases and at the same time the guardian of "socialist legality"—for which purpose he has wide control powers over the administrative and economic agencies of government.

Warrants. A warrant empowers a police officer to arrest a suspected criminal or to search premises and seize property for the purpose of obtaining evidence. Although arrests may be made without a warrant, particularly when a person is seen committing a criminal act, the laws of most countries limit the power of arrest in order to prevent the police from interfering with individual liberty except in cases of apparent necessity.

The judge's role. In Anglo-American law there may be a pretrial hearing before a judge in order to show the accused the nature of the charge and to determine whether or not the evidence presented by the prosecutor is sufficient to justify further action. This takes the form of an adversary proceeding by the attorneys for the prosecution and the defense. In continental law, however, the pretrial proceeding is inquisitorial—that is, it is an inquiry conducted by the prosecuting authority. In France it is conducted by a special judge, the *juge d'instruction;* in West Germany it is carried out almost entirely by the police under the authority of the prosecuting attorney.

Defense counsel. The defense counsel is of the highest importance during the pretrial proceedings. It is his task to protect the defendant against unfounded testimony, to introduce evidence in his favour, to employ various procedural devices for his benefit, and to provide moral support. All legal systems give defendants the right to counsel, and normally the law requires that the prosecuting authority inform a defendant of this right. In the United States several supreme court decisions have broadened the rights of the defendant with respect to pretrial investigation; in *Miranda* v. *Arizona,* (1966) the court attached a number of restrictions to the admissibility of confessions obtained by police interrogation. There have been similar tendencies in European countries, especially in amendments made to the West German code of criminal procedure.

Preparation of charges. In all legal systems, criminal proceedings begin with a formal accusation made to the court. In the United States it is the responsibility of the prosecutor to prepare formal charges; he may decide which charges to press and may even refrain from pressing charges at all, or he may suspend a trial that is already under way. In England most charges are preferred and most prosecutions are undertaken by the police; the director of public prosecutions handles only the more difficult and important cases. In France, on the other hand, the *procureur de la République* is entrusted with the formal act of accusation; because of his discretionary power to decide whether or not charges are to be pressed, even in serious cases, he is called the judge of the advisability of the prosecution. In West Germany, again, the situation is quite different; the German prosecutor, like his colleague in Italy, is bound by the "legality principle," under which he is obliged to start an action whenever he can prove his case in court, except for petty matters.

Role of the police

In some countries the police have a discretionary influence in the pressing of charges. In the United States they frequently reduce the charges in order to make the prosecution easier. In England, the police can decide whether or not to prosecute depending on the importance of the case and their assessment of the evidence. In European countries, however, the discretionary role of the police is restricted to only minor cases.

The Anglo-American system provides for pretrial hearings, subject to waiver by the defendant, in the presence of the court. In these hearings evidence may be excluded if it does not meet prescribed rules. In England the pretrial examination may take place in writing with the exchange of documents and depositions without formal hearing. In those European countries in which the examining judge may also conduct the preliminary inquiry, particularly in France and Spain, he may handle the entire business of collecting the evidence and committing the accused person to the trial court if the charges seem warranted.

Private charges. Private citizens are always entitled to file charges. In England such a complaint, when signed by an appropriate officer of the court, has the force of an indictment. In France the victim of an offense may initiate action by filing a civil party complaint before the *juge d'instruction* or the trial court; in this way he exercises considerable control over whether or not charges will be pressed. In West German law some offenses are not prosecuted unless the victim makes formal complaint. In Communist-governed countries a deliberate effort is made to involve the public in combatting crime; heads of governmental departments, factory directors, and other administrative officials are often required to discuss with the procurator major crimes committed in the areas of their responsibility.

The criminal trial

Trial procedure. The place of trial may be determined according to where the crime was committed, the defendant's place of residence, the place at which he was apprehended, or some other consideration. When an offense is committed on a moving vehicle, such as a train or a

boat, it may be impossible to determine precisely where it occurred; in such cases some laws require that the trial be held in a county or district through which the vehicle passed. In Anglo-American law the place of trial is called venue because formerly jurors had to "come" (in Latin *venire*) from the locality where the crime occurred, since they based their verdict on personal knowledge of the facts. But the traditional common-law rule that the defendant had to be tried at the place where the crime was committed has been amended by statute both in England and the United States for reasons of convenience or to avoid the influence of local resentments; indeed, the accused himself may request a change of venue because of prejudice against him in the community or in the court.

The jury. In a jury trial the functions of adjudication are divided between the presiding judge, who controls the admission of evidence and instructs the jury as to the applicability of the law, and the jury consisting usually of 12 laymen who decide the question of guilt. If the defendant is declared guilty, the judge usually determines his sentence.

The jury and the common law are sometimes said to be the two distinctive Anglo-American legal institutions. The Sixth Amendment to the U.S. Constitution guarantees the right of jury trial in criminal cases, as do the state constitutions. In the U.S., the defendant may waive his right to trial by jury, and consequently many trials are conducted entirely by the judge. In both the U.S. and England, no jury is called if the defendant enters a plea of guilty.

In Europe, however, the jury has been almost entirely abandoned. France abolished it in 1941, Italy in 1931, Spain in 1936; in Germany the jury was replaced in 1924 by a mixed court in which three professional judges deliberate and decide the questions of guilt and sentencing in common with six lay judges chosen at random. The jury survives today only in Austria, Belgium, Norway, and four Swiss cantons.

Something analogous to the jury has been introduced in the Communist-governed countries of Europe, where "comrades' courts," factory courts, and collective probation boards composed of laymen are used in order to secure public participation in the trial of petty offenses.

In the U.S. a jury is usually required to return a unanimous verdict. In England, under certain conditions, 10 jurors out of a panel of 11 or more can determine the verdict. Both the prosecution and the defense have the right to challenge a limited number of the jurors without giving a reason; and they have an unlimited right of challenge for cause (on the ground that a prospective juror is biassed because of his religion, race, language, or some other factor). Objections may also be raised against the trial judge if there is reason to suspect him of bias or prejudice. In European law, judges may also be challenged (see also JURY).

Public trial. The guarantee of a public trial is generally recognized as an important human right, often provided for in constitutions. Only reasons such as state security or public morality are invoked to justify trials *in camera* or in secret. Trial must be in open court, which means that the public is admitted but that filming and broadcasting are generally not permitted.

Right of counsel. Legal systems generally guarantee the right of counsel in criminal prosecutions. This right is difficult to assure when the defendant is poor, but some countries now provide free legal aid for the indigent.

Presentation of evidence. In Anglo-American law, the presentation of evidence is left to the parties themselves. Witnesses are examined and cross-examined by counsel, not by the court. The functions of the trial judge are to enforce the rules governing evidence and to ask supplementary questions if he feels that the parties have failed to clarify the facts. The defendant may testify as a witness, if he chooses to, but he is not examined by the judge as he is in European procedure. If the defendant pleads guilty, and his plea is accepted by the court, no jury verdict is called for. A confession made outside the court but not accompanied by a guilty plea must be corroborated by other evidence.

European or civil-law procedure is quite different. One of the main tasks of the presiding judge is to elicit evidence by putting questions to witnesses and experts. (Exceptions to this are Spain, some Scandinavian practice, and Japan.) The defendant does not have the right to take the witness stand and testify as in Anglo-American law, but he is examined by the presiding judge. He may remain silent if he chooses.

Establishing guilt. A basic principle of both Anglo-American and European procedure is that in criminal cases guilt must be established beyond a reasonable doubt. The burden of proof rests upon the prosecution. On the continent this is true even in cases involving insanity, drunkenness, self-defense, or necessity, and in cases in which the question of intent arises. Anglo-American law is generally satisfied with the formula that the actor "is presumed to intend the natural and probable consequences of his act," but in continental law the prosecution must prove intent. Otherwise the axiom *in dubio pro reo*, corresponding to the Anglo-American "presumption of innocence," requires a ruling in favour of the defendant. In reaching a verdict as to the guilt of the defendant, U.S. law generally requires that the jury be unanimous; in Europe, however, a two-thirds majority of the judges' bench is generally sufficient.

Sentencing. A guilty verdict does not necessarily mean that the defendant will be sentenced to prison. In Anglo-American jurisdictions the court may order an investigation of the offender to see whether or not his background and character justify placing him on probation. The sentence of a person on probation is conditionally suspended upon promise of good behaviour and agreement to accept supervision and abide by specified requirements including, usually, reporting to a probation officer or to the court at regular intervals. European law emphasizes conditional sentencing rather than probation; under its procedure the sentence is imposed but with provision for suspending execution of the punishment subject to the offender's good behaviour.

Postconviction procedures. After conviction, a defendant may move in the trial court to arrest judgment or he may file a motion for a new trial. Among the grounds most frequently asserted in such motions are that the verdict is not supported by the law or the evidence; that newly discovered evidence has come to light; and that the court erred in its rulings on the admission of evidence.

Motions for a new trial

The legality of the conviction may also be challenged by appeal to a higher court. Criminal appeals were unknown in the traditional common law, and even today the U.S. Supreme Court does not consider them to be required by that country's constitution. Not until 1879 were appellate proceedings in criminal cases introduced into U.S. federal courts, although in many states criminal review had been practiced much earlier. Under U.S. federal procedure, appeals may be made on two levels: the first is that of the court of appeals; the second is that of the Supreme Court, which may grant a review on a writ of certiorari if it feels some important reason requires it; for example, if there is a conflict between the decision rendered by the lower court and the provisions of the Constitution. In the U.S., as a rule, the prosecution cannot enter a motion of appeal after the defendant has achieved an acquittal because of the broad interpretation given to the protection against "double jeopardy." The defendant may appeal a conviction on the ground that the rules of evidence have been violated by the court or that new evidence has come to the attention of the defendant. In addition, the writ of habeas corpus may be employed to attack a conviction by challenging the jurisdiction of the trial court.

In England, the Criminal Appeal act of 1907 established an elaborate system of appellate procedure, proceeding from magistrates' courts all the way to the House of Lords, the supreme court of England. Extraordinary remedies in English procedure include the writ of habeas corpus, on the ground of unlawful detention, and the orders of mandamus, certiorari, and prohibition, which are directed against improper handling of fundamental questions of jurisdiction.

In continental countries every criminal judgment is in principle subject to review except, in some countries, jury acquittals. Decisions of the supreme court sitting in first instance (mostly in political cases) are equally final and cannot be appealed. The principle of "equal arms" is widely applied in such a way that the appeal is open to the defendant and to the prosecuting authority as well. In some European countries there are two different kinds of appeal, the first pertaining to questions combining facts and law and the second relating to questions of law alone.

In the Communist-governed countries of Europe there is a supervisory appeal beginning after sentencing and designed to enforce the principle of socialist legality. The supervisory appeal is open only to the prosecutor and court chairmen, not to the defendant. It can be lodged either in favour of the defendant or against him and may deal with both the findings of fact and the interpretation of the law on the part of the lower court.

THE RELATIONSHIP OF CRIMINAL LAW TO THE SOCIAL SCIENCES

Criminal law has been strongly influenced in the last century or two by the social sciences, especially criminology, sociology, and psychology. The empirical methods of the social sciences have been introduced into legal research and have done much to improve the courts' approach to sentencing, as well as the planning methods of law-enforcement agencies.

Criminal law and human behaviour

Behavioral norms. The fact that in recent decades the crime rates in many countries have risen faster than the population has brought into question the relevance of the law itself, and whether or not laws against crime actually influence behaviour. Various large-scale inquiries have been made into the relation between law and civil order: in the United States, the President's Commission on Law Enforcement and Administration of Justice; in Europe, several research studies sponsored by the Council of Europe; in Germany, the hearings of the Criminal Reform Commission of the Bundestag. One conclusion emerging from these inquiries is that criminal legislation ought to be restricted to acts that pose a serious threat to public order and that can be effectively dealt with by the police, the courts, and various correctional institutions. The effort to punish all behaviour that is considered immoral or deviant, such as drunkenness, gambling, disorderly conduct, vagrancy, and petty sex offenses, multiplies the number of crimes without changing the norms of behaviour.

The effectiveness of statutes. Man's conduct is determined by many factors that are not responsive to criminal statutes. Thus it appears that introducing or abolishing the death penalty does not have any appreciable effect on the murder rate. On the other hand, the severe British legislation on road safety in 1967, with its introduction of the "breath test" for drunkenness, apparently brought about a significant decrease in traffic accidents. This may have been less a consequence of the severity of the law than of its correspondence to generally accepted moral standards (that drivers should be careful) and its educational impact on the public. Much depends also on the way in which laws are enforced. Any inquiry into the effectiveness of criminal statutes must examine the way in which police, attorneys, and the courts actually operate—for example, the manner in which they investigate suspects, gather evidence, instruct juries, and use their discretionary powers in "plea bargaining" and in sentencing.

The effectiveness of punishment. It is difficult to measure scientifically the extent to which punishment serves to deter convicted offenders from committing further crimes. The requirements of justice forbid any experimentation along that line. The prevailing opinion among criminologists is that short-term sentences are particularly harmful because they tear the offender away from his family and occupation and expose him to criminal indoctrination in prison and to social obloquy after his release. The new general part of the West German criminal code replaces short-term imprisonment (less than six months) in most cases by fines. Long-term sentences are also viewed with growing skepticism, despite more than 150 years of prison reform, because of the adverse side effects of even the best institutions. These ill effects include acclimatization to the prison atmosphere, association with prison subcultures, infantilism, mental illness, and in general a decline in fitness for responsible life in a free community. It is now considered preferable to treat the convicted criminal in open institutions if the seriousness of the crime and the personality of the offender do not make this impossible.

Treatment of mental deficiency and juvenile delinquency. A large area of criminal behaviour involving mental deficiency and diminished responsibility cannot be dealt with through ordinary prison sentences. Mentally disabled offenders require hospitalization and psychiatric treatment; this is usually handled through the probation mechanism or by commitment to a hospital for the criminally insane. Similar problems arise in the case of crimes resulting from narcotics addiction; prison terms for addicts and "pushers" do not make sense unless some effort is made to treat the underlying condition.

Juvenile delinquency has been increasing throughout the world, particularly in urban and industrialized areas. Young offenders are generally dealt with through separate juvenile courts and sent to detention centres, training centres, and part-time homes. Others are handled through conditional sentencing and probation (see also CRIME AND DELINQUENCY; PRISONS AND PENOLOGY).

BIBLIOGRAPHY

General and comparative works: J.A. COUTTS (ed.), *The Accused* (1966); H. JONES, *Crime and the Penal System*, 3rd ed. (1965); LUIS JIMENEZ DE ASUA, *Tratado de derecho penal*, 2nd ed. 5 vol. (1957–63), vol. 1–2, 3rd ed. (1964); H. MANNHEIM, *Comparative Criminology* (1965); G.O.W. MUELLER and F. LE POOLE-GRIFFITHS, *Comparative Criminal Procedure* (1969); S.S. NAGEL, *The Legal Process from a Behavioral Perspective* (1969).

Treatises on particular countries: M. ANCEL *et al.*, *Le droit pénal des pays scandinaves* (1969); J. ANDENAES, *The General Part of the Criminal Law of Norway*, trans. by T.P. OGLE (1965); R. ARGUILE, *Criminal Procedure* (1969); G. BETTIOL, *Diritto penale*, 7th ed. (1969); P. BOUZAT and J. PINATEL, *Traité de droit pénal et de criminologie*, 2nd ed., 3 vol. (1970); R. CONQUEST, *Justice and the Legal System in the USSR* (1968); S. DANDO, *Japanese Criminal Procedure*, trans. by B.J. GEORGE (1966); F.J. FELDBRUGGE, *Soviet Criminal Law* (1964); J.N. HAZARD and I. SHAPIRO, *The Soviet Legal System* (1962); H.H. JESCHECK, *Lehrbuch des Strafrechts: Allgemeiner Teil* (1969); D. KARLEN, *Anglo-American Criminal Justice* (1967); G. LEONE, *Trattato di diritto processuale penale*, 3 vol. (1961); R. MAURACH, *Deutsches Strafrecht: Allgemeiner Teil*, 4th ed. (1971); R. MERLE and A. VITU, *Traité de droit criminel* (1967); L.B. ORFIELD, *Criminal Procedure from Arrest to Appeal* (1947), *Criminal Procedure and the Federal Rules*, 7 vol. (1966–68); K. PETERS, *Strafprozess: Ein Lehrbuch*, 2nd ed. (1966); ROSCOE POUND, *Criminal Justice in America* (1945); E.W. PUTTKAMMER, *Administration of Criminal Law* (1953); L. RADZINOWICZ, *A History of English Criminal Law and its Administration from 1750*, 4 vol. (1948–68); J.M. RODRIGUEZ DEVESA, *Derecho penal español: Parte general* (1970); J.F. STEPHEN, *A History of the Criminal Law of England*, 3 vol. (1883); R.J. and M.G. WALKER, *The English Legal System*, 2nd ed. (1970); G.L. WILLIAMS, *The Proof of Guilt: A Study of the English Criminal Trial*, 3rd ed. (1963).

Works on substantive criminal law: M. CHERIF BASSIOUNI, *Criminal Law and its Processes: The Law of Public Order* (1969); J. HALL, *General Principles of Criminal Law*, 2nd ed. (1960); S.H. KADISH and M.G. PAULSEN, *Criminal Law and its Processes: Cases and Materials*, 2nd ed. (1969); C.S. KENNY, *Outlines of Criminal Law*, 19th ed. by J.W. CECIL TURNER (1966); R.M. PERKINS, *Criminal Law*, 2nd ed. (1969); J.C. SMITH and B. HOGAN, *Criminal Law*, 2nd ed. (1969); G.L. WILLIAMS, *Criminal Law: The General Part*, 2nd ed. (1961).

Works in criminology and sociology: M. ANCEL, *Social Defence*, trans. and rev. by J. WILSON and the author (1965); COUNCIL OF EUROPE (ed.), *The Effectiveness of Punishment and other Measures of Treatment* (1967); M. GRUNHUT, *Penal Reform: A Comparative Study* (1948); N. JOHNSTON *et al.* (eds.), *The Sociology of Punishment and Correction*, 2nd ed. (1970); G.K. STURUP, *Treating the "Untreatable": Chronic Criminals at Herstedvester* (1968).

(H.-H.J.)

Criminology

Criminology is the scientific study of the nonlegal aspects of crime (including juvenile delinquency). In its wider sense, embracing penology, it is thus the study of the causation, correction, and prevention of crime—seen from the viewpoints of such diverse disciplines as ethics, anthropology, biology, ethology (the study of character), psychology and psychiatry, sociology, and statistics. Whereas the traditional legal approach to crime focusses on the action of crime and the protection of society, criminology focusses on the person of the criminal and the essential interests of the individuals of whom society consists. Whereas criminal law has been a relatively conservative force, often slow to change even where change has seemed imperative, criminology as a part of the developing social sciences of the past hundred years has been a revolutionary force—its object being not to replace the legal system in dealing with crime and punishment but to supplement it, making it less rigid and more sympathetic to approaches wider than strictly legal ones.

Criminology and criminal law compared

THE NATURE OF CRIMINOLOGY

Functions and scope. Although variously opposed to the rigidity of some of the criminal law, the social sciences are nevertheless in many respects incapable of forming a united opposition. Whereas psychiatry, psychoanalysis, and (to a lesser extent) psychology, for instance, emphasize the individual, sociology and anthropology focus more on society as a whole, its institutions and groups of individuals. Not only do many of the problems of the respective disciplines differ but also their techniques in trying to solve them vary greatly. In their scientific languages they are often far apart, and students in a few of them are even unwilling to accept certain other disciplines as scientific in any sense at all. True, one of the ambitions of criminology has been to provide a neutral territory for discussion and perhaps resolution of some of these difficulties of communication, but the fact is that criminologists themselves—coming as they do from diverse scientific backgrounds and thus torn between conflicting loyalties—are in need of "integration." The situation, however, has been improving with the growth of criminological studies in the 20th century and with the multiplication of institutes, congresses, conferences, journals, and literature; and, in some countries at least, criminologists seem to have become more than mere interpreters and coordinators and arrived as collectors and developers of their own vast body of scientific knowledge.

Role of applied research. Without denying the value of "pure research," one must point to criminology and particularly penology as primarily practical subjects or "applied" disciplines. This practical value of criminological research can make itself felt in several ways. Its accumulated findings can give judges, prosecutors, lawyers, probation officers, and prison officials better understanding of crime and criminals, leading hopefully to more effective and humane sentencing and methods of treatment. Criminological research and knowledge can be equally at the disposal of legislators and administrators to assist in their task of reforming the law and improving penal and reformatory institutions. Essentially this purveyance of information represents a neutral role for criminologists; they garner the facts, and the various governmental officials decide for themselves what kind of practical conclusions to draw from the facts. Increasingly, however, some criminologists—like their counterparts in such fields as the atomic sciences—are demanding that scientists fully shoulder the moral and political responsibilities for their discoveries and for the use made of them instead of leaving vital decisions entirely to their governments. Thus some criminologists, for instance, insist upon actively campaigning against capital punishment, given the facts as they see them. Opponents of this activist role, on the other hand, contend that penological arguments are not sufficient but must be weighed along with political, social, religious, and moral arguments and that this all-round consideration should be left to responsible political bodies. The view does not deny the right of criminologists to express their opinions as ordinary citizens and voters; it does contend, nevertheless, that a government of officials responsive to the popular will, however fallible it may be, is less dangerous than a "government by experts."

The role of criminalistics and victimology

Another question involving the scope and functions of criminology is whether or not it should extend to the study of crime detection, involving such measures as photography, toxicology, fingerprint study, and the like. In several countries, notably Austria and Belgium, and at the school of criminology of the University of California at Berkeley, this so-called criminalistics has long been an important branch of criminological teaching and research, and the distinguished *Journal of Criminal Law, Criminology, and Police Science* (U.S.) devotes much of its space to criminal investigation. Actually, the only reason for excluding it from criminology is perhaps the expense of staff and equipment, which can be better borne by police colleges and similar specialized institutions. On the other hand, in recent decades criminology has undergone an important and perfectly legitimate extension of its territory by devoting much attention to so-called victimology—the study of the victim of crime, his relations to the criminal, and his role as a potential causal factor in crime.

Status among other disciplines. Although the exclusion of criminalistics makes it easier to locate criminology on the map of scientific studies, its origin in, its close relations to, and its partial dependence on so many other disciplines result in considerable diversity and confusion regarding its proper place in the academic curriculum. Universities in continental Europe, when they do not ignore criminology altogether, tend to treat it as part of legal education; even where its principal teachers are not lawyers. In Great Britain the only existing Institute of Criminology is part of the law faculty of Cambridge University; in other schools criminological research and teaching are usually divided between departments of sociology or social administration, law faculties, and institutes of psychiatry. In South America the anthropological and medical elements predominate, and in the United States, criminology, with a few notable exceptions, forms an established section of departments of sociology.

Given this situation in which criminology is submerged in other fields, it is not surprising that most teachers and researchers in criminology regard themselves first as sociologists, psychologists, lawyers, or whatever and only secondarily as criminologists. Their education contributes to this status; although a number may have pursued some criminological studies in their undergraduate years, criminology is largely a postgraduate discipline, at least in terms of major concentration for students.

This floating character of criminology weakens its position and tends to lend doubt to its claim to scientific status. Nevertheless, other disciplines—such as psychology, psychiatry, history, sociology, and social anthropology—have gone through similar birth pangs and, even after having achieved more or less assured positions, still face challenges to their claim to being scientific disciplines. The answer lies perhaps in historian H.R. Trevor-Roper's remark, "there are sciences and sciences." If the results of research can be viewed relatively, it is possible to perceive science in the criminologist's systematic application of sound research methods and his development of a body of facts from which he interprets general trends on a subject of real importance to mankind.

HISTORY

The origins of criminology are generally dated from the late 18th century, when those imbued with a spirit of humanitarianism began questioning the cruelty, arbitrariness, and inefficiency of criminal justice and prison systems. From this period arose the so-called classical school of criminology, composed of such reformers as the Italian Marchese di Beccaria and the Englishmen Sir

Samuel Romilly, John Howard, and Jeremy Bentham, all of whom may be said to have sought penological and legal reform rather than criminological knowledge per se —that is, knowledge about crime and criminals. Their principal aims were to mitigate legal penalties and subject judges to the principle of *nulla poena sine lege* or "due process of law" and also to reduce the application of capital punishment and humanize penal institutions. In all this they were moderately successful, but in their desire to make criminal justice "just," they tried to construct rather abstract and artificial equations between crimes and penalties, thereby forgetting the personal characteristics and needs of the individual criminal. Moreover, the object of punishment was seen as being primarily retribution, with deterrence occupying second place, and reformation lagging far behind.

Positivist school

By the second half of the 19th century these deficiencies, together with the influential teachings of the French sociologist Auguste Comte, had prepared the ground for the positivist school, which sought to bring a scientific neutrality into criminological studies. Instead of assuming a moral stance that focussed on measuring the criminal's "guilt" and "responsibility," the positivists attempted a morally neutral and social interpretation of crime and its treatment. Their leading figure, Cesare Lombroso (1836–1909), professor of psychiatry and anthropology at the University of Turin, sought through firsthand observation and measurement of prison inmates to determine the characteristics of criminal types. Some of his investigations led him into anthropometric interpretations—for example, his oft-criticized deduction of the "born criminal" with cranial, skeletal, and neurological malformations—but largely he and other positivists helped to introduce the ideas that crime has multiple causes and that most criminals are not born criminal but are shaped by their environmental upbringing and associations. With the positivists, therefore, the emphases in criminology had turned to experimental case studies and to preventive and rehabilitative measures. Without the upheaval caused by the positivists not only criminological research in the modern sense but also the present-day alternatives to capital punishment and old-fashioned imprisonment such as probation, suspended sentence, fines, and parole, inadequate as some of them are, would have been unthinkable.

Today, nevertheless, the feeling is widespread that the battle of ideas fought by the classical and positivist schools has not yet produced the secure foundations on which the criminology and the penal systems of the future can be built. Thus a third school, the postwar movement of "social defense," also originating in Italy, has tried to combine their best features and eliminate their excesses. This school disapproves of any rigid typology of criminals and stresses the uniqueness of human personality; it refuses the "scientism" of the positivists in favour of a strong belief in moral values—most importantly in balancing the rights of criminals and the rights of society. The school still speaks with too many voices, however, to be conclusively labelled.

Independent of these debates between schools, however, are the advances accomplished by such great figures as the statisticians Adolphe Quetelet (1796–1874) and André Michel Guerry (1802–66), the sociologists Gabriel Tarde (1843–1904) and Émile Durkheim (1858–1917), and of course Sigmund Freud, all of whom introduced a wealth of new ideas into the old problems of the social and individual characteristics of human, including criminal, behaviour.

MODERN CRIMINOLOGICAL RESEARCH

Objectives. The objectives of criminological research are sometimes said to be threefold: descriptive, causal, and normative. The descriptive aspect consists of the collection of relevant and reliable facts, together with their interpretation. The collection does not begin as a random and meaningless running after whatever phenomena happen to rouse the researcher's interest. Rather it is preceded by some hypothesis or "hunch," an assumption about what the researcher expects to find. The hypothesis "organizes" his inquiry. As the collection of facts proceeds, he may find either that his hypothesis is correct or that it requires revision or abandonment and thus the development of a new hypothesis to guide further research. The history of criminology reflects this perennial revision and renewal of inquiry, this continuous process of abandoning seemingly well-established theories in favour of new ones. Lombroso's theory of the "born criminal," the theory that all crime (or at least all economic crime) is due to poverty, and the theory that all juvenile delinquents come from broken homes had all to be drastically revised.

Forming and testing a hypothesis

The causal aspect involves relating the effects of one body of facts on another. Although regarded nowadays with some suspicion or indifference (even in the natural sciences), the search for causes is not being dispensed with altogether. So long as one does not jump to causal conclusions when arriving at statistical correlations or does not pressure "facts" into some proof of a popular theory, theories of causation can be useful in planning for the alleviation of crime and criminality.

The normative aspect, however, is more decidedly suspect. Research aimed at formulating so-called laws governing criminological phenomena has been thus far futile and does not look promising. What is sometimes regarded as a "law" has in reality been a mere trend. To diagnose statistical trends may be useful so long as their possible ephemeral nature is recognized, but trends are not laws, and there is thus but little scope for the normative approach in criminology.

Cross-disciplinary and cross-cultural approaches

The character of the research. Criminology, as suggested earlier, is cross-disciplinary and indeed draws methods or techniques from both the natural and the social sciences and must continually take heed of developments in other fields. It also depends increasingly on cross-cultural approaches; there have been recent statistical (though admittedly controversial) comparisons of "delinquent generations" in England, New Zealand, Denmark, and Poland; various studies of the sociological and statistical aspects of homicide; and studies of the ecological significance of "criminal areas" and of the possibility of predicting criminality.

In common with other disciplines, criminology must face such distinctions as between pure and applied research and between statistical and intuitive ways of thinking, but what is almost unique to criminological research is its intense involvement with society and its difficulty in achieving "detachment." Not only do society's biases toward crime and punishment influence a criminologist's choice and execution of research but also he is dependent on the willing cooperation of governmental departments and other public authorities to secure essential raw material or data. There are only a few limited areas—such as adolescent delinquency and gang activities—in which research can be pursued privately without resort to official help.

Techniques of research. Seen in the light of the previous remarks on the history of criminology, the development of criminological research can be divided into three stages: the prescientific, lacking any theoretical basis and largely identifiable with the work of the classical school; the semiscientific, possessing theories and hypotheses but without scientifically sound techniques and largely characteristic of the efforts of the positivist school; and the scientific, hopefully trying to repair earlier deficiencies and developing or improving the techniques described below.

Statistics. Often serving as the initial step in any research and regarded by some researchers, perhaps incorrectly, as the one and only reliable technique, the collection and interpretation of statistics for social and criminological purposes began in Europe early in the 19th century. The reputed "father" of this criminological method is the Belgian astronomer Adolphe Quetelet, who is perhaps best remembered for his famous "law," developed from the French and Belgian statistics, which showed that crime in any given country remains fairly

constant over the long term (short-term fluctuations being insignificant). When he qualified himself, however, by saying that the volume and kind of crime were constant only so long as society's social, economic, and political conditions remained unchanged, he deprived the "law" of much of its significance for a rapidly changing modern world.

The manner and extent of data collection today differ considerably from country to country or, in federal unions like the United States, even from state to state or province to province. They differ in how often data are collected and published, in what items are given importance, in the choice between complete listings and sample surveys, in the ratio between governmental and private research, and so forth. Such far-reaching differences, together with the differences in law and its administration and in popular views and habits, have made it so far impossible to devise a meaningful system of international criminal statistics. Generally, however, there is increasingly less tendency to collect any and all data, regardless of reliability or practical value, and to concentrate on limited, reliable data involving matters of agreed upon importance.

Comparison of police and court statistics

A noteworthy distinction to be made is between police statistics and court statistics. Police data are nearer to the event but perhaps less reliable, and they usually describe the crimes only, not the criminals. The data from the courts, being based on convictions, do deal with the persons involved but include only the material brought forward by the prosecution and the defense. All criminal statistics indeed depend entirely on human factors such as the willingness of private individuals and officials to prosecute, on the popularity or unpopularity of the criminal laws at issue, and so forth. The figures also usually fail to rate very clearly the gravity of individual cases; except for such broad categories as "petty" and "grand," theft is theft regardless of the value of the objects stolen. Only recently have more detailed categories been attempted for criminological research.

Case Study. Also called the individual "case history," the case study concentrates on the career or life of one individual or group of individuals and is the method used primarily, though not exclusively, by psychologists, psychiatrists, and psychoanalysts. If well done, such histories can give deep insights into the personalities and motives of criminals, but the method does have shortcomings. Although the volume of case histories has grown large, their reliability is sometimes suspect—partly because of a criminal's natural reluctance to expose himself completely and partly because of the nature of the publication of case histories. Their publication is comparatively rare; professional ethics often forbid the exposure of details given confidentially, and those studies actually published may be too few to be typical and may even on occasion be designedly selective because of an investigator's wish to prove a theory.

Closely related to case studies are autobiographies and other books written by ex-prisoners, but in spite of their considerable human and scientific interest, they do suffer from even greater disadvantages, chiefly questionable objectivity. Sociologists have also contributed important studies of individuals in their social environments.

Typologies. The typological method involves classifying offenses, criminals, criminal associations, criminal areas, or whatever according to some criteria of relatedness or similarity. Thus there have been attempts to dichotomize criminals as either "normal" or "abnormal," "habitual" or "professional," or to form a continuum of criminals from the "insane" at one extreme through various career criminals, petty offenders, and white-collar criminals to "organized" or "professional" criminals at the other extreme. The typological method is less impersonal and heterogeneous than the statistical method and less individual or specific than the case study. Developed mainly in Germany and Austria and more recently in the United States, the method has been disputed; psychiatrists and psychoanalysts have especially questioned its value, primarily because it attempts to reduce complex phenomena to simple terms and tends to ignore important individual differences. Nevertheless, employed with restraint, the method is indispensable as a bridge between the two extremes, and it is in fact often used unknowingly by both statisticians and case students.

Experimental methods. A controlled experiment involves taking two closely related situations or groups, subjecting one of them to a specific change and comparing the subsequent characteristics of both. In the past, so-called experiments by judicial, penal, and reformatory institutions were not really controlled or even experimental in the scientific sense, for public agencies, at least in theory, are bound by the idea of justice to give equal treatment to equals, not one kind of treatment to one group and another kind to another group. Thus, generally speaking, most controlled experiments must be left to universities and other private bodies, and indeed the need for strict control and variable treatment has been recently accepted in such researches as Harvard University's Cambridge-Somerville Youth Study, which sought the effects of counselling on "pre-delinquent" boys.

Problems of crime prediction

Prediction studies. Criminological prediction—not unlike actuarial prediction used by insurance companies—is intended to forecast, usually in percentages, the future conduct of persons under certain conditions. Based on statistics or case histories or both, the predictions attempt to indicate probabilities—how any specific individuals or groups are likely to be affected by certain conditions or treatments. Thus, for example, various categories of criminals are listed as likely to be recidivistic.

The techniques involved in constructing prediction tables are too complicated to be discussed in brief; they have been developed and refined in the past 40 years mainly by Sheldon and Eleanor Glueck of Harvard University and also by several other authors in various countries. Statistical prediction by itself can never be conclusive; it must be subjected to rigorous validation for any individual or group, and even then it can merely show certain probabilities, which should be used in penal decisions only with the greatest caution and along with the lessons of experience derived from other sources. Nevertheless, the method can be valuable in supplementing the inevitably limited personal experience of judges and administrators, and indeed in recent decades prediction research has probably nowhere in the social sciences been more popular and urgent than in criminology.

Action research. Action research, which is often contrasted with experimental research, consists of drawing upon the observations of field workers and other persons directly involved with delinquents, potential delinquents, or prisoners. Thus, for example, have social workers attempted to help slum children and adolescents with their problems and at the same time studied their delinquent behaviour, related it to their environment, and evaluated the results of the youth clubs or other services offered. The chief values of action research are that it aims at practical results through collaboration with fieldworkers, tries to build a bridge between theoretical and practical work, and may well dispense with formal hypotheses and simply aim at preventive tactics. Its best known and perhaps most successful example has so far been Clifford Shaw's Chicago Area Project, which, in close cooperation with the famous ecological studies of the University of Chicago, has tried to enlist suitable local people to deal with the social problems of their area.

Sociological research. Sociological research involves various methods—general surveys and personal interviews, as well as statistical, case-study, typological, experimental, and predictive techniques—and thus the purpose of classifying sociological research separately is chiefly to signal its focus or fields of interest. Mainly, criminology derives benefits from three fields of sociological study: (1) social institutions, involving such things as different conceptions of, or attitudes toward, property held by various societies or groups or the different effects of mass media on crime; (2) social groups, involv-

ing such things as the influence of juvenile gangs or criminal subcultures on individual criminal behaviour or the influence of prejudice on certain racial, national, or religious minorities; and (3) ecology, involving the study of different geographical areas and their rates and kinds of crime. Clifford Shaw's ecological studies of Chicago have been especially revealing in analyzing urban areas that either "breed" crime or "attract" crime.

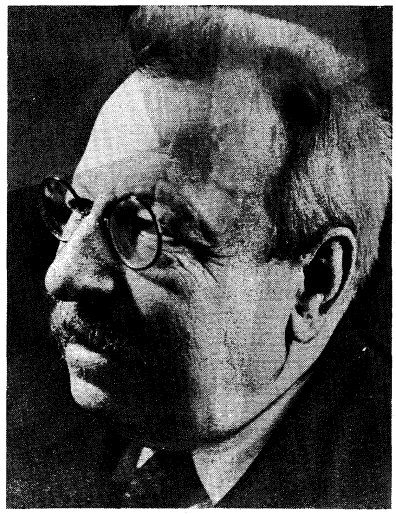
BIBLIOGRAPHY. H. MANNHEIM, *Comparative Criminology*, 2 vol. (1965), covers the whole field suggested by this article; M.E. WOLFGANG and F. FERRACUTI, *Il comportamento violento* (1966; Eng. trans., *The Subculture of Violence*, 1967), a work dealing mainly with the theme of its title but also with several methodological questions; H. MANNHEIM (ed.), *Pioneers in Criminology* (1960), critical discussions of the work of some famous criminologists and penologists; L. RADZINOWICZ, *Ideology and Crime* (1966), a survey containing much historical material; J.T. SELLIN and M.E. WOLFGANG, *The Measurement of Delinquency* (1964), an attempt to find an operational definition of serious offenses and to construct a "crime index"; H. MANNHEIM, *Social Aspects of Crime in England Between the Wars* (1940), a detailed analysis of the structure, interpretation, and contents of English criminal statistics; F.H. MCCLINTOCK and N.H. AVISON, *Crime in England and Wales* (1968), partly an updating of the preceding work; L.T. WILKINS, *Social Deviance* (1964), an attempt to bridge the gap between social research and social action from the viewpoint of the statistician; H. MANNHEIM and L.T. WILKINS, *Prediction Methods in Relation to Borstal Training* (1955), a theory of prediction, with an introductory historical survey of criminological prediction studies in various countries; S. and E. GLUECK, *Predicting Delinquency and Crime* (1959), *Ventures in Criminology* (1964), surveys of the Gluecks' prediction studies. Current information may be found in the following periodicals: *Journal of Criminal Law, Criminology, and Police Science* (quarterly); *British Journal of Criminology; Delinquency and Deviant Social Behaviour* (quarterly); *International Review of Criminal Policy* (issued irreg.); and *Canadian Journal of Corrections* (quarterly).

(H.M.)

Croce, Benedetto

Benedetto Croce, the foremost Italian philosopher of the first half of the 20th century, left his mark under a number of titles—historian, humanist, philosopher, critic, statesman; but even more influential was his stature as a person. Moral courage in defying a pretentious and oppressive political regime and again in rebuilding a nation demoralized by defeat in war made him the symbol of the spirit of modern Italy, of all that justifies its existence as a nation.

H. Roger-Viollet



Croce.

Croce belonged to a family of landed proprietors with estates in the Abruzzi region of central Italy but chiefly resident in Naples. His background was religious, monarchical, and conservative. Born at the Abruzzi town of Pescasseroli on February 25, 1866, Croce spent almost his whole life in Naples, becoming intimately identified with and a keen observer of its life and a biographer of its heroes. His life, of which he left a too-modest record in his autobiography, falls roughly into four phases; each develops the dual theme of his intellectual and moral growth and his gradual, ever-deepening identification with the moral character and destiny of the Italian nation.

First phase: agony and work

Early life and maturity. The first phase (until about 1900) was the period of Croce's agony. Orphaned (with his brother Alfonso) by the earthquake of Casamicciola in 1883, his life became, in his words, a "bad dream." The stable world of childhood and youth was shattered; he was forever marked; henceforth, he was a solitary figure, despite his considerable activity in the world.

His salvation lay in work. Disillusioned with the university, he set out upon an austere course of study, to become one of the great self-taught students of history. His writings of this period are universally alert, intelligent, and engaging; although limited in scope, they show a fine sobriety of style, as well as wit, irony, and a fiery polemical spirit; lyricism, which he eulogized, eluded him. Ostensibly, he had little taste for politics; actually, several basic attitudes were forming. Disillusioned with the nationalistic liberal leaders of the period following the Risorgimento (the 19th-century movement for Italian unity), he began to develop his own convictions on how an ethical, democratic, liberal government should be structured. He "coquetted"—according to his autobiography—with Socialism (which generally advocates public ownership and administration of the means of production and the distribution of goods) and Marxism (which adds the notion of class struggle), eventually freeing himself from this infatuation, typically, by a thorough examination and severe criticism of both positions. Nevertheless, he was subject to a constant and profound malaise. Subliminally, he desired but saw no public relevance for his activity; the limited world of erudition palled on him.

Second phase: founding of *La Critica*

He was delivered from this malaise, and the second phase of his life was opened in 1903 by the founding of *La Critica*, a journal of cultural criticism, in which, during the course of the next 41 years, he published nearly all his writings and reviewed all of the most important historical, philosophical, and literary work that was being produced in Europe at the time. At this same time he began the systematic exposition of his "Philosophy of the Spirit," his chief intellectual achievement. This term designates two distinct, but related, phases of his thought: (1) In the first phase, philosophy of spirit designates the construction of a philosophical system on the remote pattern of the Rationalism of classical Romantic philosophy. Its principle is the "circularity" of spirit within the structure of the system and in historical time. The phases, or moments, of spirit in this system are theoretical and practical; they are distinguished, respectively, into aesthetic, logical, and economic and ethical. The circular dynamic moves between both the lesser and the greater moments. The law of this circularity is that of absolute immanence. This system is documented in the volumes *Estetica come scienza dell'espressione e linguistica generale* (1902; *Aesthetics as Science of Expression and General Linguistic*, 1953), *Logica come scienza del concetto puro* (1909; *Logic*, 1917), *Filosofia della practica. Economia ed etica* (1909; *Philosophy of the Practical: Economic and Ethic*, 1967), and *Teoria e storia della storiografia* (1917; *History: Its Theory and Practice*, 1960). (2) Croce gradually abandoned, without explicitly renouncing, this schematism in response primarily to methodological considerations in history. Its moments are not dissolved but are concretized into the flow of historical action and thought. History becomes the unique mediational principle for all the moments of spirit, while spirit—*i.e.*, human consciousness—is completely spontaneous, without a predetermined structure. This change is signaled by the publication of *La storia come pensiero e come azione* (1938; *History as the Story of Liberty*, 1941). To this phase some have attached the term historical positivism, but Croce himself has called it absolute historicism and identified it as the definitive form of his thought. The

philosophy of spirit in its asystematic form produced the effective method of Croce's later work, as in the anthology *Filosofia, poesia, storia* (1951; *Philosophy, Poetry, History*, 1966).

According to Croce, "The foundation of *La Critica* marked the beginning of a new period in my life, the period of maturity or harmony between myself and reality." Through this journal he found the larger public theatre he sought. "*La Critica* was the most direct service I could render to Italian culture. . . . I was engaged in politics in the broad sense . . . uniting the role of a student and of a citizen." Through *La Critica* Croce's public role as teacher of modern Italy emerged. Count Camillo Benso di Cavour, the prime minister who presided over the formation of a unified Italy, had said, "Having made Italy, we must make Italians." *La Critica* took up this task.

The image of the Italian which animates this work is severe and beautiful. Creative effort, a passion for freedom united to a profound sense of civic duty, a life-style purged of all rhetoric and sentimental romanticism, unambiguous norms of public and private truth, a sense of history united to an obligation to the future, unceasing but constructive self-criticism: these were its elements. This image strongly reflected the personal ideal that Croce had gradually formed for himself. But history was preparing to put this ideal to the test.

Third phase: struggle with Fascism

Later life. The test was to be Fascism, the political attitude that places the nation or race at the centre of life and history and disregards the individual and his rights. So gradual was this preparation that Croce himself did not at once perceive it. He confessed that he first saw in Fascism a movement to the right of the political spectrum that might restrain and counteract the leftist tendencies toward unrestricted individual freedom released by World War I. But as the character of the regime revealed itself, his opposition hardened, becoming absolute, beyond compromise. He became, within and without Italy, the symbol of the opposition to Fascism, the rallying point of the lovers of liberty. Fascist leaders recognized the quality of his opposition; they accepted the fact that against such moral determination, their own methods were powerless. In Fascism Croce saw not merely another form of political tyranny. He saw it as the emergence of that other Italy, in which egoism displaced civic virtue, rhetoric dislodged poetry and truth, and the pretentious gesture replaced authentic action.

His consciousness of his role as the moral teacher of Italy was strengthened. Instruction now took the form of the composition of the great histories—a history of Europe in the 19th century, of Italy from 1871 to 1915, and of the Kingdom of Naples. Their didactic character was unmistakable; in them Croce pointed out how the historical path of Italy had become *la via smarrita* ("the lost way"). Moreover, the lesson was intended for Europe and for the entire Western world as well.

In the maelstrom of conflict and ambiguity that followed defeat, a voice of moral authority that could speak for the true Italy was demanded. Croce's was unanimously recognized as that voice. And with authority that voice recalled Italy to the inner spiritual resources through which it might renew itself. It matters little that Croce's own project for the rebuilding of Italy—the retention of the monarchy with certain dynastic changes, the return to the principles of a revived Liberal party in government—was not the one realized in history. More important is the fact that the new Italy, in its democratic form, was inspired by his spirit.

Fourth phase: return to studies

This last public office fulfilled, Croce returned to his beloved studies. In his own library—one of the finest collections in Europe within its own scope—he established the Istituto Italiano per gli Studi Storici (Italian Institute for Historical Studies) as a research centre. The rhythm of his own life became again what it had always been: the rhythm of peaceful, laborious days. Asked his state of health, he replied with true stoic equanimity, "I am dying at my work." Work and death were the order of that last day, November 20, 1952.

Character and significance. One traumatic experience may well determine the cast of a man's character. This was Croce's case and the traumatic event was the catastrophe of Casamicciola.

That night established him in the deep solitude that was the true key to his character. This solitude became in turn the rock upon which his absolute devotion to freedom rested. The mindless havoc wrought by nature on that night made clear to him that whatever meaning man's life might have must be the creation of his own free spirit, of his own action and decision. Because his solitude thus expressed itself in creative freedom, it did not isolate him. It provided the solid point from which he could survey all human life, suffering, and striving with universal comprehension and sympathy. His sympathy took the form not of melting compassion but of fraternal didacticism; it made him the critic, the teacher. And his solitude, moreover, sheltered a sensitivity both aesthetic and moral, both rare and unfailing; it flowered in his response to the basic beauty of all the creations of man's spirit; to him they were sacred. Inevitably, the final form of his thought and of his sentiment of life became "the religion of liberty." Religion alone could express the reverence in which he held the creative spirit of man, locked in the inviolate precinct of every individual but flowering in creations of universal meaning. This freedom alone could be ultimate and sacred.

BIBLIOGRAPHY. FAUSTO NICOLINI, *L'"editio ne varietur" delle opere di Benedetto Croce* (1960), the most scholarly bibliography; SILVANO BORSARI, *L'opera di Benedetto Croce: bibliografia* (1964), a very complete chronological listing that does not, however, supersede Nicolini.

Biography: FAUSTO NICOLINI, *Benedetto Croce* (1962), a quasi-official study, in Italian, written by a very close associate (not definitive, though sometimes spoken of as such); CECIL SPRIGGE, *Benedetto Croce: Man and Thinker* (1952), a compact, most informative account by a long-time friend and interpreter of Croce to the English-speaking world.

Works about Croce: CARLO ANTONI, *Commento a Croce* (1955), a careful analysis and criticism of salient points in Croce's thought by one of his most serious followers; ADRIANO BAUSOLA, *Etica e politica nel pensiero di Benedetto Croce* (1966), a careful and critical analysis of the most important single theme in Croce's philosophy; A. ROBERT CAPONIGRI, *History and Liberty: The Historical Writings of Benedetto Croce* (1955), considered by reviewers the most objective presentation of the whole range of Croce's historical writing, with careful evaluations relative to his theory of historiography; GIAN N.G. ORSINI, *Benedetto Croce: Philosopher of Art and Literary Criticism* (1961), the best book in English on Croce as literary critic.

(A.R.C.)

Crocodilia

The crocodiles constitute an order of the vertebrate class Reptilia. They are generally large, ponderous, amphibious animals, somewhat lizardlike in appearance, and carnivorous in habit. They have powerful jaws with many conical teeth and short legs with clawed, webbed toes. The tail is long and massive and the skin thick and plated. About 20 species are recognized.

The group is of particular interest because of its evolutionary position: the crocodiles are the last living link with the dinosaur-like reptiles of prehistoric times. They are, at the same time, the nearest living relatives of the birds. A large variety of crocodile fossils have been discovered; three of the four suborders of Crocodilia are extinct. On the basis of this extensive fossil record, it has been possible to establish well-defined relationships between the crocodiles and other vertebrate groups.

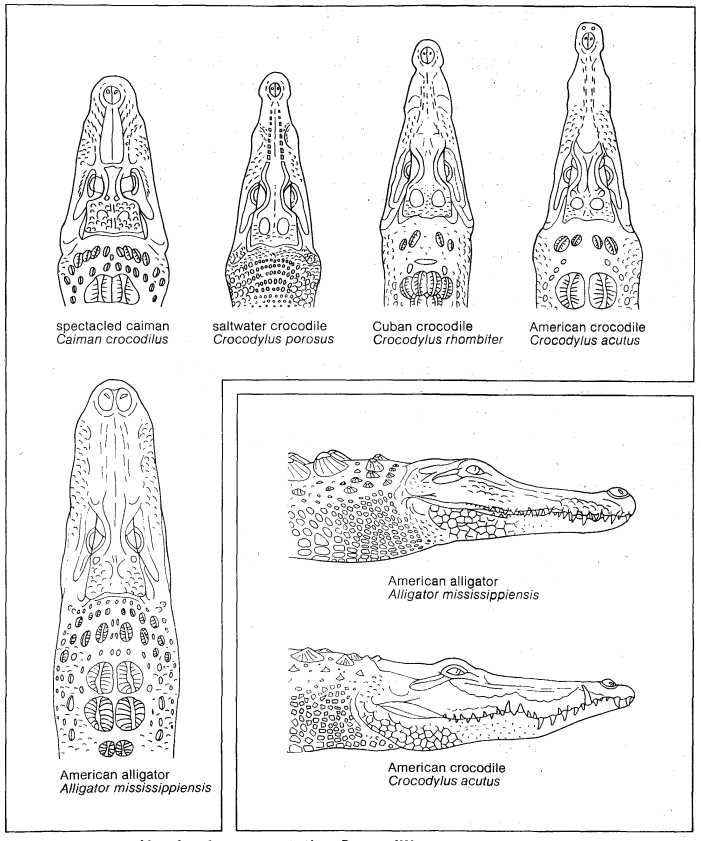
GENERAL FEATURES

Size range and diversity of structure. The crocodiles are the largest and the heaviest of present-day reptiles. In former times the Nile crocodile (*Crocodylus niloticus*) and the estuarine crocodile (*Crocodylus porosus*) attained a length of almost nine metres (about 30 feet), but today, specimens rarely exceed six metres (20 feet). Other species, for example, the smooth-fronted caiman (*Paleosuchus*) and the dwarf crocodile (*Osteolaemus tetraspis*) are about 1.7 metres (six feet) in length.

All crocodiles have a relatively long snout, or muzzle, which varies considerably in proportions and shape. The

large horny plates that cover most of the body generally are arranged in a regular pattern. Thick, bony plates occur on the back. The families and genera may be distinguished by anatomical features, principally those of the skull. Species are identified principally by the proportions of the snout; by the bony structures on the dorsal, or upper, surface of the snout; and by the number and the arrangement of the large neck scutes.

From *Mitteilungen aus dem Zoologischen Museum in Berlin*



Heads of representative Crocodilia.

Distribution and abundance. The habitat of the crocodile is mainly the tropics and subtropics of the northern and southern hemispheres. The Mississippi, or American, alligator (*Alligator mississippiensis*), the American crocodile (*Crocodylus acutus*), and the Chinese alligator (*Alligator sinensis*) are the only species found outside the tropics.

The true crocodiles (family Crocodylidae) occur in most of Africa south of the Sahara, Madagascar, India, Ceylon, Southeast Asia, the East Indies, northern Australia, Mexico and Central America, the West Indies, and most of South America east of the Andes and north of the mouth of the Río de la Plata. The caimans are confined to South America. The gavial occurs in India.

Declining populations

As recently as several decades ago, crocodiles were plentiful in much of the tropics and subtropics, but today they are almost extinct in many areas, as the inroads of civilization continue to limit their natural habitat. Thousands are killed annually by man, either for sport or for their valuable skins, which provide leather for handbags, luggage, shoes, belts, and other articles. In some localities they are killed because of their depredation of domestic cattle. The governments of some countries have sought by legislation to protect the remaining crocodile populations from hunters.

The disappearance of the Nile crocodile from parts of Africa has resulted in an overabundance of the catfish *Clarias*, which in turn has greatly diminished the supply of popular food fishes. In an attempt to restore the original balance, crocodiles are bred and raised on farms. When they have grown to a length of about 1.5 metres (5 feet) the crocodiles are released to their natural habitat.

NATURAL HISTORY

Life cycle. The young crocodile emerges from the egg with a length of 20–25 centimetres (8–10 inches). At first it remains concealed at the edge of its water habitat in order to avoid various predators. Principal among these are fishes and birds, but larger crocodiles also prey upon the young. During the first three to four years, the young increase in length by about 30 centimetres (about one foot) per year. The growth rate then gradually decreases, but growth can continue throughout life. Sexual maturity occurs at about ten years of age and at a body length of about 1.5 metres (five feet).

Of the little information available on the longevity of crocodiles, most has been gained from observation of animals in captivity. Captive animals seldom live more than 40 years, possibly because of lack of proper exercise and diet. Statistics on the growth of the Nile crocodile suggest, however, that specimens about 6 metres (20 feet) long are probably much more than 100 years old.

Behaviour. Crocodiles are predators, mostly nocturnal (*i.e.*, active at night), and spend most of their time in the water; they are also known to take rather long journeys over land. In their first weeks of life crocodiles eat mostly worms and water insects, then frogs and tadpoles; finally, their main diet is fish. Older crocodiles are more apt to prey upon waterfowl and on mammals, and occasionally a member of one of the larger species eats a human. This happens so infrequently, however, that crocodiles cannot be generally regarded as maneaters.

Diet

Crocodiles capture water animals in their jaws with a sideways movement of the muzzle. To catch land animals they remain motionless at the edge of a water hole from which the prey habitually drink, or they float passively in the water, resembling a drifting log. With a swift blow of the tail, they knock unsuspecting prey into the water. A number of crocodile species grip the legs of the victim in their jaws, then rotate themselves rapidly in the water, thus tearing the prey apart. When a crocodile cannot consume all of a victim at one time, it drags the carcass into its burrow.

The burrow is dug more or less horizontally at or just above the waterline and may extend for several metres, eventually ending in a chamber. The Mississippi alligator and the China alligator enter a state of inactivity (*i.e.*, hibernate) in these chambers during cold periods.

During the day, crocodiles often lie at the water's edge to sun themselves, frequently in large numbers. Otherwise, crocodiles live as lone individuals and establish individual territories. The extent of the territory is apparently defined for the animal's neighbours by its loud, vibrant roar, answered in kind by the crocodiles in adjacent territories.

When it roars, the crocodile tenses the musculature of its body so that the head and tail rise high out of the water; the flanks may vibrate so violently that water is sprayed high into the air from each side. Roaring can be provoked by similar noises, such as a deep note from a trumpet or the sound of artillery or supersonic aircraft. Crocodiles are also capable of deep grunting sounds, which apparently play a role during courtship, and a warning hiss. The young make squeaking sounds.

Intelligence

Of all reptile brains, the crocodile brain is the most highly developed. Crocodiles often show curiosity and are capable of being developed into tame animals—both attributes being measures of intelligence. If kept in captivity from birth, individuals of some species are known to recognize their keepers, show neither fear nor aggressiveness, beg for food, and permit themselves to be petted. In other words, they learn to adapt their behaviour to the unnatural captive environment.

Locomotion. The principal style of locomotion is that of swimming, in which the crocodile places its legs back against the sides of the body and moves forward by means of lateral wavelike motions of the tail. In walking on land, crocodiles hold themselves high on all four legs, and the body moves in waves, with a lateral swinging motion. When startled, crocodiles are able to run short distances, the two front legs moving forward and backward together as do also the two hindlegs, some-

what in the manner of a hopping rabbit. When moving quickly into the water from a bank, crocodiles slide on their bellies, pushing with the feet.

Reproduction. The sexes are outwardly different only in the Indian gavial (*Gavialis gangeticus*), in which the males, during the reproductive period, have a bulbous knob at the tip of the muzzle. Copulation occurs in the water and lasts about ten minutes. It is preceded by a courtship in which the animals rub their muzzles against each other and over the neck of the partner. The male then mounts the back of the female, and both animals rotate their tails so that the respective bodily openings are brought into contact. The distensible male reproductive organ is then inserted into the female.

All crocodiles lay hard-shelled eggs, which may number more than 100, depending upon the age and size of the female. The female builds a nest as a shelter for the eggs. The Nile crocodile female digs a trench, which it refills with dirt after laying the eggs. The female estuarine crocodile builds a mound of mud and decaying plant material, in the centre of which are the eggs. With her tail the female splashes water onto the nest. This promotes the heat-generating process of vegetative decay. The trench and mound types of nest are extremes, between which intermediate types are made by the various other species. In every case the female remains close to the nest and protects the eggs from predators until the eggs hatch.

The nest

After two or three months the young are fully developed and ready to hatch. While still in the egg they utter squeaking sounds, perhaps signalling that they are ready to emerge. The female then removes the dirt or other debris from the eggs but thereafter provides no further care for her offspring.

Ecology. Crocodiles are mainly inhabitants of swamps, lakes, and rivers, although some species make their way to brackish water or to the sea. The estuarine (or saltwater) crocodile (*C. porosus*), which lives almost entirely in the ocean, may swim miles out to sea. Such crocodiles venture upriver only as far as the limit of the salt water. The smooth-fronted caiman (*Paleosuchus*) in South America prefers rocky, fast-flowing rivers. In West Africa the dwarf crocodile (*Osteolaemus tetraspis*) is found principally in the rivers of the forest regions.

FORM AND FUNCTION

The form of the crocodile is adapted to its amphibious way of life. The elongated body with its long, muscular paddle tail is well suited to rapid swimming. Other features, enumerated below, are also adapted to the animal's amphibious habit.

Sensory structures

The external nostril openings, the eyes, and the ear openings are the highest parts of the upper side of the head. These important sense organs thus remain above the water surface even when the rest of the head is submerged. The two nostril openings are close together on a raised portion of the point of the muzzle and may be closed by membranous flaps so that no water can enter when the animal dives. A long nasal passage enclosed in bone leads from the exterior nostril openings to the interior nostril openings, or choanae, located at the extreme posterior end of the palate; a membranous flap in front of the choanae constitutes the posterior closure of the mouth cavity. Thus, the crocodile can breathe even if its mouth is open under water.

Like many nocturnal animals, crocodiles have eyes with vertical, slit-shaped pupils; these narrow in bright light and widen in darkness, thus controlling the amount of light that enters. On the back wall of the eye the so-called tapetum lucidum reflects the incoming light, thus utilizing small amounts of light to the best advantage. In addition to the protection provided by the upper and lower eyelids, the nictitating membrane, a thin, translucent structure, may be drawn over the eye from the inner corner while the lids are open. The delicate eyeball surface is thus protected under water, but a certain degree of vision is still possible.

Unlike the ears of other modern Reptilia, those of the crocodile have a movable, external membranous flap. By means of this structure the crocodile is able to protect its ears from the water.

The crocodile's relatively flat snout is usually quite long; in some species it is extremely elongated. The outer margin of the jaws in most species is jagged. Each jaw carries a row of sharp teeth, which may number more than 100 in species with very long muzzles. The teeth are held in sockets and replaced continuously, new ones growing from below and old ones being forced out. On the floor of the mouth is the thick, fleshy tongue, firmly attached and therefore almost immobile.

The posterior portion of the head forms a flat plate. Here the short, powerful neck is attached. On its upper side are two groups of knobby scales: the smaller postoccipital knobs (absent only in the estuarine crocodile); and the large nuchal knobs, which in some species may be connected to the adjacent horny plates of the back.

Body plates

The upper surfaces of the back and tail are covered with large, rather rectangular horny plates arranged regularly in longitudinal and transverse rows. Most of the dorsal plates have a longitudinal ridge, or keel. Under these plates lie bony plates of about the same size, except in the estuarine crocodile, in which the bony plates are much smaller.

The entire underside of the crocodile usually also has regularly arranged horny plates, which are smaller than those on the upper surface, entirely smooth, rectangular, and contain little or no bone material. An exception to this condition occurs in caimans of the genera *Melanosuchus*, *Caiman*, and *Paleosuchus*, in which the surface plates on the lower side are just as bony as those on the back. Slightly posterior to the attachment of the hindlegs, on the underside of the base of the tail, lies the anus, which extends longitudinally, surrounded by an oval area of small scales.

In contrast to the back and belly, the sides of the body have mostly small knobby scales. The flanks are thus distensible, a necessary condition for breathing and for the expansion of the body that occurs in the pregnant female.

The legs are short but powerful. The forefeet have five toes—the usual number for Reptilia. In their anatomical structure, however, the forelegs differ markedly from those of other reptiles. This departure suggests that crocodiles developed from ancestors with degenerate forelegs; it also supports the possibility of a close evolutionary relationship between the crocodiles and the birds, in which the forelimbs have been considerably modified. The hindlegs are more powerfully developed than the front pair. The hind feet have only four toes, which are wholly or partially webbed.

On its flat upper side the tail carries a comb which, in the anterior portion, is double, consisting of high leaf-shaped scales. Near the middle of the tail the two combs merge.

The heart of the crocodile is markedly different in structure from that of other reptiles: the two auricles and the two ventricles are completely separate.

Arterial–venous connection

There exists, nevertheless, a connection between the arterial and venous circulation by way of the so-called Panizzae foramen, which opens between the two vessels leading separately from the ventricles. This connection is necessary for equalization of pressure differences between the arterial and venous circulatory systems that arise during long dives in deep water.

EVOLUTION AND CLASSIFICATION

Paleontology. The crocodile skull exhibits distinctly developed upper and lower temporal (*i.e.*, behind the eye sockets) openings; the teeth arise from sockets, and the roof of the skull lacks an opening for the parietal organ, a median, dorsal outgrowth of the brain. The crocodiles thus show the most important characteristics of the group that includes the dinosaurs (Archosauria). Within the Archosauria the crocodiles are a separate order, since they have developed a secondary bony palate, which encloses the nasal passage from the exterior nasal openings to the choanae (internal nostrils).

These features occur even in the most primitive repre-

sentatives of the crocodile group, namely the Protosuchia of the Upper Triassic Period (about 190,000,000–200,000,000 years ago); but their muzzles were very short, and the choanae were relatively far forward on the palate. As the crocodiles continued to evolve, the openings of the choanae tended to move further back. In the Mesosuchia of the Jurassic (136,000,000–190,000,000 years ago) and Cretaceous periods (65,000,000–136,000,000 years ago)—to which the long-snouted ocean crocodiles also belong—the choanae had already moved to the posterior part of two bones of the skull (pterygoids). In the true crocodiles (Eusuchia), which appear in the Upper Jurassic, the choanae are entirely enclosed by the pterygoids. In modern species they have moved to the posterior border of the palate. In the Sebecosuchia of the Upper Cretaceous to the Miocene Epoch (7,000,000–26,000,000 years ago), a branch collateral to that of the crocodiles, the skull is laterally flattened, and the choanae lie in a depression in the anterior part of the pterygoids.

Distinguishing taxonomic features. The families and genera of the order Crocodilia are differentiated primarily by the anatomical peculiarities of their skulls. The classification of the species is based mainly upon external characteristics, such as the proportions of the snout, the bony structures on the dorsal side of the snout, the number of teeth, the number and arrangement of the large knobs on the nape of the neck, and the characteristics of the dorsal plates.

Annotated classification. Extinct groups represented only by fossils are indicated by a dagger (†).

ORDER CROCODILIA
Heavy cylindrical body; large, triangular head; legs short, toes webbed; long, muscular tail; large flat plates on belly, keeled ones on back; heart 4-chambered.

†Suborder Protosuchia
Upper Triassic; muzzle very short; choanae (internal nostrils) in region of palatine bone.

†Suborder Mesosuchia
Jurassic to Upper Cretaceous; choanae in posterior part of pterygoid bone.

†Suborder Sebecosuchia
Upper Cretaceous to Miocene; skull laterally flattened; choanae in depression in anterior part of pterygoids.

Suborder Eusuchia
Upper Jurassic to Recent; choanae entirely enclosed by pterygoids.

Family Alligatoridae (alligators)
Four genera and 7 species; teeth of lower jaw fit inside those of upper jaw.

Family Crocodylidae (true crocodiles)
Three genera and 13 species; teeth of upper and lower jaws form one interdigitating row when mouth is closed.

Family Gavialidae (gavial)
One genus and 1 species; extremely long snout, more than 22 teeth in each jaw; nasal bones separated from premaxillaries.

Critical appraisal. One authority has separated the order Crocodylomorpha into two suborders, Crocodylia and Paracrocodylia. According to this scheme the Crocodylia include as infra-orders those groups given above as suborders. This scheme also contains a suborder, Thalattosuchia.

Widely different views prevail concerning the classification of the living groups of Eusuchia—*i.e.*, the alligators, the true crocodiles, and the gavials. Of these, the alligators and the true crocodiles, without doubt more closely related to each other than to the gavials, are sometimes regarded as constituting two subfamilies of the family Crocodylidae. Some authors regard the gavials as a third subfamily. Others give the false gavial, or Sunda gavial, a special position with respect to the true crocodiles. For conciseness, these three groups have been treated here as distinct families.

BIBLIOGRAPHY. M.M. COHEN and C. GANS, "The Chromosomes of the Order Crocodilia," *Cytogenetics*, 9:81–105 (1970), pictures of karyotypes of all recent species of crocodiles and an attempt to interpret them from the viewpoint of evolution; H.B. COTT, "Scientific Results of an Inquiry into the Ecology and Economic Status of the Nile Crocodile (*Crocodilus niloticus*) in Uganda and Northern Rhodesia," *Trans. Zool. Soc. Lond.*, 29:211–356 (1961), one of the most recent and comprehensive accounts of the behaviour and territorialism of the Nile crocodile; O. KUHN, *Die vorzeitlichen Krokodile* (1968), a survey of all recent and fossil crocodiles and a proposal of a common classification for both; F.J. MEDEM, "The Crocodilian Genus *Paleosuchus*," *Fieldiana, Zool.*, 39: 227–247 (1958), the first comprehensive field work on smooth-fronted caimans, especially on their ecology; W.T. NEILL, *The Last of the Ruling Reptiles: Alligators, Crocodiles, and Their Kin* (1971), a richly illustrated and extremely careful compendium of the biology of all recent crocodiles; A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966), the classic work on the classification and phylogeny of the vertebrates and their mutual relations; A.D. WALKER, "A Revision of the Jurassic reptile *Hallopus victor* (Marsh), with Remarks on the Classification of Crocodiles," *Phil. Trans. R. Soc.*, Series B, 257:323–372 (1970), a discussion of the division of crocodiles (Crocodylomorpha) in the suborders Crocodylia and Paracrocodylia based on anatomical examination of some fossil forms; H. WERMUTH, "Systematik der rezenten Krokodile," *Mitt. Zool. Mus. Berl.*, 29:375–514 (1950), a review of recent crocodiles and a discussion of the taxonomic importance of their characteristics, with figures of heads and skulls of all species; "Farbwechsel und Lernfähigkeit bei Krokodilen," *Dt. Aquar.-Terrar.-Z.*, 16:90–92 (1963), observations on the intelligence of a tame, seven-foot, spectacled caiman; O. VON WETTSTEIN, "Crocodilia," *Handb. Zool.*, 7:236–424 (1931), a complete monograph of the recent crocodiles, with an extensive bibliography.

(H.F.We.)

Cro-Magnon Man

Cro-Magnon man is the name that was first given to several prehistoric skeletons found in a rock shelter at Cro-Magnon, near Les Eyzies-de-Tayac, Dordogne, France, in 1868. Sent to the site, French geologist Louis Lartet began excavations in which he established the existence of five archaeological layers covered with ash. The age of the human remains found in the topmost layer—along with worked flint and the bones of animals of species now extinct—is Upper Paleolithic (35,000–10,000 years ago), but the attribution of these to a clearly defined Upper Paleolithic culture is less definite. Traditionally regarded as Aurignacian, since typically Aurignacian artifacts were found in the rock shelter, they could be more recent, and one study suggests that they should be assigned to the Perigordian (a separate industry covering approximately the same time period as the Aurignacian), which would suggest an age of about 25,000 BC.

On the basis of the Cro-Magnon remains and those of Neanderthal man, which had been discovered a little earlier (1856), a general overview of prehistoric man was presented in 1882, in which the Cro-Magnon remains were asserted to be representative of a Cro-Magnon race.

Examination and analysis of Cro-Magnon fossils. Although it appears that, at the time of discovery, the remains of about 15 individuals existed at Cro-Magnon, only fragments from some six individuals were preserved, among them the cranium and mandible of a male about 50 years old. Considered representative of the Cro-Magnon type, this specimen is known as the Old Man of Cro-Magnon. Also preserved were skull fragments of about four other individuals, some bits of bone from a fetus or newborn child, and an assortment of bones attributed variously to the individuals mentioned above. The first subject mentioned, the Old Man of Cro-Magnon, has been regarded as typical of the Cro-Magnon race.

The skull is longheaded and as seen from above has a pentagonal outline, with outward bulging of the parietal bones (at the sides of the skull). The forehead is straight, the brow ridges only slightly projecting, the cranial vault noticeably flattened, and the occipital bone (at the back of the head) projects backward. The cranial capacity is large, about 1,600 cubic centimetres (about 100 cubic inches). Although the skull is relatively long and narrow, the face appears quite short and wide. This is often regarded as a common feature of the Cro-Magnon race.



Reconstruction of the appearance of Cro-Magnon man.
By courtesy of the American Museum of Natural History, New York

Typical skull features

The forward projection of the upper jaw (maxilla) is also distinctive. The eye sockets are low-set, wide, and rather square in shape; and the nasal aperture of the skull is narrow and strongly projecting. The mandible is robust, with massive ascending ramus (the upward projection of the lower jaw, where it attaches to the skull), strongly developed points of muscular attachment, and a quite prominent chin.

The root of only one molar tooth remains in the jaw of the Old Man of Cro-Magnon, a fact that contributed to the idea of his extreme age. In fact, it is probable that the loss of the majority of his teeth occurred after death. The teeth of the other individuals found at Cro-Magnon, which are similar to the teeth of other fossil humans classed as Cro-Magnon, show that the dentition of Cro-Magnon man was nearly identical to that of modern man. Most of the teeth, however, especially the last molars, are distinctly larger than those of most modern peoples. Dental caries is sometimes apparent, and tooth wear was often extreme.

The remainder of the Cro-Magnon skeleton is not fully known from the remains found at the original site, which are incomplete and poorly preserved. Skeletal material attributed to the Cro-Magnon race from other sites, however, affords the general impression of robustness, probably combined with powerful musculature. The forearm is relatively long, as is the thigh; the femur (thighbone) has a very prominent linea aspera (a bony ridge that runs lengthwise down the back of the femur), and the tibia is flattened from back to front (platycnemy). The hand skeleton is large with short fingers, and the foot has a prominent heel.

Cro-Magnon man's height

Early investigators were impressed by the stature of Cro-Magnon man, as some reconstructions suggest that the Old Man of Cro-Magnon may have been as much as 1.90 metres (six feet three inches) tall. A recent restudy, however, suggests that the stature of the original Cro-Magnon remains varied from 1.66 to 1.71 metres (five feet five inches to five feet seven inches). The stature of several skeletons from the Grimaldi Caves (in Italy, near the French frontier), which show clear affinities to those of Cro-Magnon, was noticeably greater, with an average height of 1.77 metres (five feet eight inches). It is thus reasonable to conclude that, on the whole, the Cro-Magnon peoples were relatively tall.

Lesions noted in the Cro-Magnon skeletal remains have been attributed to wounds, but one analysis has suggested that these lesions are pathological in origin and may have resulted from the action of a toxic mushroom, *Actinomyces israeli*.

Variant Cro-Magnon types. Two French prehistorians, A. de Quatrefages and Ernest Hamy, in 1882 took the Cro-Magnon fossils to be prototypes of a Cro-Magnon (or Les Eyzies) race. As opposed to the Neanderthal (or Cannstadt) race, the Cro-Magnon were then considered to be the most ancient form of *Homo sapiens*. To the Cro-Magnon race were assigned other remains discovered before 1868 at Paviland, Glamorganshire; Engis, Belgium; La Madeleine, France; and Bruniquel, France. Subsequently, further finds of human skeletal remains extended the geographical range of the Cro-Magnon peoples through much of Europe and into Asia and North Africa.

Relation of Cro-Magnon man to earlier fossil forms

The place of Cro-Magnon man in human evolution. The question of the relation of Cro-Magnon man to earlier forms of *Homo sapiens* is still unclear. Thousands of years before the appearance of Cro-Magnon peoples, certain modern traits are seen in human remains of Middle Pleistocene age (500,000–100,000 years old). Such remains, however, are rare and usually fragmentary, so the question of the relationship of Cro-Magnon peoples to the Middle Pleistocene remains, often assigned to an early form of *Homo sapiens*, and to the distinctive Neanderthal peoples who preceded Cro-Magnon man in Europe, is still to be settled.

Perhaps as complex as the question of origin is that of the duration of Cro-Magnon culture and the disappearance of Cro-Magnon man. The latter question is, in a sense, more a matter of genetics than of descriptive anthropology, since it is assumed that the Cro-Magnon race or type was simply absorbed into later European populations. Individuals nearly resembling Cro-Magnon types are found in the Mesolithic (in Europe, 8000–*c.* 5000 BC), for example, at Muge, Portugal, and in the Neolithic (in Europe, roughly from 5000 BC to about 2000 BC). Human groups, more or less homogeneous and structured, that have retained a close relationship to Cro-Magnon types, at least in their cranial morphology, are still in existence today. Such groups have been noted and studied in France, England, Spain, and North Africa; but the most remarkable examples are the Dal race from Dalecarlia (now Dalarna, Sweden) and the Guanches of the Canary Islands, who may represent the last relatively pure Cro-Magnon stock.

The culture of Cro-Magnon man. The ties between *Homo sapiens*, and particularly Cro-Magnon peoples and the various Upper Paleolithic cultures (Castelperronian, Aurignacian, Gravettian, etc.—based on classification of stone and bone tools), are not entirely clear. It is currently impossible to establish a general outline of physical types and cultures for this period. There do appear, however, to be some detectable differences between populations of different late Pleistocene ages in western Europe and between roughly contemporaneous populations in western Europe as contrasted with those of central or eastern Europe.

Cro-Magnon tools and weapons

Toolmaking. The Cro-Magnon peoples are generally associated with the Aurignacian culture tool industry, and perhaps with the Gravettian (also called upper Perigordian). The Aurignacian tool industry is characterized by retouched blade tools, end scrapers and "nosed" scrapers, burins (chisel-like tools), and fine bone tools, in particular long, flat points (spearheads) with cleft bases. Other bone and reindeer-horn implements are also seen: awls, tools for smoothing or scraping leather, and the so-called *bâtons de commandement*—bars of antler or bone with holes drilled in them, the use of which is still uncertain, but perhaps used for straightening arrow or spear shafts.

Dwellings. The dwellings of Cro-Magnon man were most often caves and shelters made by rock overhangs, but it is apparent that huts were also made; sometimes these were simply lean-tos against rock walls, but foundation stones and "pavements" of stone in the shape of houses are evidence of complete huts. These houses are not a new development with the Cro-Magnon people, however; both the Neanderthal people and earlier peoples of the Middle Pleistocene are associated with similar remains. It seems probable that the Cro-Magnon peoples lived fairly settled lives. Studies of occupation sites and the types and extent of remains found in these sites sug-

gest that the rock shelters were inhabited all through the year rather than seasonally, and it is likely that these Paleolithic hunters moved their homes only when hunting or environmental conditions forced them to do so.

Animals hunted

Hunting techniques. The climate in the habitable parts of Europe at the time of Cro-Magnon man was cool to cold. Plants and animals of types associated with tundra and steppe environments were usual. Bone remains found at Cro-Magnon occupation sites indicate that they were successful hunters of such animals as reindeer, bison, wild horse, and even mammoth. As yet, very little is known of Cro-Magnon hunting methods—individual or collective, use of the bow, use of traps, etc. It is obvious from the animal remains, however, that hunting techniques must have been efficacious.

Aesthetics and religion. Although earlier human groups certainly had religious practices of some sort—the Neanderthal people buried their dead, a practice merely continued and elaborated by Cro-Magnon and later peoples—and no doubt had some appreciation of aesthetics as well, the first examples of prehistoric art are Cro-Magnon. Small engravings, reliefs, and sculptures of animals have been found, as well as a few later statuettes of ivory or stone and occasional engravings in stone of female figures. These figures are usually large-breasted, wide-hipped, and most often apparently pregnant; they are assumed to be some sort of fertility symbol, perhaps used in religious or magical rituals intended to promote the fertility of the group. The many fine paintings of animals found in caves in France and Spain are generally considered to have magical efficacy also, perhaps to ensure the presence of game, perhaps to aid the hunters in catching it; but it is by no means true that art for these prehistoric people had only magical or religious significance. The Cro-Magnon people also appreciated the decorative aspects of art, as demonstrated by their frequent use of animal pictures and simple geometrical designs to ornament tools and weapons. Moreover, whatever the motivation of such art, it is surprising by its quality. Whether it is a great composition of cave art or a carved or painted art object, it does not represent a first artistic attempt but artistic work that reaches the level of the masterpiece, by its artistic realism or by its power of movement. Sculpture like the "Venus of Brassempouy" or the head of a neighing horse at Le Mas-d'Azil (Ariège) or the bison at La Madeleine are worthy of a place among the masterpieces of world art.

"Venus" figurines and animal paintings

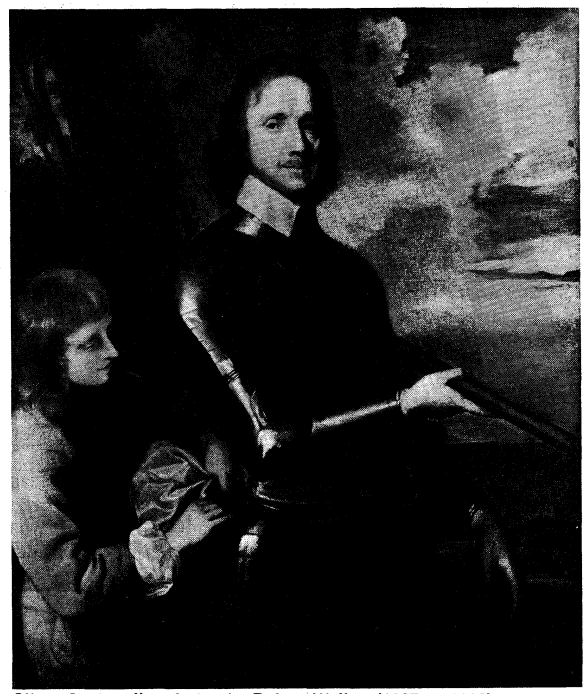
BIBLIOGRAPHY. M. BOULE and H.V. VALLOIS, *Les hommes fossiles*, 5th ed. rev. (1957; Eng. trans., *Fossil Men: A Textbook of Human Paleontology*, 1957), a well-illustrated general introduction to the field of fossil man, including an extended discussion of the Cro-Magnon peoples and the history of their discovery; J. PIVETEAU (ed.), *Traité de paléontologie*, vol. 7, *Primates, paléontologie humaine* (1957), a basic encyclopaedic source on primate and human evolution, extensively illustrated; F.C. HOWELL, *Early Man* (1965), a general and popular treatment, well-illustrated, of the field of early man studies; F. BORDES, *La Paléolithique dans le monde* (1968; Eng. trans., *The Old Stone Age*, 1968), an introduction to the material culture and the occupation places of Pleistocene peoples; *L'homme de Cro-Magnon, anthropologie et archéologie* (1970), a series of papers by specialists in physical anthropology and prehistoric archaeology on the skeletal biology and culture of the Cro-Magnon peoples; A. LEROI-GOURHAN, *Préhistoire de l'art occidental* (1965; Eng. trans., *Treasures of Prehistoric Art*, 1967), a magnificently illustrated volume on the art of the Cro-Magnon peoples; H.V. VALLOIS and G. BILLY, "Nouvelles recherches sur les hommes fossiles de l'abri de Cro-Magnon," in *L'Anthropologie*, vol. 69 and 73 (1965, 1969), the most thorough study of the skeletal biology of the human remains from the Cro-Magnon shelter; W.W. HOWELLS, *Mankind in the Making*, rev. ed. (1967), an extremely well-written popular account of human evolution.

(H.J.De.)

Cromwell, Oliver

Oliver Cromwell, an English soldier and statesman of outstanding gifts and a forceful character shaped by a devout Calvinist faith, was lord protector of the republican Commonwealth of England, Scotland, and Ireland from 1653 to 1658. One of the leading generals on the parliamentary side in the English Civil War against King Charles I, he helped to bring about the overthrow of the Stuart monarchy, and, as lord protector, he raised his country's status once more to that of a leading European power from the decline it had gone through since the death of Queen Elizabeth I. Cromwell was one of the most remarkable rulers in modern European history; for although a convinced Calvinist, he believed deeply in the value of religious toleration. At the same time his victories at home and abroad helped to enlarge and sustain a Puritan attitude of mind both in Great Britain and in North America that continued to influence political and social life until recent times.

By courtesy of the National Portrait Gallery, London



Oliver Cromwell, painting by Robert Walker (1607–c. 1660). In the National Portrait Gallery, London.

YOUTH AND EARLY PUBLIC CAREER

Cromwell was born at Huntingdon in eastern England on April 25, 1599, the only son of Robert Cromwell and Elizabeth Steward. His father had been a member of one of Queen Elizabeth's parliaments and, as a landlord and justice of the peace, was active in local affairs. Robert Cromwell died when his son was 18, but his widow lived to the age of 89. Oliver went to the local grammar school and then for a year attended Sidney Sussex College, Cambridge. After his father's death he left Cambridge to look after his widowed mother and sisters but is believed to have studied for a time at Lincoln's Inn in London, where country gentlemen were accustomed to acquire a smattering of law. In August 1620 he married Elizabeth, daughter of Sir James Bourchier, a merchant in the City of London. By her he was to have five sons and four daughters.

Upbringing

Formative influences. Both his father and mother came from Protestant families who had profited from the destruction of the monasteries during the reign of King Henry VIII, and it is probable that they influenced their son in his religious upbringing. Both his schoolmaster in Huntingdon and the Master of Sidney Sussex College were enthusiastic Calvinists and strongly anti-Catholic. In his youth Cromwell was not notably studious, being fond of outdoor sports, such as hunting; but he was an avid reader of the Bible, and he admired Sir Walter Raleigh's *The History of the World*. From his teachers and from his reading Cromwell learned that the sins of man were punishable on earth but that God, through His Holy Spirit, could guide the elect into the paths of righteousness.

During his early married life Cromwell, like his father,

was profoundly conscious of his responsibilities to his fellow men and concerned himself with affairs in his native fenlands, but he was also the victim of a spiritual and psychological struggle that perplexed his mind and damaged his health. He does not appear to have experienced conversion until he was nearly 30; later he described to a cousin how he had emerged from darkness into light. Yet he had been unable to receive the grace of God without feeling a sense of "self, vanity and badness." He was convinced that he had been "the chief of sinners" before he learned that he was one of God's Chosen.

Career as a country squire

Cromwell also had financial worries until, at the age of 39, he inherited property at Ely from his mother's brother. Like other lesser gentry, he contended with bad harvests and a variety of taxes and impositions, such as ship money, exacted by the monarchy not only to pay for the upkeep of the navy but to sustain the lavish tastes of the court. Though in 1628 he had been elected a member of Parliament for the borough of Huntingdon, King Charles I dissolved this Parliament in 1629 and did not call another for 11 years.

During the interval, country gentlemen like Cromwell accumulated grievances. The Cromwell family was but one of a network of dissatisfied gentry who belonged to what one might call the political nation: for example, John Hampden, the wealthy Buckinghamshire squire who brought a test case against the crown over the levying of ship money, was Cromwell's first cousin. Thus, when in the spring of 1640 Cromwell was elected member of Parliament for the borough of Cambridge, partly because of the important social position he held in Ely and partly because of his fame as "Lord of the Fens," he found himself among a host of friends at Westminster who, led by John Pym, a veteran politician from Somerset, were highly critical of the monarchy. Little was achieved by the Short Parliament (dissolved after three weeks), but, when in November 1640 Cromwell was again returned by Cambridge to what was to be known as the Long Parliament, which sat until 1653, his public career began.

Cromwell in Parliament. Cromwell had already become known in the Parliament of 1628–29 as a fiery and somewhat uncouth Puritan, who had launched an attack on Charles I's bishops. He believed that the individual Christian could establish direct contact with God through prayer and that the principal duty of the clergy was to inspire the laity by preaching. Thus he had contributed out of his own pocket to the support of itinerant Protestant preachers or "lecturers" and openly showed his dislike of his local bishop at Ely, a leader of the High Church party, which stood for the importance of ritual and episcopal authority. He criticized the bishop in the House of Commons and was appointed a member of a committee to investigate other complaints against him. Cromwell, in fact, distrusted the whole hierarchy of the Church of England, though he was never opposed to a state church. He therefore advocated abolishing the institution of the episcopate and the banning of a set ritual as prescribed in *The Book of Common Prayer*. He believed that Christian congregations ought to be allowed to choose their own ministers, who should serve them by preaching and extemporaneous prayer. Though he shared the grievances of his fellow members over taxes, monopolies, and other burdens imposed on the people, it was his religion that first brought him into opposition to the King's government. When in November 1641 John Pym and his friends presented to King Charles I a "Grand Remonstrance," consisting of over 200 clauses, among which was one censuring the bishops "and the corrupt part of the clergy, who cherish formality and superstition" in support of their own "ecclesiastical tyranny and usurpation," Cromwell declared that had it not been passed by the House of Commons he would have sold all he had "the next morning, and never have seen England more; . . ."

Attacks on the church

The Remonstrance was not accepted by the King, and the gulf between him and his leading critics in the House of Commons widened. A month later Charles vainly attempted to arrest five of them for treason: Cromwell was not yet sufficiently prominent to be among these. But when in 1642 the King left London to raise an army, and events drifted toward civil war, Cromwell began to distinguish himself not merely as an outspoken Puritan but also as a practical man capable of organization and leadership. In July he obtained permission from the House of Commons to allow his constituency of Cambridge to form and arm companies for its defense, in August he himself rode to Cambridge to prevent the colleges from sending their plate to be melted down for the benefit of the King, and as soon as the war began he enlisted a troop of cavalry in his birthplace of Huntingdon. As a captain he made his first appearance with his troop in the closing stages of the Battle of Edgehill (October 23, 1642) where Robert Devereux, 3rd earl of Essex, was commander in chief for Parliament in the first major contest of the war.

MILITARY AND POLITICAL LEADER

During 1643 Oliver acquired a reputation both as a military organizer and a fighting man. From the very beginning he had insisted that the men who served on the parliamentarian side should be carefully chosen and properly trained, and he made it a point to find loyal and well-behaved men regardless of their religious beliefs or social status. Appointed a colonel in February, he began to recruit a first-class cavalry regiment. While he demanded good treatment and regular payment for his troopers, he exercised strict discipline. If they swore, they were fined; if drunk, put in the stocks; if they called each other Roundheads—thus endorsing the contemptuous epithet the Royalists applied to them because of their close-cropped hair—they were cashiered; and if they deserted, they were whipped. So successfully did he train his own cavalrymen that he was able to check and re-form them after they charged in battle. That was one of Cromwell's outstanding gifts as a fighting commander.

Military organization

Throughout 1643 he served in the eastern counties that he knew so well. These formed a recognized centre of parliamentary strength, but, unwilling to stay on the defensive, Cromwell was determined to prevent the penetration of Yorkshire Royalists into the eastern counties and decided to counterattack. By re-forming his men in a moment of crisis in the face of an unbeaten enemy, he won the Battle of Gainsborough in Lincolnshire on July 28. On the same day he was appointed governor of the Isle of Ely, a large plateau-like hill rising above the surrounding fens, that was thought of as a possible bastion against advancing Royalists. In fact, however, Cromwell, fighting alongside the parliamentary general Sir Thomas Fairfax, succeeded in stemming the royalist attacks at Winceby in Lincolnshire and then successfully besieged Newark in Nottinghamshire. He was now able to persuade the House of Commons, well pleased with these victories, to create a new army, that would not merely defend eastern England but would march out and attack the enemy.

This new army was formed under the command of Edward Montagu, 2nd earl of Manchester, early in 1644. Appearing in the House of Commons, Cromwell, besides commending Manchester for the command, accused some of his fellow officers as incompetents or being "profane" and "loose" in their conduct. Although not all members of the House of Commons approved of Cromwell's using his political position to defame other officers, his friends rallied round him, and in 1644 he was appointed Manchester's second in command, with the rank of lieutenant general, and paid five pounds a day. After an alliance had been concluded with the Scots, he was also appointed a member of the Committee of Both Kingdoms, which became responsible for the overall strategy of the Civil War. But since he was engaged at the front during the campaigning season, Cromwell took little part in its deliberations.

After Manchester's army had stormed Lincoln in May 1644, it marched north to join the Scots and the Yorkshire parliamentarians at the siege of York. But Charles I's commander in chief, Prince Rupert, raised the siege. He was, however, defeated in the Battle of Marston Moor, July 2, 1644, that in effect gave the north of En-

gland to Parliament. Cromwell had again distinguished himself in the battle, and when Manchester's army returned to eastern England to rest on its laurels, Cromwell criticized his superior officer for his slowness and lethargy. He did not believe that Manchester really wanted to win the war, and in mid-September he laid his complaints before the Committee of Both Kingdoms. The quarrel was patched up, but after the defeat at Newbury, caused largely by the earl of Manchester's refusal to support Cromwell's cavalry with his infantry, it broke into the open once more.

Removal of Lord Manchester

Cromwell now expounded his detailed complaint about Manchester's military conduct in the House of Commons. Manchester retorted by attacking Cromwell in the House of Lords. It was even planned to impeach Cromwell as "an incendiary." Once again, however, these quarrels were patched up. In December 1644, Cromwell proposed that in the future no members of either house of Parliament should be allowed to hold commands or offices in the armed forces; his proposal was accepted, and it was also agreed that a new army should be constituted under Sir Thomas Fairfax. Cromwell, an admirer of Fairfax, put forward his name and then busied himself with planning the new army, from which, as a member of Parliament, he himself was excluded. But, significantly, the post of second in command was left open, and, when the Civil War reached its climax in the summer of 1645, Fairfax insisted that Cromwell should be appointed to it. He then fought at the battles of Naseby and Langport, where Charles I's last two field armies were destroyed. In January 1646 the House of Commons awarded Cromwell £2,500 a year in confiscated Royalist land for his services and renewed his commission for a further six months. Thus he was able to join Fairfax in the siege of Oxford, from which Charles I escaped before it surrendered.

Cromwell was delighted with the way in which the war had gone since Fairfax had taken command of the new army and the lethargic earls of Essex and Manchester had been removed from their commands. He attributed these victories to the mercy of God and demanded that the men who had served the country so faithfully should have their due reward. After Naseby he wrote to the Speaker of the House of Commons urging that such "honest men" should not meet with discouragement: "He that ventures his life for the liberty of his country, I wish he trust God for the liberty of his conscience, and you for the liberty he fights for."

Postwar discord

But once the war was over the House of Commons wanted to disband the army as cheaply and quickly as possible. Disappointed, Cromwell told Fairfax in March 1647 that "never were the spirits of men more embittered than now." He devoted himself to trying to reconcile the Parliament with the army and was appointed a parliamentary commissioner to offer terms on which the army could be disbanded except for those willing to take part in a campaign in Ireland. As late as May he thought that the soldiers might agree to disband but that they would refuse to serve in Ireland and that they were "under a deep sense of some sufferings." When the civilian leaders in the House of Commons decided that they could not trust the army and ordered it disbanded, while they hired a Scottish army to protect them, Cromwell, who never liked the Scots and thought that the English soldiers were being disgracefully treated, left London and on June 4, 1647, threw in his lot with his fellow soldiers.

Mediation and the Second Civil War. For the remainder of this critical year he attempted to find a peaceful settlement of the kingdom's problems, but his task seemed insoluble; and soon his good faith was freely called into question. The army was growing more and more restive, and on the day Cromwell left London, a party of soldiers seized Charles I. Cromwell and his son-in-law, Henry Ireton, interviewed the King twice, trying to persuade him to agree to a constitutional settlement that they then intended to submit to Parliament. At that time Cromwell, no enemy of the King, was touched by his devotion to his children. His main task, however, was to overcome the general feeling in the army that neither the King nor Parliament could be trusted. When, under pressure from the rank and file, General Fairfax led the army toward the houses of Parliament in London, Cromwell still insisted that the authority of Parliament must be upheld; and in September he also resisted a proposal in the House of Commons that no further addresses should be made to the King. Just over a month later he took the chair at meetings of the General Council of the Army (which included representatives of the private soldiers known as Agitators), assured them that he was not committed to any particular form of government and had not had any underhand dealings with the King. On the other hand, fearing anarchy, he opposed extremist measures such as the abolition of the monarchy and the House of Lords and the introduction of a more democratic constitution.

Cromwell as mediator

But all Cromwell's efforts to act as a mediator between army, Parliament, and King came to nothing when Charles I escaped from Hampton Court Palace, where he had been kept in honourable captivity, and fled to the Isle of Wight to open negotiations with Scottish commissioners offering to restore him to the throne on their terms. On January 3, 1648, Cromwell abandoned his previous position and, telling the House of Commons that the King was "an obstinate man, whose heart God had hardened," agreed to a vote of no addresses, which was carried. The Royalists, encouraged by the King's agreement with the Scots and the failure of Cromwell to unite Parliament and the army, took up arms again and the Second Civil War began.

The arrest of Charles I

General Fairfax first ordered Cromwell into Wales to crush a rising there and then sent him north to fight the Scottish army that invaded England in June. Though his army was inferior in numbers to that of the Scots and northern Royalists, he defeated them both in a campaign in Lancashire; then he entered Scotland and restored order there; finally he returned to Yorkshire and took charge of the siege of Pontefract. The correspondence he conducted during the siege with the governor of the Isle of Wight, whose duty it was to keep watch on the King, reveals that he was increasingly turning against Charles. Parliamentary commissioners had been sent to the island in order to make one final effort to reach an agreement with the King. But Cromwell told the governor that the King was not to be trusted, that concessions over religion must not be granted, and that the army might be considered a lawful power capable of ensuring the safety of the people and the liberty of all Christians.

While Cromwell, still not entirely decided on his course, lingered in the north, his son-in-law Ireton and other officers in the southern army took decisive action. They drew up a remonstrance to Parliament complaining about the negotiations in the Isle of Wight and demanding the trial of the King as a Man of Blood. While Cromwell still felt uncertain about his own views, he admitted that his army agreed with the army in the south. Fairfax now ordered him to return to London; but he did not arrive until after Ireton and his colleagues had removed from the House of Commons all members who favoured continuing negotiations with the King. Cromwell asserted that he had not been acquainted with the plan to purge the House, "yet since it was done, he was glad of it, and would endeavour to maintain it." Hesitating up to the last moment, Cromwell, pushed on by Ireton, by Christmas Day finally accepted Charles's trial as an act of justice. He was one of the 135 commissioners in the High Court of Justice and, when the King refused to plead, he signed the death warrant.

First chairman of the Council. After the British Isles were declared a republic and named the Commonwealth, Oliver Cromwell served as the first chairman of the Council of State, the executive body of a one-chamber Parliament. During the first three years following Charles I's execution, however, he was chiefly absorbed in campaigns against the Royalists in Ireland and Scotland. He also had to suppress a mutiny, inspired by a group known as Levellers, an extremist Puritan party said to be aiming at a "levelling" between rich and poor, in the Common-

Cromwell in Ireland

wealth army. Detesting the Irish as primitive, savage, and superstitious, he believed they had carried out a huge massacre of English settlers in 1641. As commander in chief and lord lieutenant, he waged a ruthless campaign against them, though when he refused quarter to most of the garrison at Drogheda near Dublin in September 1649, he wrote that it would "tend to prevent the effusion of blood for the future, . . . which otherwise cannot but work remorse and regret." On his return to London in May 1650 Cromwell was ordered to lead an army into Scotland, where Charles II had been acknowledged as its new king. Fairfax had refused the command; so on June 25 Cromwell was appointed captain general in his place. He felt more tender toward the Scots, most of whom were fellow Puritans, than toward the Catholic Irish. The campaign proved difficult, and during the winter of 1650 Cromwell was taken ill. But he defeated the Scots with an army inferior in numbers at Dunbar on September 3, 1650, and a year later, when Charles II advanced into England, Oliver destroyed his army at Worcester.

This battle ended the civil wars. Cromwell now hoped for pacification, a political settlement, and social reform. He pressed through an "act of oblivion" (amnesty), but the army became more and more discontented with Parliament. It believed that the members were corrupt and that a new Parliament should be called. Once again Cromwell tried to mediate between the two antagonists, but his sympathies were with his soldiers. When he finally came to the conclusion that Parliament must be dissolved and replaced, he called in his musketeers and on April 20, 1653, expelled the members from the House. He asserted that they were "corrupt and unjust men and scandalous to the profession of the Gospel"; two months later he set up a nominated assembly to take their place. In a speech on July 4 he told the new members that they must be just, and, "ruling in the fear of God," resolve the affairs of the nation.

The "Little Parliament"

Cromwell seems to have regarded this "Little Parliament" as a constituent body capable of establishing a Puritan republic. But just as he had considered the previous Parliament to be slow and self-seeking, he came to think that the Assembly of Saints, as it was called, was too hasty and too radical. He also resented the fact that it did not consult him. Later he described this experiment of choosing Saints to govern as an example of his own "weakness and folly." He sought moderate courses and also wanted to end the naval war begun against the Dutch in 1652. When in December 1653, after a coup d'etat planned by Major General John Lambert and other officers, the majority of the Assembly of Saints surrendered power into Cromwell's hands, he decided reluctantly that Providence had chosen him to rule. As commander in chief appointed by Parliament, he believed that he was the only legally constituted authority left. He therefore accepted an "Instrument of Government" drawn up by Lambert and his fellow officers by which he became lord protector, ruling the three nations of England, Scotland, and Ireland with the advice and help of a council of state and a Parliament, which had to be called every three years.

ADMINISTRATION AS LORD PROTECTOR

Before Cromwell summoned his first Protectorate Parliament on September 3, 1654, he and his Council of State passed more than 80 ordinances embodying a constructive domestic policy. His aim was to reform the law, to set up a Puritan Church, to permit toleration outside it, to promote education, and to decentralize administration. The resistance of the lawyers somewhat dampened his enthusiasm for law reform, but he was able to appoint good judges both in England and Ireland. He was strongly opposed to severe punishments for minor crimes, saying: "to see men lose their lives for petty matters . . . is a thing that God will reckon for." For him murder, treason, and rebellion alone were subject to capital punishment. During his Protectorate, committees known as Triers and Ejectors were set up to ensure that a high standard of conduct was maintained by clergy and schoolmasters. In spite of resistance from some members of his council Cromwell readmitted Jews into the country. He concerned himself with education, was an excellent chancellor of Oxford University, founded a college at Durham, and saw to it that grammar schools flourished as they had never done before.

Foreign and economic policies. In 1654 Cromwell brought about a satisfactory conclusion to the Anglo-Dutch War, which, as a contest between fellow Puritans, he had always disliked. The question then arose of how best to employ his army and navy. His Council of State was divided, but eventually he resolved to conclude an alliance with France against Spain. He sent an amphibious expedition to the Spanish West Indies, and in May 1655 Jamaica was conquered. As the price for sending an expeditionary force to Spanish Flanders to fight alongside the French he obtained possession of the port of Dunkirk. He also interested himself in Scandinavian affairs; although he admired King Charles X of Sweden, his first consideration in attempting to mediate in the Baltic was the advantages that would result for his own country. In spite of the emphasis Cromwell laid on the Protestant interest in some of his speeches, the guiding motive in his foreign policy was national and not religious benefit.

Conquest of Jamaica

His economic and industrial policy followed mainly traditional lines. But he opposed monopolies, which were disliked by the country and had only benefitted the court gentry under Queen Elizabeth and the first two Stuarts. For this reason the East Indian trade was thrown open for three years, but in the end Cromwell granted the company a new charter (October 1657) in return for financial aid. Satisfactory methods of borrowing had not yet been discovered; hence—like those of practically all European governments of his time—Cromwell's public finances were by no means free from difficulties.

Relations with Parliament. When Cromwell's first Parliament met he justified the establishment of the Protectorate as providing for "healing and settling" the nation after the civil wars. Arguing that his government had prevented anarchy and social revolution, he was particularly critical of the Levellers who, he said, wished to destroy well-tested institutions "whereby England hath been known for hundreds of years." He believed that they wanted to undermine "the 'natural' magistracy of the nation" as well as "make the tenant as liberal a fortune as the landlord." He also thought that the spiritual anarchy that followed the destruction of the old Anglican Church had gone too far, for now ordained preachers were frequently interrupted or shouted down in their pulpits. A radical in some directions, such as in seeking the reform of the laws, Cromwell now adopted a conservative attitude because he feared that the overthrow of the monarchy might lead to political collapse.

But vociferous republicans, who became leaders of this newly elected Parliament, were unwilling to concentrate on legislation, questioning instead the whole basis of Cromwell's government. Cromwell insisted that they must accept the "four fundamentals" of the new constitution that, he argued, had been approved both by "God and the people of these nations." The four fundamentals were government by a single person and Parliament; the regular summoning of parliaments, which must not be allowed to perpetuate themselves; the maintenance of "liberty of conscience"; and the division of the control of the armed forces between the protector and Parliament. Oliver said that he would sooner be "rolled into my grave and buried with infamy, than I can give my consent" to the "wilful throwing away of this Government, . . . so owned by God, so approved by men." He therefore required all members of Parliament, if they wished to keep their seats, to sign an engagement to be faithful to a protector and Parliament and to promise not to alter its basic character. Except for 100 convinced republicans, the members agreed to do so but were still more concerned with rewriting the constitution than reforming the laws as desired by the protector. As soon as he could legitimately do so (January 22, 1655), Cromwell dissolved Parliament.

The four fundamentals

But with his second Parliament, which he convened in 1656, he encountered exactly the same difficulty in the

end, for the republican leaders, when they were allowed to resume their seats, tried to destroy the Protectorate on the ground that they were being forced to return to "an Egyptian bondage." Once again Cromwell emphasized that he had been "called" to power and that anarchy or an invasion from abroad would follow if his authority were not upheld. Thus in February 1658 he felt himself driven again to dissolve Parliament even though, as a former member, he understood only too well the gravity of his action.

Death and burial

Ever since the campaign in Ireland Cromwell's health had been poor. In August 1658, after his favourite daughter, Elizabeth, died of cancer, he was taken ill with malaria and taken to London with the intention of living in St. James's Palace. But he died in Whitehall at three o'clock on September 3, the anniversary of two of his greatest victories. His body was secretly interred in Westminster Abbey on November 10, 13 days before his state funeral. In 1661, after the restoration of King Charles II, Cromwell's embalmed remains were dug out of the tomb and hung up at Tyburn where criminals were executed. His body was then buried beneath the gallows. But his head was stuck on a pole on top of Westminster Hall where it is known to have remained until the end of Charles II's reign.

Assessment. Oliver Cromwell was by no means an extreme Puritan. By nature he was neither cruel nor intolerant. He cared for his soldiers, and when he differed from his generals he did not punish them severely. (For example, when he dismissed John Lambert he gave him a generous pension.) He was devoted to his old mother, his wife, and family. (The stories spread by Royalists that he was an admirer of a number of ladies have little substance to them.) While he concerned himself with the spiritual welfare of his children because he believed that "often the children of great men have not the fear of God before their eyes," he committed the mistake of not preparing for the practical tasks of government his eldest son, Richard, whom in the last days of his life he nominated to succeed him as protector. Music and hunting were among his recreations. He delighted in listening to the organ and was an excellent judge of horses. He was known to smoke, to drink sherry and small beer, and to prefer English food; he permitted dancing at the marriage of his youngest daughter. In his younger days he indulged in horseplay with his soldiers, but he was a dignified ruler. Sir Peter Lely, the famous Dutch painter, pictured him as he was in his prime (although the portrait was apparently not painted from life); the numerous paintings from life by Robert Walker dating from the beginning of the Civil War show him looking more of a fanatic.

Private life and religious beliefs

As lord protector, Cromwell was much more tolerant than in his fiery Puritan youth. Once bishops were abolished and congregations allowed to choose their own ministers, he was satisfied. Outside the church he permitted all Christians to practice their own religion so long as they did not create disorder and unrest. He allowed the use of *The Book of Common Prayer* in private houses and even the English Roman Catholics were better off under the protectorate than they had been before.

Although many Quakers were kept in prison for disturbing the peace, Cromwell was on friendly terms with George Fox, the founder of the society of Friends, and explored religious questions with him. When in the winter of 1656 a Quaker entered Bristol in imitation of Christ's entry into Jerusalem, Cromwell tried, though unsuccessfully, to save him from the fury of Parliament, which voted heavy punishments on the blasphemer. The year before, Oliver interviewed two of the leaders of the Fifth Monarchy Men, an extreme sect: he pointed out to them that they had been imprisoned for sedition but emphasized that no one would hinder them from preaching the Gospel of Christ.

Political views

In politics Cromwell held no fixed views except that he was opposed to what he called arbitrary government. Before the execution of Charles I he contemplated the idea of placing one of Charles's sons upon the throne. Oliver also resisted the abolition of the House of Lords. In 1647 he said that he was not "wedded and glued" to any particular form of government. After the Assembly of Saints failed, he summoned two elected parliaments (1654–55 and 1656–58), but he was never able to control them. His failure to do so has been attributed to "lack of that parliamentary management by the executive which, in correct dosage, is the essential nourishment of any sound parliamentary life" (H.R. Trevor-Roper). In between these two parliaments (1655–56) he sanctioned the government of the country by major generals of the Horse Militia who were made responsible for law and order in groups of counties. But he soon abandoned this experiment when it met with protests and reverted to more normal methods of government. In the spring of 1657 he was tempted by an offer of the crown by a majority in Parliament on the ground that it fitted in better with existing institutions and the English common law. In the end he refused to become king because he knew that it would offend his old republican officers. Nevertheless, in the last year and a half of his life he ruled according to a form of government known as "the Petition and Advice." This in effect made him a constitutional monarch with a House of Lords whose members he was allowed to nominate as well as an elected House of Commons. But he found it equally difficult to govern either with or without parliaments.

Reputation

Although in the late 17th century Cromwell was execrated as a brave bad man, it was admitted that he had made his country great. In the 18th century, on the other hand, he was considered a nauseating hypocrite, while the 19th century, under the influence of the writer and historian Thomas Carlyle, regarded him as a constitutional reformer who had destroyed the absolutism of Charles I. Modern critics are more discriminating. His belief in God's providence is analyzed in psychological terms. Marxists blame him for betraying the cause of revolution by suppressing the radical movement in the army and resisting the policy of the Levellers. On the whole, he is regarded only in a very limited sense as a dictator, but rather as a patriotic ruler who restored political stability after the civil wars and contributed to the evolution of constitutional government and religious toleration.

BIBLIOGRAPHY. SIR CHARLES H. FIRTH, *Oliver Cromwell and the Rule of the Puritans in England* (1900, many times reprinted), is generally accepted as the best and fairest biography. Other modern biographies include R.S. PAUL, *The Lord Protector* (1955), excellent on Cromwell's religious attitude; and MAURICE ASHLEY, *The Greatness of Oliver Cromwell* (1957 and 1966), embodying some of the results of research since Firth's day. CHRISTOPHER HILL, *God's Englishman* (1970), places Cromwell in perspective from a Marxist point of view. Other books on Cromwell include three by distinguished statesmen, French, American, and British: FRANCOIS GUIZOT, *Oliver Cromwell and the English Revolution*, 2 vol. (Eng. trans. 1854); THEODORE ROOSEVELT, *Oliver Cromwell* (1900); and JOHN MORLEY, *Oliver Cromwell* (1900). Cromwell's letters and speeches are collected in *The Writings and Speeches of Oliver Cromwell*, 4 vol., ed. by W.C. ABBOTT (1937–47); these volumes also contain a sound and detailed narrative of his career.

(M.As.)

Cromwell, Thomas

Thomas Cromwell, as principal adviser to King Henry VIII, was virtual ruler of England from 1532 to 1540. Purposeful, decisive, and essentially ruthless, Cromwell during these years established the English Reformation, confiscated the wealth of the monasteries, and effected a revolution in the administration of the kingdom. He was born in Putney, the son (probably) of Walter Cromwell, alias Smyth, a brewer, blacksmith, and cloth maker. His early life is obscure. It appears that he went abroad at an early age and spent some time in Italy. For several years after 1510 he was resident in the Low Countries, and he seems to have been closely connected with the London Merchant Adventurers. By 1520 he had entered Cardinal Wolsey's service as his solicitor, and from that time his career is well documented. Wolsey employed him in 1525 in the dissolution of some lesser monasteries, in which work he earned a good deal of dislike. The Car-

dinal, however, continued to favour him, and Cromwell soon became his confidential adviser.

By courtesy of The Frick Collection, New York



Thomas Cromwell, painting by Hans Holbein, the Younger (1497/98–1543). In the Frick Collection, New York.

When Wolsey fell into disgrace in 1529, Cromwell entered Parliament, where his remarkable ability attracted the notice of the King. For nearly three years he worked his way up in the royal favour, entering Henry's service early in 1530. He was sworn into the council toward the end of that year and reached the inner circle of confidential advisers a year later. All the time, he was establishing his ascendancy in the House of Commons. In 1532 he obtained office as master of the jewels. Other offices soon followed: principal secretary and master of the rolls in 1534 and lord privy seal in 1536. The last office was combined with a peerage, and he took the title of Lord Cromwell of Wimbledon.

Cromwell and the Reformation

Cromwell's part in the English Reformation has been much debated. At one time he was credited with supplying Henry with a complete plan of action as early as 1529; later it became usual to see in him nothing but the King's most competent executive agent. The truth seems to be that he was in no way in charge until early in 1532, taking over when the King's policy of forcing the Pope to come to terms had proved to be a failure. It was, to all appearances, Cromwell who then came forward with a clear notion of how to achieve Henry's purpose without the Pope. His policy consisted in making a reality of some large and vague claims to supreme power that Henry had uttered at intervals. He proposed to destroy Rome's power in England and to replace it by the royal supremacy in the church. He was behind the first attacks on the papacy (1532) and the act against the payment by bishops of their first year's revenue to Rome. He secured the submission of the clergy to the King in matters of legislation, and, in 1533 he secured the passage of the Act in Restraint of Appeals to Rome, preventing appeals to Rome in matrimonial and testamentary cases. Its preamble embodied his political theory of the sovereign national state. Thereafter he was in complete control of the government, though he remained careful to pretend to be acting on the King's authority. In 1534 he completed the erection of the royal supremacy with the passage of the Act of Supremacy.

Political and financial reasons counselling an attack on the monasteries, Cromwell was appointed the King's vicar general with powers to visit and reform all monastic institutions. Despite serious opposition, especially in the north, the task was carried out relentlessly. During 1536–40 the surrender of the greater houses was obtained by pressure and persuasion, and by 1540 all monastic institutions had ceased to exist and their property had been vested in the crown. Cromwell and other crown officials obtained valuable grants as rewards, but while the minister lived, the new wealth was not squandered.

In 1536 Cromwell was also appointed the King's deputy as head of the church. Cromwell's own religious views have been in much doubt. They certainly were not very strong, and his essentially secular temper subordinated religious to political considerations. Nevertheless, he came to be firmly associated with a radical policy of reform and Reformation. In the main, this resulted from difficulties abroad. While hostility between France and Spain had prevented foreign intervention during the critical years of the Reformation, 1533–36, there seemed a danger of an alliance against England after that date. Cromwell, whose forthright and clear-sighted temper was less well suited to the conduct of foreign affairs than Henry VIII's skillful opportunism, involved himself in projects of a Lutheran alliance distasteful to the King who wished to stand on Catholic orthodoxy. In 1539 Cromwell made the mistake of trying to force the King to his side by compelling him to marry Anne of Cleves. The King from the first hated his fourth wife, and by February 1540 it was clear that the alliance with the German princes that she represented was unnecessary. Thereafter, Cromwell's fall came quickly. He fought back for a few months, being created earl of Essex and lord great chamberlain in April 1540, but early in June his enemies persuaded Henry that his vicegerent was a heretic and a traitor. He was arrested on June 10, condemned without a hearing, and executed on July 28. His fall did not end the Reformation, but it marked the end of competent government and purposeful policy in Henry's reign.

Foreign policy

The basis of Cromwell's thought was the notion of the sovereign national state that in practice he established by the expulsion of the papacy. He was greatly interested in political theory, though it seems improbable that he knew Machiavelli's writings, as was once alleged. In his conception of the English state and monarchy, his central idea was that of the supremacy and omnipotence of statute, or (as it came to be called) legislative sovereignty of the king in Parliament. In other words, he wanted to establish unlimited sovereignty in the hands of a monarchy limited by dependence on consent. His work in Parliament—managing elections, drafting statutes, piloting legislation—makes him the first of a long line of English parliamentary statesmen. He also demonstrated his awareness of the need of providing practical management of a new kind. No minister before him, and few after, exercised such pervasive influence over every detail of administration. Cromwell began, and to a large extent carried through, a reconstruction that replaced administration by the king's household with a national administration divorced from the person of the king and dependent on civil service departments. This aspect of his work endured, through many reforms, until the great changes of the 19th century.

Accomplishments

BIBLIOGRAPHY. R.B. MERRIMAN, *The Life and Letters of Thomas Cromwell* (1902), prints the bulk of the surviving letters, but the biography, though full of relevant facts, is badly marred by the author's dislike of his subject and his little understanding of the age. A.G. DICKENS, *Thomas Cromwell and the English Reformation* (1959), provides a sympathetic and sensible picture, especially of Cromwell's contribution to the religious Reformation. Various aspects of Cromwell's career are dealt with in W.G. ZEEVELD, *Foundations of Tudor Policy* (1948), which throws light on Cromwell's political thought; G.R. ELTON, *The Tudor Revolution in Government* (1953), which analyzes his work as an administrator; and the same author's *Policy and Police* (1972), which deals with the problems of enforcement raised by the early Reformation and with Cromwell's dominant part in solving them.

(G.R.E.)

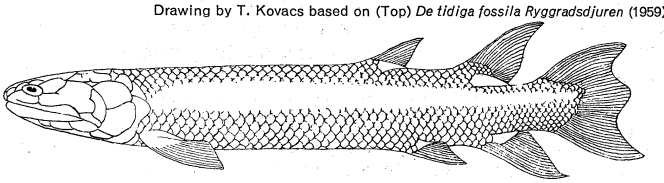
Crossopterygii

The Crossopterygii constitutes a largely extinct subclass of the bony fishes that appeared at the beginning of the Devonian Period (395,000,000 years ago) but are now represented only by the coelacanth *Latimeria chalumnae*.

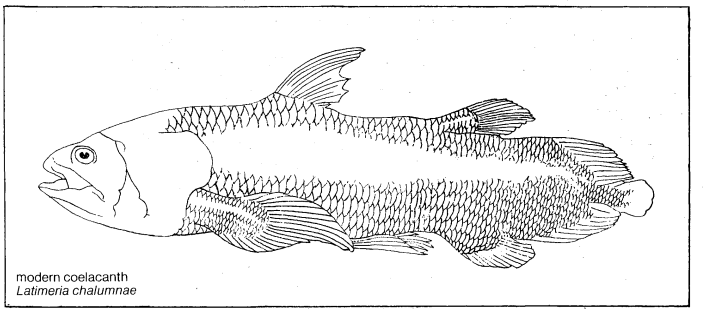
General features. One major trait of the subclass is the division of the skull into an anterior, or ethmosphe-

noidal, unit and a posterior, or oto-occipital, unit, similar to the two cartilaginous prototypes found in the embryonic cranium. A strong joint unites the two regions at each side. The base of the skull and the vertebral column, which are incompletely ossified, allow the persistence, to various degrees, of the initial skeletal axis, or notochord. The subclass comprises three orders: Rhipidistia, Actinistia, and Struniiformes. Some authorities consider the Crossopterygii to be an order and what are treated here as orders to be suborders (see FISH). After being widely distributed around the world in the Mesozoic Era, which began about 225,000,000 years ago, the crossopterygians underwent a rapid decline and then almost became extinct after the Triassic Period (190,000,000 years ago).

Drawing by T. Kovacs based on (Top) *De tidiga fossila Ryggradsdjuren* (1959)



Devonian rhipidistian
Eusthenopteron foordi



modern coelacanth
Latimeria chalumnae

Representative crossopterygians.

Importance in the evolution of land animals

The Rhipidistia, predatory fishes of the Mesozoic, were ancestral to the terrestrial vertebrates, lived in freshwater, and probably had two respiratory apparatuses, a branchial (gill) system for aquatic respiration and a pulmonary (lung) system for air breathing. To facilitate air breathing the nasal cavities were provided with posterior nares (nostrils) homologous with the primary choanae (internal openings to the pharynx) of more advanced vertebrates. The skeletal structure of the paired fins clearly shows the ability for locomotion both on solid ground and in the water. The rhipidistians are thus credited, in the history of vertebrate evolution, with having made the great transition in anatomy and physiology involved in the emergence from water and resulting in the evolution of the amphibians. A Swedish paleontologist, E. Jarvik, has suggested that the rhipidistian Osteolepiformes gave rise to the Stegocephalia (the extinct ancestors of reptiles, birds, and mammals) and the Anura (frogs) and that the suborder Porolepiformes gave rise to the Urodela (salamanders).

The Actinistia, especially the family Coelacanthidae, unlike the Rhipidistia, have exhibited exceptional evolutionary stability. The same fossil deposits contain both marine and freshwater types, both already specialized during the Devonian. They were thought to have disappeared 50,000,000 to 70,000,000 years ago, but in 1938 a live specimen was taken in the Indian Ocean. South African ichthyologist J.L.B. Smith identified it as a member of the Coelacanthidae and named it *Latimeria chalumnae*, the generic name in honour of Miss Courtenay Latimer, an associate who first brought the strange fish to his notice, the species name recalling its capture near the mouth of the Chalumna River. Between 1952 and 1970, more than 60 specimens of *Latimeria* were caught on the volcanic slopes of the Comoro Islands, at depths of 200 to 300 metres (650 to 1,000 feet).

Form and function. *Latimeria* has made it possible to reconstruct, with a high probability of accuracy, the anatomy of the coelacanths in general, in particular that of the perishable organs. Among the most striking characteristics are those of the head. The brain exhibits a relatively simple and harmonious structure, has an extremely small volume by comparison with the cranial capacity, and shows considerable displacement of the forebrain relative to the floor of the skull. The snout contains a special sensory organ, the rostral organ. At the intracranial articulation are attached some fibrous connective tissues as well as a pair of powerful longitudinal muscles, the subcranial muscles. This pair of muscles encloses the imposing notochord (a slender skeletal structure), the morphology of which must have been largely the same throughout the crossopterygian group. The heart of the *Latimeria*, which is very primitive, exhibits almost perfect bilateral symmetry (mirror-image form). It lies within a substantial pericardial cavity that retains the primitive continuity with the peritoneal (abdominal) cavity. There is a series of small valves near the exit from the heart, and several small contractile organs attached to the branchial arteries apparently fulfill the necessary function of assisting the propulsion of blood. On the whole, an embryonic condition co-exists with specialized arrangements.

An enormous cylinder of adipose (fat) tissue, aligned with a short median diverticulum (a blind pouch) of the ventral wall of the esophagus, lies above the abdominal organs. It apparently is a result of the degeneration of a lung apparatus. The extraordinary size of this cylinder is related to a displacement of the kidneys that undoubtedly occurs in the course of development, these organs occupying an unusual ventral position, posterior to the anus. Segments of the sympathetic nervous system are carried along in this movement.

Fins

The body is covered with large rough scales. The powerful tail fin has three lobes, lying in the median plane. The posterior end of the notochord extends into the middle lobe, which is by far the smallest. Two pairs of fins, the pectoral and the pelvic, are attached to their respective girdles. The base of each fin consists of a fleshy stalk supported by several successive segments of bone or cartilage that are not homologous with the similar parts of the Rhipidistia. Median fins similarly formed grow from the posterior part of the body, the posterior dorsal (above) and anal (below) fins. Finally, there is an anterior dorsal fin which, in contrast to the foregoing, is of ray-finned (actinopterygial) type—*i.e.*, lacking the fleshy, supportive stalk.

The Struniiformes, discovered only recently, lived in the Devonian. Their bony remains indicate considerable differences from both the Rhipidistia and the Actinistia. The fossil remains indicate, however, that they possessed the major characteristic of the subclass, the division of the cranium into an anterior and a posterior part.

BIBLIOGRAPHY. There is virtually no popular literature dealing with this group of fishes. The following are technical in nature: E. JARVIK, "On the Structure of the Snout of Crossopterygians and Lower Gnathostomes in General," *Zool. Bidr. Upps.*, 21:235–675 (1942); J.P. LEHMAN, "Crossopterygii," in *Traité de paléontologie IV*, 3:301–412 (1966); J. MILLOT, and J. ANTHONY, "*Latimeria chalumnae*, dernier des Crossoptérygiens," in *Traité de zoologie*, vol. 13, pp. 2553–2597 (1958); *Anatomie de* Latimeria chalumnae, 2 vol. (1958–66); J.L.B. SMITH, "A Living Coelacanthid Fish from South Africa," *Trans. R. Soc. S. Afr.*, 28:1–106 (1940); K.S. THOMSON, "The Comparative Anatomy of the Snout in Rhipidistian Fishes," *Bull. Mus. Comp. Zool. Harv.*, 131: 313–357 (1964).

(J.D.A./Ja.M.)

Crusades

The term crusade is commonly used to refer to military expeditions organized by Western Christians against Muslim powers in order to take possession of or maintain control over the Holy City of Jerusalem and the places associated with the earthly life of Jesus Christ. Between 1095, when the First Crusade was launched, and 1291, when the Latin Christians were finally expelled from their bases in Syria, historians have formally enumerated eight major expeditions. Many other lesser ven-

tures also took place, and even after 1291 there were attempts to recover what had been lost.

This period of roughly two centuries was one of significant social, economic, and institutional growth in western Europe. As a consequence, each of the Crusades reflected the particular conditions prevailing in Europe at the time, and their impact on Europe varied as new situations developed in the East.

This article is divided into the following sections:

I. The crusading movement and the first four Crusades
The First Crusade and the establishment of the Latin states
Background of the First Crusade
Preparations for the Crusade
The siege of Antioch
The siege of Jerusalem
The establishment of the crusader states
The era of the Second and Third Crusades
The Second Crusade
The crusader states to 1187
The institutions of the First Kingdom
The Third Crusade
The Fourth Crusade and the Latin Empire of Constantinople
II. The decline of the crusading movement
The later Crusades and the decline of the Latin enclaves
The Latin East after the Third Crusade
The Fifth Crusade
The Sixth Crusade
The Crusade of Louis IX of France
The final loss of the crusader states
The later Crusades and the Kingdom of Cyprus
The results of the Crusades

I. The crusading movement and the first four Crusades

THE FIRST CRUSADE AND THE ESTABLISHMENT OF THE LATIN STATES

Background of the First Crusade. Politically, western Europe in the last quarter of the 11th century comprised several kingdoms loosely describable as feudal. While certain monarchies were already developing better integrated systems of government, many problems associated with feudalism, such as vassalage and inheritance, endemic private warfare, and brigandage, were still present. There were, of course, many variations, and in certain cases special conditions in a country precluded its significant participation in a crusade. Such, for example, were the problems of England's readjustment after the Norman Conquest in 1066, Spain's preoccupation with Muslim incursions from Africa, and civil strife in Germany consequent on long controversy with the popes. In fact, the principal support for the First Crusade came from France; and it was in France that the most typical aspects of feudalism, both negative and positive, could be found.

The economic revival of the 11th century

Europe in the 11th century was feeling the impact of a population growth that had begun toward the end of the 10th century and would continue well into the 13th century. Younger sons of the nobility, unless they married advantageously or entered the religious life, were more than ever disposed to adventure, organized or otherwise. At the same time, an economic revival was in full swing well before the First Crusade. Forest lands were being cleared, frontiers pushed forward, and markets organized, and the Muslim predominance in the Mediterranean was being challenged by Italian shipping. Nobles, bourgeois, and peasants were all seeking new outlets.

The 11th century also witnessed a far-reaching religious renewal. Especially significant for the Crusade was a general overhauling of the ecclesiastical structure, which enabled the popes to assume a more active role in society. In 1095, for example, Urban II, though still meeting resistance from the German Emperor, who opposed papal reform policies, was in a strong enough position to convoke two important ecclesiastical councils.

Thus it was that in the closing years of the 11th century western Europe was abounding in energy and full of confidence. What is more, as is evident in such achievements as the Norman Conquest of England, Europeans possessed the capacity to launch a major military undertaking. Why this energy was channelled into a holy war against Muslims in a distant land is uncertain, though the answer in part involves a new challenge from Eastern Islām presented by the incursions of the Seljuq Turks. The Seljuqs, originally one of several Turkish tribes on the northeastern borders of the Muslim world, had embraced Islām and by the 11th century were moving south and west into Iran and beyond. In large part, however, the answer also lies in the nature of popular religious life and feeling in the West. The two developments are not unconnected.

Factors in the growth of religious enthusiasm

The impact of religion on the lay person of the time is not easy to determine: his religion was unsophisticated, and he was moved by tales of signs and wonders and attributed natural disasters to supernatural intervention. At the same time, lay people were not indifferent to movements of reform, and there were instances of townsmen agitating against clergy whom they regarded as unworthy. There also developed, especially in France, under the leadership of certain bishops but with considerable popular support, what could be described as a peace movement. Many local areas proclaimed the Truce of God and the Peace of God, designed to halt or at least limit fratricidal strife and protect the lives of clergy, travellers, and other persons unable to protect themselves against brigandage. It is particularly interesting to note that the Council of Clermont at which Urban II preached the First Crusade in 1095 renewed and generalized the Truce of God.

While it may seem paradoxical that a council both promulgated peace and officially sanctioned war, it must be remembered that the peace movement was designed to protect those in distress, and a strong element in the Crusade was the idea of giving aid to fellow Christians in the East. Furthermore, there had developed in western Europe, partly under the influence of the Reconquista (the movement to reconquer Spain from the Muslims), the idea that war to defend Christian society was not only justifiable but a holy work, and therefore pleasing to God.

Closely associated with this Western concept of holy war was another popular religious practice, pilgrimage to a holy shrine. Eleventh-century Europe abounded in local shrines housing relics of saints, but three great centres of pilgrimage stood out above all the others: Rome, with the tombs of SS. Peter and Paul; Santiago de Compostela, in northwestern Spain; and Jerusalem, with the Holy Sepulchre of Christ's entombment. Moreover, pilgrimage, which had always been considered an act of devotion, had come to be regarded as a more formal expiation for serious sin, even occasionally prescribed as a penance to the sinner by his confessor.

The importance of Jerusalem

Yet another element in the popular religious consciousness of the 11th century, one associated with both Crusade and pilgrimage, was a widespread belief that the end of the world was imminent. This belief was once thought to have been associated with the year 1000 (the millennium). Scholars now, however, tend to discount the idea that contemporaries had so precise a date in mind and emphasize the continuance of the idea through the 11th century and even beyond. Moreover, in certain late-11th-century portrayals of the end of all things, the "last emperor," now popularly identified with the "king of the Franks," the final successor of Charlemagne, was to lead all the faithful to Jerusalem and there await the Second Coming of Christ. Manifestly, Jerusalem, as the earthly symbol of the heavenly city, figured prominently in Western consciousness. And it was becoming clear, as the number of pilgrimages to Jerusalem increased in the 11th century, that any interruption would have serious repercussions.

By the middle of the 11th century the Seljuqs had wrested political authority from the ʿAbbāsid caliphs of Baghdad. Their policy, originally directed southward against the Fāṭimids of Egypt, was increasingly diverted by the pressure of Turkman raids into Anatolia and Byzantine Armenia. A badly organized Byzantine army was

defeated at Manzikert in 1071, opening Asia Minor to eventual Turkish occupation. Meanwhile, many Armenians south of the Caucasus migrated south to join others in the region of the Taurus Mountains and to form a colony in Cilicia.

Seljuq expansion southward continued, and in 1085 the capture of Antioch in Syria was another blow to Byzantine prestige. Thus, although the Seljuq Empire was never successfully held together as a unit, the Byzantines, driven from all but the coastal regions of Asia Minor, and with Nicaea in Turkish hands by 1092, faced a resurgent Islām perilously close to the capital. It was this danger that prompted the Byzantine emperor Alexius Comnenus to seek aid from the West, and by 1095 the West was ready to respond.

The turmoil of these years disrupted normal political life and made the pilgrimage to Jerusalem difficult and often impossible. Stories of dangers and molestation reached the West and remained in the popular mind even after conditions improved. Further, informed authorities began to realize that the revived power of the Muslim world now seriously menaced the West as well as East. It was this realization that stimulated official and organized action.

Alexius' appeal came at a time when relations between the Eastern and Western branches of the Christian world were improving. Difficulties between the two in the middle years of the century had resulted in a de facto, though not formally proclaimed, schism, and ecclesiastical disagreements had been accentuated by Norman occupation of formerly Byzantine areas in southern Italy. A campaign of the Norman adventurer Robert Guiscard against the Greek mainland further embittered the Byzantines, and it was only after Robert's death in 1085 that conditions for a renewal of normal relations between East and West were reasonably favourable. Envoys of Emperor Alexius Comnenus thus arrived at the Council of Piacenza in 1095 at a propitious moment; and it seems probable that Pope Urban II viewed military aid as a means toward restoring ecclesiastical unity.

The Council of Clermont

The Council of Clermont convoked by Urban on November 18, 1095, was attended largely by bishops of southern France. There was also a scattering of representatives from the north and elsewhere. After the transaction of many important items of ecclesiastical business, which resulted in a series of canons, among them one renewing the Truce of God, there was included a canon that granted a plenary indulgence (the remission of all penance for sin) to those who undertook to aid Christians in the East. Then in a great outdoor assembly the pope, himself a Frenchman, addressed a large crowd. We are not certain what language Urban spoke to these people, but apparently it was a language they could understand.

Precisely what the Pope said can never be known, since the only surviving accounts of his speech were written years later. But he apparently stressed the plight of Eastern Christians, the molestation of pilgrims, and the desecration of the holy places. He urged those of his hearers who were guilty of disturbing the peace to turn their warlike energies toward a holy cause. He emphasized the need for penance, for the acceptance of suffering. No one should undertake this pilgrimage for any but the most exalted of motives.

The response was immediate and overwhelming, probably far greater than Urban had anticipated. Cries of "Deus volt" ("God wills it") were heard everywhere, and it was decided that those who agreed to go were to wear a cross. Moreover, it was not only warrior knights who responded; a popular element, apparently unexpected and probably not desired, also came forward.

The era of Clermont witnessed the concurrence of three significant developments: first, there existed as never before a popular religious fervour that was not without marked eschatological tendencies in which the Holy City of Jerusalem figured prominently; second, war against the infidel had come to be regarded as a religious undertaking, a work pleasing to God; and finally, from an organizational standpoint, ecclesiastical as well as secular, western Europe possessed the capacity to plan such an enterprise and carry it through.

Preparations for the Crusade. Following Pope Urban's speech, preparations began in both East and West. Emperor Alexius, who had doubtless anticipated some sort of auxiliary force, apparently soon realized that he would have to provide for and police a much larger influx of warriors. In the West, as the leaders began to assemble their armies, persons who took the cross sought to raise money, often by selling or mortgaging property, both for the immediate purchase of equipment and for the long-term needs of the journey.

As this was being done, several less organized and generally miscellaneous bands, including some armed men, commonly known as the "People's Crusade," set out across Europe. The most famous of these, brought together by a remarkable popular preacher, Peter the Hermit, and his associate Walter the Penniless, arrived at Constantinople after causing considerable disorder in Hungary and Bulgaria. Alexius received Peter cordially and advised him to await the arrival of the main Crusade force. But the rank and file grew unruly, and on August 6, 1096, they were ferried across the straits. As Peter returned to beg for aid, they were ambushed at Cibotus (called Civetot by the Franks) and all but annihilated by the Turks. Other, even less well-disciplined bands, largely from northwest Germany, where they committed atrocities against the Jewish communities, were dispersed before reaching the Byzantine frontier.

The People's Crusade

The main Crusading force, which began to move as Urban had directed in August 1096, consisted of four major contingents. Hugh of Vermandois, brother of King Philip I of France, left first with a small following that was reduced by shipwreck while crossing the Adriatic from Bari to Dyrrhachium. Godfrey of Bouillon, duke of Lower Lorraine since 1089, the only major prince from the German kingdom, though he and his associates largely spoke French, was joined by his brothers, Eustace and Baldwin, and a kinsman, Baldwin of Le Bourg. Taking the land route, Godfrey crossed Hungary without incident. Markets and provisions were supplied in Byzantine territory, and, except for some pillaging, the army reached Constantinople without serious incident in late December.

Second, Bohemond, a Norman from southern Italy and the son of Robert Guiscard, was thus on familar ground when he crossed the Adriatic, where he had fought with his father and was understandably feared by the Byzantines. But he was now 40 years old and determined to come to profitable terms with his former enemy. He arrived at Constantinople on April 9, 1097.

Third, the largest army was that assembled by Raymond of Saint-Gilles, the count of Toulouse. Fifty-five years of age, he was the oldest and most prominent of the Crusading princes, and he aspired and perhaps expected to become the leader of the entire expedition. He was accompanied by Adhémar, bishop of Le Puy, whom the Pope had named as legate for the Crusade. Raymond led his followers, including a number of noncombatant pilgrims whom he supported at his own expense, across north Italy, around the head of the Adriatic, thence southward into Byzantine territory. This large body caused considerable trouble in Dalmatia and with Byzantine troops policing the area nearer the capital, where Raymond arrived on April 27.

Meanwhile, the fourth army, under Robert of Flanders, had crossed the Adriatic from Brindisi. Accompanying Robert were his cousin Robert of Normandy (brother of King William II Rufus of England) and Stephen of Blois (the son-in-law of William I the Conqueror). No king took part in the First Crusade, and the predominantly French-speaking participants came to be known as Franks.

The presence near Constantinople of massive military forces, numbering perhaps 4,000 mounted knights and 25,000 infantry, posed a serious problem for Alexius, and there was occasional disorder. Forced to consider the permanent imperial interests, which, it soon became evident, were different from the objective of the crusad-

Latin–Byzantine friction

ers, the Emperor required each Crusade leader to promise under oath to restore any conquered territory that had belonged to the empire before the Turkish invasions and to swear allegiance to him for any lands occupied beyond the former frontiers. It was a reasonable demand, and all the leaders took the oath before crossing the straits, though Godfrey and Raymond did so only after some hesitation. Raymond, in fact, agreed only to a modified form of oath common in southern France, and in the end he remained of all the Crusade leaders the most loyal to Alexius.

The siege of Antioch. Late in May 1097, crusaders, accompanied by a Byzantine contingent, appeared before Nicaea. On June 19, when the city fell, it was surrendered, as had been agreed, to the Byzantines. Crossing arid and mountainous Anatolia proved a difficult task. At Dorylaeum, on July 1, 1097, Turks attacked the advance column of the crusader army. Despite the heat and a rain of arrows that fell on them, the crusaders held their ground, and, when the rest of the army drew up, the Turks were routed. A major victory in open warfare had been achieved by cooperation among the separate contingents and the Greeks.

Further advance across Anatolia was even more arduous, and it was only after suffering many casualties, especially in the region of the Taurus Mountains, that the crusaders arrived near Antioch on October 20. Meanwhile, Godfrey's brother Baldwin left the main army to become involved in Armenian politics and he became ruler of Edessa.

One of the great cities of the Levant, Antioch was surrounded by a great circle of walls, and despite reinforcements and supplies by Genoese and English ships and later by the Patriarch of Jerusalem, then in Cyprus, the siege proved long and difficult. Spring brought the threat of attack by a relief force under Karbūgah of Mosul, and some of the crusaders, among them Stephen of Blois, became discouraged and left for home. In Asia Minor, Stephen met Alexius, who had set out as he had agreed to do, and convinced him that Antioch's cause was hopeless. The Emperor's decision to turn back, however justified tactically, was a diplomatic blunder; when the crusaders learned of the Emperor's move, they felt free from any obligation to return the city to him.

Bohemond meanwhile had proposed that whoever first entered the city should have possession, provided the Emperor did not make an appearance. The Norman had, in fact, already made contact with a discontented commander within, who proceeded to admit him over a section of the walls. The other crusaders followed into the city on June 3, 1098, and massacred the Muslim inhabitants. Only the citadel held out.

Thus, Antioch was occupied, and it is significant that the Greek Patriarch who had been imprisoned was restored, suggesting that neither Adhémar nor presumably Urban II had any plan to substitute a Latin incumbent. The victory, however, was incomplete, and Karbūgah soon blockaded the city. Disagreements among the leaders persisted and were accentuated by unseemly arguments over the validity of what had come to be called the Holy Lance, which a Provençal priest found below the cathedral and insisted was the lance that, according to the Gospels, had pierced the side of Christ when he hung on the cross. Nonetheless, on June 28 the crusader army moved out of the city and after a day of fierce fighting defeated Karbūgah's forces. The citadel then surrendered to Bohemond, and its garrison was permitted to leave.

The death of Adhémar of Le Puy

Rejoicing was tempered by a devastating epidemic that took many lives, including that of the legate, Adhémar of Le Puy, who, as the spiritual leader of the Crusade, had been a wise counsellor and a stabilizing influence whom the leaders could ill afford to lose.

As the leaders continued to confer on the final disposition of Antioch, the rank and file of the army grew impatient. In January, following the siege of Ma'arrat an-Nu'mān, some miles to the south, there was a popular demonstration in favour of moving on. The army then set out for Jerusalem under the leadership of Raymond of Saint-Gilles. As they moved south, Tancred and Robert of Normandy and, later, Godfrey and Robert of Flanders joined them. Bohemond, fearing for the safety of Antioch, turned back soon after leaving.

The siege of Jerusalem. Not far from Beirut, the army entered the territory of the Fāṭimid caliphs of Cairo, who, as Shī'ite Muslims, were enemies of the Sunnite Seljuqs and the caliphs of Baghdad. In August 1098 the Fāṭimids had occupied Jerusalem. The final drive of the First Crusade, therefore, was against the Fāṭimids of Egypt, not the Seljuqs.

On June 7, 1099, the Christian army, by then considerably reduced to perhaps 1,200–1,500 cavalry and 12,000 foot soldiers, encamped before Jerusalem, the governor of which was well supplied and confident that he could withstand a siege until a relief force arrived from Egypt. The crusaders, on the other hand, were short of supplies and would be until six vessels arrived at Jaffa (Yafo) and managed to unload before the port was blockaded by an Egyptian squadron. On July 8 a strict fast was ordered, and, with the Muslims scoffing from the walls, the entire army, preceded by the clergy, marched in solemn procession around the city, thence to the Mount of Olives, where Peter the Hermit preached with his former eloquence.

Siege towers were brought up to the walls on July 13–14, and on July 15 Godfrey's men took a sector of the walls, and others followed on scaling ladders. When the nearest gate was opened, Tancred and Raymond entered, and the Muslim governor surrendered to the latter in the Tower of David. The governor along with his bodyguard was escorted out of the city. Tancred promised protection in the Aqṣā Mosque, but his orders were disobeyed. All Muslims, men, women, and children, as well as Jews, perished in the general slaughter that followed.

Massacre of Jerusalem's inhabitants

Thus it was that, three long years after they had set out, the crusaders attained their goal. Doubtless some participants had been moved more by hopes of material gain or sheer love of adventure than by religious devotion. Nonetheless, it is difficult to believe that the Crusade could have succeeded without an extraordinary dedication to the ultimate objective, the liberation of the Holy City of Jerusalem.

The establishment of the crusader states. After a successful surprise attack on the Egyptian relief army, which confirmed the occupation of Palestine, most of the crusaders, having fulfilled their vows of pilgrimage, departed for their homes and left the problem of governing the conquered territories in the hands of the few who remained. Initially, there was disagreement as to the nature of the government to be established, and some held that the Holy City should be ruled under ecclesiastical authority. As an interim measure, Godfrey was elected to govern, and took the modest title of defender of the Holy Sepulchre. Nevertheless, Raymond, who had previously declined the offer of the crown, chose to leave immediately.

In December 1099, in the midst of this somewhat confused situation, Bohemond and Baldwin of Edessa arrived in Jerusalem to fulfill their crusader vows. Accompanying Bohemond was Daimbert, the archbishop of Pisa, who was chosen patriarch and proceeded to receive the homage of both Godfrey and Baldwin. If Daimbert had ambitions to govern Jerusalem, they were thwarted when, on Godfrey's death, his brother Baldwin was summoned back to Jerusalem, where he assumed the title of king (November 11, 1100). Thus, there had come into being not a church state but a feudal Kingdom of Jerusalem.

Though the Crusade of 1101, which Pope Paschal II had organized to bring reinforcements, collapsed in Asia Minor, King Baldwin profited from the chronic rivalries among his Muslim neighbours. He was further aided in extending the coastline by Italians and in one instance by a Norwegian squadron that arrived under King Sigurd in 1110. By 1112 Arsuf, Caesarea, Acre, Beirut, and Sidon had been taken, and the entire coast except for Ascalon and Tyre was in Latin hands. Meanwhile, castles had been built in Galilee and the frontier pushed southward.

These same years witnessed the formation of three oth-

The County of Edessa

er crusader states to the north. The County of Edessa, an ill-defined domain extending into the upper Euphrates region with a population consisting mainly of Armenians and Syrians, had been established by Godfrey's brother Baldwin. When Baldwin left to become ruler of Jerusalem he bestowed the county, under his suzerainty, on his cousin Baldwin of Le Bourg.

Antioch had not been returned to the Emperor, and Bohemond had consolidated his position there. The city was predominantly Greek in population, though there were also Syrians and Armenians, and the latent Greek–Latin friction was intensified when Bohemond replaced the Greek Patriarch by a Latin. When Bohemond was captured by the Muslims in 1100, his nephew Tancred became regent and expanded the frontiers of the principality to include the important port of Latakia, taken from the Byzantines in 1103. Not long after his release in 1103, Bohemond travelled to Europe, where he succeeded in winning over Pope Paschal II to the idea of a new Crusade. Whatever the original intention, there resulted not an expedition against Muslims but an attack on the Byzantine city of Dyrrhachium. Like its predecessor, the ill-fated campaign of 1081, the enterprise failed, and in 1108 Bohemond was forced to take an oath of vassalage to the Emperor for Antioch and to return to Italy, where he died in 1111. Though Tancred, again in power, disregarded his uncle's oath, Antioch and its patriarchate remained a source of controversy.

A fourth crusader state was established on the coast in the vicinity of Tripoli (Arabic: Tarābulus) by Raymond of Saint-Gilles, who had been outmanoeuvred in Jerusalem and had returned to Constantinople hoping for aid from the Byzantine Emperor to whom he had always been loyal. But the Crusade of 1101, which he led, had been another blow to his prestige. In 1102 he returned to Syria, took Tortosa (Ṭarṭūs), and began the siege of Tripoli. But he died in 1105, and it remained for his descendants to finish the task. Tripoli fell in 1109.

The establishment and protection of the frontiers was, for the new states, a problem conditioned by two factors, geography and the politics of Levantine Islām. From Antioch south, the crusaders had occupied a narrow strip of coastland bounded by mountains to the north and by the Jordan Valley in the south. To the east beyond the Syrian desert lay the Muslim cities of Aleppo, Ḥamāh, Homs, and Damascus. Though the Franks did push southward to Aylah (or Elim, modern al-ʿAqabah), all attempts to move eastward failed, and it was necessary to erect castles at vulnerable points along the eastern frontier as well as along the coast and inland. Among the most famous of these were Krak de Montréal, in the Transjordan, and Krak des Chevaliers, in the County of Tripoli. Meanwhile, the hostility between Shīʿite Egypt and Sunnite Baghdad continued for some time, while the emirates in between remained divided in their allegiance and those in the north fearful of the Seljuqs of Iconium.

After Baldwin I's death in 1118, the throne passed to his cousin Baldwin of Le Bourg, who left Edessa to another cousin, Joscelin of Courtenay. In 1124, Tyre, the last great city north of Ascalon still in Muslim hands, was taken with the aid of the Venetians, who, as was customary, demanded a section of the city. Baldwin II was succeeded by Fulk of Anjou, a newcomer recommended by Louis VI of France, who married Baldwin's daughter Melisend. In 1131 Baldwin and Joscelin both died; they were the last of the first generation of crusaders, and with their passing the formative period in the history of the crusader states came to an end.

End of the period of expansion

Fulk's policies represented an end of expansion and a stabilization of gains made, a wise course, because his reign coincided with the rise of Zangī, *atabeg* (governor) of Mosul, whose achievements earned him the reputation of a great champion of the *jihād* (holy war) against the Franks. When Zangī moved against Damascus, the recognition of their common danger resulted in a Jerusalem–Damascus alliance, a kind of diplomacy by then common with the second-generation Franks.

The north, however, was in great danger. The Byzantines had recovered their influence in Anatolia and were putting pressure on Armenia and Antioch. Emperor Manuel Comnenus forced Prince Raymond of Antioch to acknowledge the imperial suzerainty. But the greater danger to both was dramatically brought home by Zangī's capture of Edessa in 1144. Attempts at recovery failed, and the northernmost crusader state was subsequently overrun.

THE ERA OF THE SECOND AND THIRD CRUSADES

The Second Crusade. It had long been apparent that Edessa was vulnerable, but its loss came as a shock to Christians both East and West. Urgent pleas for aid soon reached Europe, and Pope Eugenius III in 1145 issued a formal crusade bull, the first of its kind, with precisely worded provisions designed to protect crusaders' families and property and reflecting contemporary advances in canon law. Legates were designated. The Crusade, energetically supported by Louis VII, was preached by St. Bernard of Clairvaux in France and, with interpreters, even in Germany. As in the First Crusade, many simple noncombatant pilgrims responded. But, since Emperor Conrad III, though at first reluctant to leave Germany, had been won over by Bernard's eloquence, the Second Crusade, unlike the First, was led by two of Europe's major rulers.

Conditions in the East in mid-12th century

The situation in the East was also different. Manuel Comnenus was favourably disposed toward the West and, because of his interest in Antioch, was concerned over the power of Zangī, the Muslim ruler of Aleppo. He was able to assist the crusaders with guides and supplies but contributed no troops. In anticipation of the arrival of the western armies and more aware than they of the delicate power balance in the Levant, he had made peace with the Seljuqs of Iconium in 1146. Above all, Manuel was alarmed by the possibility of an attack by Roger II, the Norman king of Sicily.

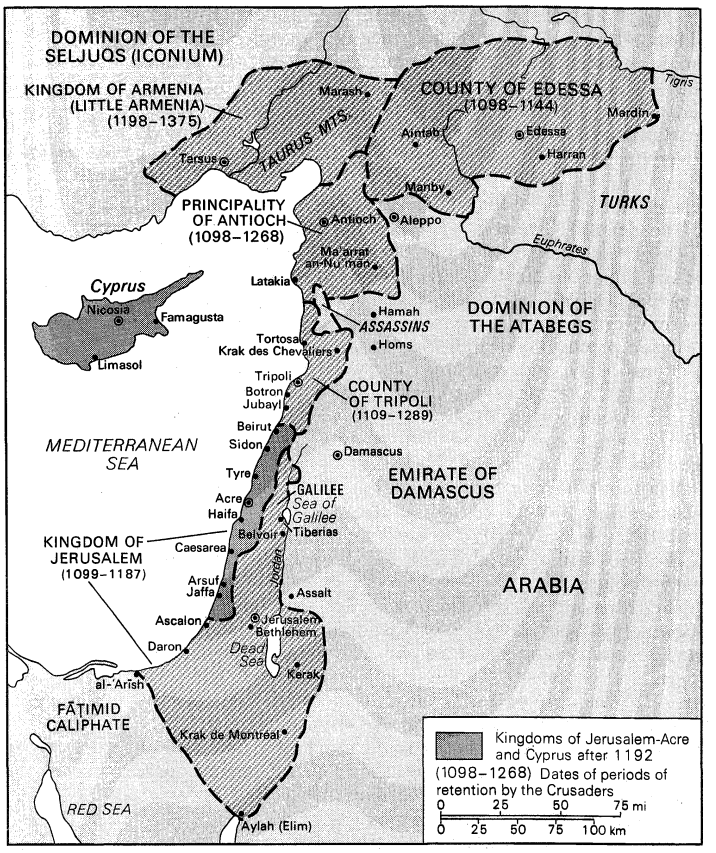
Conrad left in May 1147, accompanied by many German nobles, the Kings of Poland and Bohemia, and Frederick of Swabia, his nephew and heir. On leaving Hungary and entering Byzantine territory, he agreed to an oath of noninjury. Conrad's poorly disciplined troops created tension with the Byzantines in Constantinople, where they arrived in September. Both Conrad and Manuel, however, remained on good terms, and both were apprehensive about the moves of Roger of Sicily, who during these same weeks seized Corfu and attacked the Greek mainland.

Conrad, rejecting Manuel's advice to follow the coastal route around Asia Minor, moved his main force past Nicaea directly into Anatolia. On October 25, at Dorylaeum, not far from the place where the first crusaders won their victory, his army, weary and without adequate provisions, was set upon by the Turks and virtually destroyed. Conrad, with a few survivors, retreated to Nicaea.

Meanwhile, about a month behind the Germans, Louis VII, accompanied by his wife Eleanor of Aquitaine, followed the land route across Europe and arrived at Constantinople on October 4. A few of his more hotheaded followers, on hearing that Manuel had made a truce with the Turks of Iconium and totally misunderstanding the motives, accused the Emperor of treason and urged the French King to join Roger in attacking the Byzantine emperor. Louis preferred the opinion of his less volatile advisers and agreed to restore any imperial possessions he might capture.

In November the French reached Nicaea, where they learned of Conrad's defeat. Louis and Conrad then started along the coastal route, the French now in the van, and reached Ephesus. Conrad became seriously ill and returned to Constantinople to the ministrations of Manuel, who prided himself on his medical knowledge. He eventually reached Acre by ship in April 1148.

The French passage from Ephesus to Antioch in midwinter was harrowing in the extreme. Supplies ran short, and the Byzantines were blamed. But Antioch was finally reached in March 1148, and the crusaders were welcomed by Prince Raymond, Queen Eleanor's uncle. Raymond urged an attack on Aleppo, the centre of power of Nu-



The crusader states of the 12th century.
From W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

reddin, son and successor of Zangī. But King Louis, who resented Eleanor's open espousal of Raymond's project, left abruptly for Jerusalem and forced the Queen to join him.

At Jerusalem, where Conrad had already arrived, a brilliant assembly of French and German notables assembled with Queen Melisend, her son Baldwin III, and the barons of Jerusalem to discuss how best to proceed. Despite the absence of the northern princes and the losses already suffered by the crusaders, it was possible to field an army of nearly 50,000 men, the largest Crusade army so far assembled. After considerable debate, which revealed the conflicting purposes of crusaders and Jerusalem barons, it was decided to attack Damascus.

Attack on Damascus

How the decision was reached is not known. Damascus was undoubtedly a tempting prize, but Unur, the Turkish commander, also fearful of the expanding power of Nureddin and the one Muslim ruler most disposed to cooperate with the Franks, was now forced to seek the aid of his former enemy. And Nureddin was not slow to move toward Damascus. Not only was the campaign mistakenly conceived, it was badly executed. On July 28, after a five-day siege, with Nureddin's forces nearing the city, it became evident that the crusader army was dangerously exposed, and a retreat was ordered. It was a humiliating failure, attributable largely to the conflicting interests of the participants.

Conrad left immediately and stopped at Constantinople, where he agreed to join the Emperor against Roger of Sicily. Louis's reaction was different. His resentment against Manuel, whom he blamed for the failure, was so great that he accepted Roger's offer of ships to take him home and agreed to a plan for a new Crusade against Byzantium, a plan conceived in frustration and fortunately soon abandoned.

The Second Crusade had been promoted with great zeal and had aroused high hopes. Its collapse caused deep dismay, and subsequent expeditions, perhaps wisely, were more limited in their objectives. The Muslims, on the other hand, were enormously encouraged. They had confronted the danger of another major Western expedition and had triumphed over it.

The crusader states to 1187. During the 25 years following the Second Crusade the Kingdom of Jerusalem was governed by two of its ablest rulers, Baldwin III and Amalric I. In 1153 King Baldwin captured Ascalon—the last major conquest of the Franks. The victory extended their coastline southward. But its possession was offset in the next year by the occupation of Damascus by Nureddin, one more stage in the encirclement of the crusader states by a single Muslim power.

In 1160–61 the possibility that the Fāṭimid caliphate in Egypt, shaken by palace intrigues and assassinations, might collapse and fall under the influence of Muslim Syria caused anxiety in Jerusalem. Thus, in 1164, when Nureddin sent his lieutenant Shīrkūh to Egypt accompanied by his own nephew, Saladin, King Amalric decided to intervene. After some manoeuvring, both armies withdrew, as they were to do again three years later.

Meanwhile, Amalric, realizing the necessity of Byzantine cooperation, had sent Archbishop William of Tyre as envoy to Constantinople. But in 1168, before the news of the agreement that William of Tyre had arranged reached Jerusalem, the King set out for Egypt. The reasons are not clear, and there had been considerable division among the barons. At any rate, the venture failed, and Shīrkūh entered Cairo. On his death (May 23, 1169), Saladin, then Nureddin's deputy, was left to overcome the remaining opposition and master Egypt.

The accession of Saladin

When the Byzantine fleet and the army finally arrived in 1169, there was some delay, and both armies were forced by inadequate provisions and seasonal rains to retreat once again, each side blaming the other. In 1171, Saladin obeyed Nureddin's order to have the prayers in the mosques mention the Caliph of Baghdad instead of the Caliph of Cairo, who was then in his final illness. Thus ended the Fāṭimid caliphate and the great division in Levantine Islām from which the Latins had profited.

Ominous developments followed the deaths of both Amalric and Nureddin in 1174. In 1176 the Seljuqs of Iconium defeated the armies of Manuel Comnenus at Myriocephalon. It was a shattering blow reminiscent of Manzikert a century earlier. When Manuel died in 1180, all hope of effective Byzantine–Latin cooperation vanished. Three years later Saladin occupied Aleppo, virtually completing the encirclement of the Latin states. In 1185 he agreed to a truce and left for Egypt.

In Jerusalem, Amalric was succeeded by his son Baldwin IV, a boy of 13 suffering from leprosy. Despite the young king's extraordinary fortitude, his precarious health necessitated occasional regencies and created a problem of succession until his sister Sibyl bore a son, the future Baldwin V, to William of Montferrat. Her marriage in 1180 to Guy of Lusignan, a newcomer to the East and brother of Amalric, the constable, accentuated existing rivalries among the barons. A kind of "court party" centring around the queen mother, Agnes of Courtenay, her daughter Sibyl, and Agnes' brother, Joscelin III, and now including the Lusignans was often opposed by another group composed mostly of the so-called native barons—old families, notably the Ibelins, Reginald of Sidon, and Raymond III of Tripoli, who through his wife was also lord of Tiberias. In addition to these internal problems, the kingdom was more isolated than ever. Urgent appeals to the West and the efforts of Pope Alexander III had brought little response.

Baldwin IV died in March 1185, leaving, according to previous agreement, Raymond of Tripoli as regent for the child king Baldwin V. But when Baldwin V died in 1186 the court party outmanoeuvred the other barons and, disregarding succession arrangements that had been formally drawn up, hastily crowned Sibyl. She in turn crowned her husband, Guy of Lusignan.

In the midst of near civil war, Reginald of Châtillon, lord of Kerak and Montréal, broke the truce by attacking a caravan. Saladin replied by proclaiming *jihād* against the Latin kingdom. In 1187 Saladin left Egypt, crossed the Jordan south of the Sea of Galilee, and took up a position near the riverbank. Near Saffuriyah (mod-

ern Zippori) the crusaders mobilized an army of perhaps 20,000 men and including some 1,200 heavily armed cavalry, probably the equal of Saladin's. In a spot well chosen and adequately supplied with water and provisions, they awaited Saladin's first move.

On July 2 Saladin blocked the main road to Tiberias and sent a small force to attack the town, hoping, since Count Raymond's wife was there, to lure the crusaders into the open. Nevertheless, it was Raymond who at first persuaded the King not to fall into the trap. But, late that night, others, accusing the Count of treason, prevailed upon the King to change his mind. It was a fateful decision. For, after an exhausting day's march on July 3, a terrible night spent without water, surrounded and constantly harassed, and a long day of fighting near the Horns of Ḥaṭṭin on July 4 with smoke from grass fires set by the enemy pouring into their faces, the foot soldiers broke and fled, destroying the essential coordination with the cavalry. When Saladin's final charge ended the battle, most of the knights had been slain or captured. Only Raymond of Tripoli, Reginald of Sidon, Balian of Ibelin, and a few others escaped.

Defeat of the crusaders

The King's life was spared, but Saladin killed Reginald of Châtillon and ordered the execution of some 200 Templars and Hospitallers. Other captive knights were treated honourably, and most were later ransomed. Less fortunate were the foot soldiers, most of whom were sold into slavery. Virtually the entire military force of the Kingdom of Jerusalem had been destroyed.

Saladin quickly followed up his victory at Ḥaṭṭin. Pausing only to take Tiberias, he moved toward the coast and seized Acre. By September 1187 he and his lieutenants had occupied most of the major strongholds in the kingdom and all of the ports south of Tripoli, except Tyre, and Jubayl and Botron (al-Batrūn) in the county of Tripoli. On October 2, Jerusalem, then defended by only a handful of men under the command of Balian of Ibelin, capitulated to Saladin, who agreed to allow the inhabitants to leave after paying a ransom. Though Saladin offered to release the poor for a specified sum, several thousand apparently were not redeemed and probably were sold into slavery. In Jerusalem, as in most of the cities captured, those who elected to remain were Syrian or Greek Christians. Somewhat later Saladin permitted a number of Jews to settle in the city.

Meanwhile, Saladin continued his conquests in the north, and by 1189 all the kingdom was in his hands except Belvoir (modern Kokhov ha-Yarden) and Tyre. The County of Tripoli and the Principality of Antioch were each reduced to the capital city and a few outposts. The larger part of the 100-year-old Latin establishment in the Levant had been lost.

The institutions of the First Kingdom. The four principalities established by the crusaders—three after the loss of Edessa in 1144—were loosely connected, and such limited suzerainty as the king of Jerusalem exercised over the others became largely nominal after the midcentury. Each of the states was organized into a pattern of feudal lordships by the ruling minority. The institutions of the Kingdom of Jerusalem are best known, partly because its history figures more prominently in both Arab and Christian chronicles but especially because its documents were better preserved. In the 13th century there was compiled a collection of laws, the famous *Assises de Jérusalem*. Though this collection reflects a later situation, certain sections and many individual enactments can be traced back to the 12th century, the period known as the First Kingdom.

In the first half of the 12th century the kingdom presented the appearance of a typical feudal monarchy on the European model, with lordships owing military service and subject to fiscal exactions. There were, however, important differences, not only in the large subject population of diverse ethnic origins but also with respect to the governing minority. No great families with extensive domains emerged in the early years. Further, the typical noble's residence was not, as in Europe, the rural castle or manor house. Castles there were, but they were garrisoned by knights and, increasingly as the century advanced, by the religio-military orders. But most Jerusalem barons lived in the fortified towns. The kings, moreover, possessed a considerable domain and retained extensive judicial rights. As a consequence, the monarchy was a relatively strong institution in early Jerusalem.

Toward the middle of the century this situation began to change. Partly as a consequence of increased immigration from the west, the baronial class grew, and there was formed a relatively small group of magnates with large domains. As individuals, they were less disposed to brook royal interference, and as a class and in the court of barons (Haute Cour, or High Court) they were capable of presenting a formidable challenge to royal authority. The last of the kings of Jerusalem to exercise effective power was Amalric I in the 12th century. In the final years of the First Kingdom, baronial influence was increasingly evident and dissension among the barons, as a consequence, more serious.

Another serious obstacle to the king's jurisdiction and one not found, at least in the same form, in the West was the extensive authority of the two religio-military orders. The older of the two, the Knights of St. John of Jerusalem, or Hospitallers, was founded in the 11th century by the merchants of Amalfi to provide hospital care for pilgrims. The order never abandoned its original purpose, and, in fact, as its superb collection of documents reveals, the order's philanthropic activities expanded. But during the 12th century, in response to the military needs of the kingdom, the Hospitallers followed the example of the Templars and became, in addition, an order of knights.

The military orders

The Poor Knights of Christ and of the Temple of Solomon, so called because of their headquarters in the former temple of Solomon, originated as a monastic–military organization to protect pilgrims on the way to Jerusalem. Its rule, composed by St. Bernard of Clairvaux, was officially sanctioned by the Council of Troyes (1128). Although Templars and Hospitallers took monastic vows, their principal function was soldiering.

The orders grew rapidly, and, as they acquired castles at strategic points in the kingdom and in the northern states and maintained permanent garrisons, they supplemented the otherwise not entirely adequate forces of the barons and king. Moreover, since they were soon established in Europe as well, they became international in character. Virtually independent, sanctioned and constantly supported by the papacy, and exempt from local ecclesiastical jurisdiction, they aroused the jealousy of the clergy and constituted a serious challenge to royal authority.

The crusaders introduced into the conquered lands a Latin ecclesiastical organization and hierarchy. The Greek patriarch of Antioch was, as has been mentioned, removed, and all the subsequent incumbents were Latin, except for one brief period before 1170, when imperial pressure succeeded in installing a Greek. In Jerusalem the Orthodox patriarch had left before the conquest and died soon after. All his successors were Latin.

Under the Latin jurisdiction was the entire Latin population and those natives (Greek in Antioch and Greek or native Syrian [Melkite] in Jerusalem) who had remained Orthodox. Outside was a larger number of Monophysites (Jacobite or Armenian) and some few Nestorians, all adherents of doctrines that had deviated from the decisions of 5th-century ecumenical councils. A number of Maronites of the Lebanon region accepted the Latin obedience late in the 12th century. After some initial confusion, the native hierarchies were able to resume their functions.

As in the West, the church had its own courts and possessed large properties. But each ecclesiastical domain was required to furnish soldiers, and the charitable foundations were considerable. The hierarchy of the Latin states was an integral part of the church of the West. Papal legates regularly visited the East, and bishops from the crusader states attended the Third Lateran Council in 1179. Western monastic orders also appeared in the crusader's states.

In addition to the nobles and their families who had

The bourgeois

settled in the kingdom, a substantially larger number of persons was classified as bourgeois. A small number had arrived with the First Crusade, but most were later immigrants from Europe, representing nearly every nationality, but predominantly from rural southern France. In the East they became town dwellers, but a few were agriculturalists, proprietors of small estates, rarely themselves tillers of the soil, inhabiting the more modest towns. Some immigrants, it appears, perhaps poor pilgrims who remained, failed to obtain a reasonably settled status and could not afford the relatively small ransom offered by Saladin in 1187.

The townsmen of the First Kingdom did not, like their counterparts in Europe, aspire to political autonomy. There were no communal movements in the 12th century. The bourgeois were, therefore, subject to king or seigneur. Some did military service as sergeants; *i.e.*, mounted auxiliaries or foot soldiers. Bourgeois were recognized as a class in the more than 30 "courts of the bourgeois" according to procedures laid down in the *Assises de la cour des bourgeois*, which, unlike other parts of the *Assises*, reflects the traditions of Roman law in southern France.

The Italians, because they supplied indispensable naval aid and shipping essential to regular contact with Europe, had been able to acquire exceptional privileges in the ports. These privileges usually included a quarter that they maintained as a virtually independent enclave. Closely associated with the mother city, Venice, Genoa, Pisa, etc., its status guaranteed by treaty between the kingdom and the mother city, the Italian enclave more nearly resembled the modern overseas colony than any other establishment of the crusaders.

The European settlers in the crusader states constituted only a small minority of the population. If the early crusaders were ruthless, their successors, except for occasional outbursts during campaigns, acquired not only a remarkable tolerance but also a notable flexibility in dealing with the diverse sectors of the native population. Those Muslim town dwellers who had not fled were captured and put to menial tasks. Some, it is true, appeared in Italian slave marts, but royal and ecclesiastical ordinances at least laid down limits to what slave owners might do. Baptism brought with it immediate freedom.

Not all Muslims were slaves. Most of those who remained were peasants who for centuries had constituted a large part of the rural population and who were permitted to retain their holdings, subject to fiscal impositions not unlike those of the European serf, and usually identical with those originally levied by their former proprietors on all non-Muslims. Muslim nomads, or Bedouins, who from time immemorial had moved with their herds as changing seasons provided pasturage, continued to enjoy pasturage rights within the kingdom guaranteed by the king.

Most mosques were appropriated during the conquest, but some were restored, and no attempt was made to restrict Muslim religious observance. Occasionally a *miḥrāb* (prayer niche) was retained for Muslim worshippers in a church that had formerly been a mosque. The tolerance of the Franks, noted by Arab visitors, often surprised and disturbed newcomers from the West.

Legal practices

The native Christians, predominantly townsmen, were governed according to the *Assises de la cour des bourgeois*. Each national group retained its customary institutions. The Syrians, for example, maintained a *cour du rais*, the *rais* (*raʾīs*) being a community chieftain and often a person of considerable importance under the Frankish regime. An important element in the kingdom's army, the corps of Turcopoles, lightly armed cavalry units, was also composed largely of native Christians, including, apparently, converts from Islām. The principle of personality of law applied to all. The Jew took oath on the Torah, the Samaritan on the Pentateuch, the Muslim on the Qur'ān, and the Christian on the Gospels.

The Jewish community of Palestine, which had already begun to diminish in the 11th century, was drastically reduced by the First Crusade. But gradually, as the Latin kingdom settled into a routine of government, the situation improved. Indeed, there is reason to believe that the later, more stable regime made possible a not inconsiderable Jewish immigration, not, it seems, as in earlier times, from the neighbouring lands of the Middle East, but from Europe.

Thus, by the 1170s the crusader states of Outremer, as the area of Latin settlement came to be called, had developed well-established governments. With allowance made for regional differences (*e.g.*, Antioch in its early years under the Norman dynasty was somewhat more centralized), the institutions of the northern states resembled those of Jerusalem. The governing class of Franks was no longer made up of conquerors from abroad but of local residents who had learned to adjust themselves to a new environment, and were concerned with administration. A few—as, for example, Reginald of Sidon and William of Tyre, archbishop and chancellor—were fluent in Arabic. Many others knew enough of the language to deal with natives. Franks adopted native dress, ate native food, employed native physicians, and married Syrian, Armenian, or convert-Muslim women.

The Franks' cultural identity

But the Franks of Outremer, though they sometimes acquired a love of luxury and comfort, did not lose the will or ability to confront danger; nor did they "go native." In fundamentals, they continued to adhere to the traditions of their French forebears. They were Latin Christians. Documents were drawn up in Latin. The *Assises* were in French. William of Tyre, born in the East but educated in Europe, wrote a celebrated *History of Deeds Done Beyond the Sea* in the Latin style of the 12th century.

Artists and architects were influenced by Byzantine and Arab craftsmen, but Oriental motifs in their work are usually limited to details, such as doorway carvings. A psalter done for Queen Melisend in the 12th century, for example, shows certain Byzantine characteristics, and the artist may have lived in Constantinople, but the manuscript is in the current tradition of French art. Castles followed Byzantine models, often being built on the old foundations, but Western ideas were also incorporated. New churches were built or additions made to existing structures, as for example the Church of the Holy Sepulchre, in the Romanesque style of the homeland.

All in all, the Franks of the First Kingdom were developing a distinctive culture and achieving a real sense of identity. Until baronial dissensions weakened the monarchy in the later years—a human failing, not a deep cultural maladjustment—the Latin kingdom showed remarkable vitality and exceptional ingenuity. It was one of the more sophisticated governmental achievements of the Middle Ages.

The Third Crusade. The news of the fall of Jerusalem reached Europe even before the arrival there of Archbishop Josius of Tyre, whom the crusaders had sent with urgent appeals for aid. Pope Urban III soon died, shocked, it was said, by the sad news; his successor, Gregory VIII, issued a Crusade bull and called for fast and penitence.

Before a new Crusade could be organized, a modest recovery had begun in the East. Scarcely two weeks after Ḥaṭṭīn, Conrad of Montferrat, Baldwin V's uncle, had landed at Tyre with a small Italian fleet and a number of followers. He immediately established himself sufficiently to stave off an attack by Saladin, who was then intent on reaching Jerusalem without delay. When toward the end of 1188 Saladin released King Guy as he had promised, Conrad refused to submit to the King.

Guy's siege of Acre

In a daring move to establish his authority, Guy suddenly gathered his few followers and besieged Acre. Saladin was taken completely by surprise. When he finally moved his army toward the city, the crusaders in their camp outside had begun to receive reinforcements from the West, many under the banner of Henry of Champagne. Since Saladin could neither enter the city nor dislodge the crusaders, by winter of 1190–91 neither side had made progress, but the crusaders had suffered losses from disease and famine.

Among the victims of disease was Guy's wife, Sibyl, the source of his own claims to the throne. Many of the older barons who had thus far supported him now turned to Conrad. The marriage of Sibyl's sister, Isabel, to Humphrey of Toron was forthwith annulled, and she was constrained to marry Conrad. But Guy refused to abandon his claim to the throne. Such was the situation in May 1191, when ships arrived off Acre bringing welcome supplies and news of the approach of the armies of the Third Crusade.

The first ruler to respond to the papal appeal was William II of Sicily, who immediately abandoned a conflict with Byzantium and equipped a fleet that soon left for the East, though William himself died in November 1189. English, Danish, and Flemish ships also departed. Meanwhile, Gregory VIII had sent a legation to the Holy Roman emperor Frederick Barbarossa, now nearly 70 years old and approaching the end of an eventful career, which had included a protracted controversy with the papacy. But Frederick had made peace with the church, had participated in the Second Crusade, and for some time had been genuinely desirous of taking the cross again.

Frederick set out in May of 1189 with the largest Crusade army so far assembled and crossed Hungary into Byzantine territory. His troops caused disturbances in the Balkans, and there were subsequent misunderstandings with Emperor Isaac Angelus. Frederick was eventually persuaded to avoid Constantinople and cross the Dardanelles. In May 1190 he reached Iconium after defeating a Seljuq army. Thence his army crossed into Armenian territory. On June 10 Frederick, who had ridden ahead with his bodyguard, was drowned while attempting to swim a stream. His death broke the morale of the German army, and only a small remnant, under Frederick of Swabia and Leopold of Austria, finally reached Tyre. To Saladin and the Muslims, who had been seriously alarmed by Frederick's approach, Frederick's death seemed an act of God.

Death of Frederick Barbarossa

In Europe, Archbishop Josius had won over Philip II Augustus of France and Henry II of England. Henry died in 1189, leaving the cause to his son and successor, Richard I the Lion-Heart. But pending Anglo-French disputes were not easily settled. The extensive holdings of the English Angevin kings in France and especially Philip's desire to recover Normandy posed problems difficult if not impossible to lay aside even during a common enterprise. Thus, it was not until July 4, 1190, three years after Ḥaṭṭin, that the two kings met at Vézelay, prepared to move with their armies.

The two kings who finally led the Third Crusade were very different persons. Richard was in many ways an unstable character. He had opposed his father and was distrustful of his brothers. He could be lavishly generous even to his adversaries but often violent to anyone who stood in his way. His abilities lay not in administration, for which he had no talent, but in war, at which he was a genius. The favourite son of Eleanor of Aquitaine, Richard epitomized the chivalrous crusader and represented the contemporary troubadour view of war with all its aristocratic *courtoisie*. Richard could honour his noble Muslim opponents but be utterly ruthless to low-born captives.

The roles of Richard I and Philip II in the Third Crusade

Philip II Augustus, with his untidy hair and defective vision, was not an attractive person. Unlike Richard, he had been king for ten years and was a skilled and unscrupulous politician. He had no love for ostentation. Though no warrior himself, he was adept at planning sieges and designing siege engines. But he was a reluctant crusader whose real interests lay in the expansion of his own domains.

According to the original suggestion of King William II, the two kings met at Messina, in Sicily, where they signed an agreement outlining their mutual obligations and rights on the crusade. Philip arrived with the French fleet at Acre on April 20, 1191, and the siege was begun again in earnest.

After a stormy passage, Richard put in at Cyprus, where his sister Joan and his fiancée, Berengaria of Navarre, had been shipwrecked and held by the Byzantine ruler, the rebel prince, Isaac Comnenus. Isaac underestimated Richard's strength and attacked. Not only did Richard defeat and capture him, but he proceeded to conquer Cyprus. He then set sail and arrived on June 8 at Acre, where he gave new vigour to the siege.

A month later, after constant battering at the walls by siege engines and after Saladin's nephew had failed to fight his way into the city, the garrison surrendered in violation of Saladin's orders. Saladin was shocked by the news, but nevertheless he ratified the agreement, which provided for an exchange of prisoners and the restoration of the relic of the True Cross, which he had captured in 1187.

As the crusaders entered the city, there were disputes over the disposal of areas. Richard managed to offend Leopold of Austria, and Philip, who felt that he had fulfilled his crusader's vow and who was, in addition, unwell, left for home in August.

Internal disputes

Though the English resented Philip's departure, it did leave Richard in control. When the first prisoners were returned by Saladin, Richard found some flaw in the selection of them and refused to free the Saracen captives. Subsequent negotiations failed, and Richard, impatient to move toward Jerusalem, ordered them all killed, along with their wives and children. Thus, the truce arrangements were voided, and Saladin did not return the True Cross.

The English king was never to reach Jerusalem. He won a victory at Arsuf in 1191, reoccupied Jaffa as a base, and took Daron in 1192 but failed to progress further. Messages from home were urging his return. He had been in constant communication with Saladin and his brother al-ʿĀdil, and various peace proposals were made, which included marriage alliances. In fact, there seemed to be warm cordiality and considerable mutual respect. Finally, on September 2, 1192, the two opponents signed a treaty of peace to last five years. The coast from Jaffa north remained in Christian hands, but Ashkelon was to be demolished. Pilgrims were to have free access to the holy places. On October 9 Richard left. He was shipwrecked and finally fell into the hands of Leopold of Austria, who had not forgotten the slight at Acre.

The Third Crusade had failed of its major objective, the capture of Jerusalem. But the possession of Acre and much of the coast permitted the continued existence of a titular kingdom. Before he left, Richard had consented to the request that Guy, who had managed to lose the support of nearly all the barons, be deposed and Conrad immediately be accepted as king. No sooner was this done than Conrad was assassinated. Isabel was persuaded to marry Henry of Champagne, and Guy was given the governorship of Cyprus, where his record was far more successful than his ill-starred career in Jerusalem.

The most lasting achievement of the Third Crusade, however, was Richard's capture of Cyprus. In succeeding decades it gained in importance, not only as an outpost of the coastal possessions but also as a base for future crusades and finally as a kingdom in its own right.

THE FOURTH CRUSADE AND THE LATIN EMPIRE OF CONSTANTINOPLE

Pope Innocent III, despite manifold problems in the West, was the first pope since Urban II to be both anxious and able to consider the crusade a major papal concern. In 1198 he broached the subject of a new expedition through legates and encyclical letters. In 1199 a tax was levied on all clerical incomes—later to become a precedent for systematic papal income taxes—and Fulk of Neuilly, a popular orator, was commissioned to preach. At a tournament held by Thibaut III of Champagne, several prominent French nobles took the cross, and others joined later. Among them was Geoffrey of Villehardouin, who was to write one of the principal accounts of the Crusade. Contact was made with Venice to provide transport.

The involvement of Venice proved to be fateful. The republic had acquired considerable trading privileges

within the Byzantine Empire, and the growing number of Venetian merchants had long incurred the hostility of the Greeks. In 1171 Manuel Comnenus had ordered the arrest of Venetians, and 11 years later an aroused citizenry massacred a large number of Latins in Constantinople and insulted a papal legate. Further, following the advice of recently returned crusaders, the new Crusade was to be directed against Egypt, now the centre of Muslim power in the Levant but a government with which Venice was closely related commercially. Venetian policy under the aging and blind but ambitious Doge Enrico Dandolo was therefore potentially at variance with that of the Pope and the crusaders. Nevertheless, an agreement was made providing for payments to the Venetians for transportation and an equal division of conquests.

The Venetians' involvement

The crusader army that arrived at Venice in the summer of 1202 was somewhat smaller than had been anticipated, since some of the crusaders were travelling directly from France. Even so, there were not sufficient funds to pay the Venetians. Accordingly, the crusaders accepted the suggestion that in lieu of payment they assist the Venetians in the capture of the Hungarian city of Zara. This was done despite the opposition of many crusaders both to the diversion of the enterprise and to the attack on a Christian city. Innocent was informed of the plan, but his veto was disregarded. Reluctant to jeopardize the Crusade, Innocent gave conditional absolution to the crusaders—not, however, to the Venetians.

The fall of Zara (November 1202) added to the Pope's already considerable misgivings over the transformation of the whole undertaking from a Crusade under papal auspices to one under lay direction. After the death of Thibaut of Champagne, the leadership of the Crusade passed to Boniface of Montferrat, a friend of the Hohenstaufen Philip of Swabia. Both Boniface and Philip had married into the Byzantine imperial family. In fact, Philip's brother-in-law, Alexius, son of the deposed and blinded emperor Isaac Angelus, had appeared in Europe seeking aid and had made contacts with the crusaders.

Innocent was aware of a plan to divert the Crusade to Constantinople in order to gain the throne for Alexius, who, in turn, promised subsidies to the crusaders. Accordingly, Innocent ordered Boniface of Montferrat to publish immediately his original letter excommunicating the Venetians, which he had deliberately refused to do, and expressly forbade any attack on Constantinople. But the papal letter arrived after the fleets had left Zara, and in the summer of 1203 Constantinople fell, Alexius III was deposed, and "Alexius IV" was crowned as co-emperor with his father. All Innocent could do was reprimand the leaders and order them to proceed forthwith to the Holy Land. No doubt he hoped that a union of the churches would result and the Crusade be thereby promoted. A few crusaders did leave, but most did not.

The takeover of Constantinople

Following the assassination of the new "emperor" by a resentful Greek population, the Venetians and crusaders themselves took over the city and the government of the empire. It was decided that 12 electors (six Venetians and six crusaders) should choose an emperor who would have one-quarter of the imperial domain. The other three-quarters were to be divided. The clergy of the party not belonging to the emperor elect were to have Hagia Sophia and choose a patriarch. A small amount of property was specifically designated to support the clergy. The rest was to be considered booty and divided. On April 13, 1204, the great city fell and with the permission of the leaders was subjected by the rank and file to pillage and massacre for three days. Many priceless icons, relics, and other objects later turned up in western Europe, a large number in Venice.

However much Innocent III may have hoped that a friendly Constantinople would aid the Crusade and promote the reunion of Eastern and Western churches, he was aghast at the sack of Constantinople and castigated the crusaders and Venetians in no uncertain terms. But the situation was beyond his control, especially after his legate, on his own initiative, had absolved the crusaders from their vow to proceed to the Holy Land.

When order had been restored, the crusaders and the Venetians proceeded to implement their agreement; Baldwin of Flanders was elected emperor and the Venetian Thomas Morosini chosen patriarch. But the lands parcelled out among the leaders did not include all the former Byzantine possessions. The imperial government continued in Nicaea, and an offshoot empire of Trebizond, at the eastern end of the Black Sea, lasted until 1461. There was also established a Byzantine Despotate of Epirus, and the Bulgarians remained hostile. Various Latin–French lordships throughout Greece—in particular, the duchy of Athens and the Principality of the Morea—did provide cultural contacts with western Europe and promoted the study of Greek. There was also a French impact on Greece. A collection of laws, the *Assises de Romanie*, was edited. The *Chronicle of the Morea* appeared in both French and Greek (and later Aragonese) versions. Impressive remains of crusader castles and Gothic churches can still be seen. Nevertheless, the Latin Empire always rested on shaky foundations, and in 1261 a sadly diminished domain was reconquered from the Latins by Michael Paleologus with the aid of the Genoese, the traditional rivals of Venice.

The survival of imperial authority

It is difficult to see in the Fourth Crusade anything but unmitigated disaster to the cause of either Christianity or the Crusade. Not only had a promising Crusade been diverted, but the new Latin principalities proved to be more attractive to Westerners who might otherwise have gone to the Levant and even to a number of former residents of the states of Syria. The rift between the Eastern and Western churches widened, and Greek popular resistance to any schemes of reunion with the empire intensified. The Byzantine Empire, for centuries a bulwark against invasion from the East, was damaged beyond repair.

II. The decline of the crusading movement

THE LATER CRUSADES AND THE DECLINE OF THE LATIN ENCLAVES

The Latin East after the Third Crusade. Saladin died on March 3, 1193, not long after the departure of the Third Crusade. One of the greatest of the Muslim leaders and a man devoutly religious and deeply committed to *jihād* against the infidel, he was, yet, respected by his opponents. His death led once again to divisions in the Muslim world, and his Ayyūbid successors were willing to continue a state of truce with the crusaders, which lasted into the early years of the 13th century. The truce was politically and economically advantageous to both sides, and the Italians were quick to make profitable trade connections in Egypt. The Franks were able to adjust themselves to a new situation and to organize what in effect was a new titular kingdom of Jerusalem centring at Acre, generally known as the Second Kingdom.

In 1194 Amalric of Lusignan succeeded his brother Guy as ruler of Cyprus, where he later accepted investiture as king from the chancellor of the emperor Henry VI. In 1197, following the death of Henry of Champagne, Amalric succeeded to the throne of Jerusalem-Acre, and in 1198 he married the thrice-widowed Isabel. He chose, however, to govern his two domains separately, and in Acre he proved to be an excellent administrator. The *Livre au roi* ("Book of the King"), an important section of the *Assises de Jérusalem*, dates from his reign. He also dealt wisely with Saladin's brother, al-ʿĀdil of Egypt. On Amalric's death in 1205 the kingdoms of Cyprus and Jerusalem-Acre were divided, and in 1210 the latter was given to John of Brienne, a French knight nominated by Philip Augustus, who came east and married Conrad's daughter, Mary.

There were also readjustments in the two now-reduced northern states. When Raymond III of Tripoli died (1187), his county passed to a son of Bohemond III of Antioch, thus uniting the two principalities. In general, Antioch-Tripoli continued to pursue the relatively independent course laid down by Bohemond III.

Armenia was more closely involved in Latin politics, partly as a result of marriage alliances with the house of Antioch-Tripoli. King Leo II of Armenia joined the cru-

saders at Cyprus and Acre. Desirous of a royal crown, he had approached both Pope and Emperor, and in 1198, with papal approval, royal insignia were bestowed by Archbishop Conrad of Mainz, in the name of Henry VI. At the same time, the Armenian Church officially accepted a union with Rome, which, however, was never popular with the lower clergy and people.

The Fifth Crusade. In Europe after the Fourth Crusade something of the original crusading fervour could still be found in certain areas of society. Even children became the victims of mass hysteria, and in 1212, in the so-called Children's Crusade, thousands of youngsters from France and Germany set out to free the Holy Land, only to be lost, shipwrecked, or sold into slavery. But there was considerable disillusionment among the nobility, especially when the same religious indulgence was promised to participants in a "Crusade" against heretics in southern France and later against secular opponents of the popes. Innocent III, nevertheless, despite his preoccupation with the kings of England and France, a civil war in Germany, heresy, the advance of Islām in Spain, and his strenuous efforts to promote widespread ecclesiastical reform, renewed his efforts to organize another expedition. Preachers were designated; even troubadours contributed to the propaganda. The final canon promulgated by the Fourth Lateran Council of 1215 was an elaborate Crusade plan repeating earlier prohibitions on the transport of military supplies to Muslims. Levies on clerical incomes were again authorized.

The Children's Crusade

The young emperor Frederick II had promised Innocent to go, but Honorius III permitted him repeated postponements to settle the affairs of Germany. Thus, the first contingents of the Fifth Crusade left without him. Nor did they accomplish anything significant until the arrival of a Frisian fleet with more German crusaders enabled an impressive force to set out for Egypt in May 1218 under the leadership of John of Brienne. An attack on Egypt had been in the forefront of all planning since the time of the Third Crusade, not as the objective of permanent conquest but as a bargaining point for the recovery of Jerusalem. By August the crusaders had captured a strategic tower at Damietta (Dumyāṭ).

The Siege of Damietta

In September the expedition organized under papal auspices and consisting mainly of French crusaders arrived under Cardinal Pelagius as legate. Since Pelagius regarded the crusaders as being under the jurisdiction of the church, he declined to accept the leadership of John of Brienne. Moreover, he was an imperious person who did not hesitate to interfere in military decisions.

By February 1219 the Muslims were seriously alarmed and offered peace terms that included the cession of Jerusalem. King John and many of the crusaders were eager to accept. But Pelagius, supported by the military orders and the Italians, refused. Damietta was finally taken on November 5, 1219. For over a year no progress was made, though Pelagius remained optimistic, still expecting the arrival of Frederick II and convinced, by rumour, of the imminent approach of a legendary oriental Christian "King David." In July 1221 he ordered an advance toward Cairo, but Nile floods forced him to retreat and to agree to a truce of eight years and an exchange of prisoners, terms far less favourable than those he had previously rejected.

The Fifth Crusade, the last in which the papacy took an active part, was an impressive effort against a divided Muslim opposition. Coming close to success, it failed largely because of divided leadership and the frequently unwise decisions of Pelagius. It might perhaps have succeeded if Frederick II had set out as promised, and it is significant that disillusioned critics blamed Emperor and Pope as well as Pelagius. All in all, it was a dreary episode, relieved only by the presence of Francis of Assisi, to whom Pelagius reluctantly gave permission to cross the lines, where he was courteously received by the sultan al-Kāmil. Francis' visit to the East was a preliminary step in the establishment of a Franciscan province in the Holy Land, a step soon imitated by the Dominicans.

The arrival of missionary orders

The Sixth Crusade. The failure of the Fifth Crusade placed a heavy responsibility on Frederick II. His motives as a crusader are difficult to assess. Always a controversial figure, he was to some the archenemy of the papacy, to others the greatest of emperors. His intellectual interests included Islām, and his attitude might seem to be more akin to that of the Eastern barons than the typical crusader from the West. Through his marriage with John of Brienne's daughter Isabella (Yolande), he had a claim first to the kingship and then, on her death in 1228, to the regency of Jerusalem (Acre) for his son Conrad. As emperor he could claim a suzerainty over Cyprus by virtue of the cession by his predecessor Henry VI. Frederick had finally agreed to terms that virtually placed his expedition under papal jurisdiction. Yet his entire Eastern policy was inextricably connected with his European concerns: Sicily, Italy and the papacy, and Germany. Cyprus-Jerusalem became, as a consequence, part of a greater imperial design.

Most of the fleet of the Sixth Crusade had left Italy in the late summer of 1227, but Frederick was delayed by illness. During the delay he received envoys from Sultan al-Kāmil of Egypt, who, threatened by the ambitious designs of his Ayyūbid brothers, was disposed to negotiate. Meanwhile, Pope Gregory IX, less patient than his predecessor, had rejected the Emperor's plea of illness and excommunicated him. Thus, when Frederick departed in the summer of 1228 with the remainder of his forces, he was in the equivocal position of a crusader under the ban of the church. He arrived in Cyprus on July 21.

In Cyprus, John of Ibelin, the leading member of the widely influential Ibelin family, had been named regent for the young Henry I. Along with most of the other barons, he was willing to recognize the Emperor's rights as suzerain in Cyprus. But in Acre, since news of Isabel's death had arrived, the Emperor could, according to the barons' lawyers, claim only a regency for his infant son. John obeyed the Emperor's summons to meet him in Cyprus, but despite intimidation refused to surrender his lordship of Beirut and insisted that his case be brought before the High Court of barons. The matter was set aside, and Frederick left for Acre.

At Acre, Frederick met more opposition. News of his excommunication had already arrived, and many who might otherwise have supported him now refused to cooperate. Dependent, therefore, on the Teutonic Knights, an organization formed by Germans who remained in the east after an expedition in 1197 and now under the direction of Hermann of Salza, and his own small contingent of German crusaders, he was forced to attempt what he could by diplomacy. Negotiations, accordingly, were reopened with al-Kāmil of Egypt.

The treaty of 1229

The resulting treaty of 1229 is unique in the history of the Crusades. By diplomacy alone and without major military confrontation, Jerusalem, Bethlehem, and a corridor running to the sea were ceded to the Kingdom of Jerusalem. Exception was made for the Temple area, the Dome of the Rock, and the Aqṣā Mosque, which the Muslims retained. The peace was to last for ten years.

The benefits of the treaty of 1229 were more apparent than real. The areas ceded were not easily defensible, and Jerusalem soon became a prey to disorder. Furthermore, the treaty was denounced by the devout of both faiths. When Frederick, still under excommunication, entered the city, the Patriarch placed it under interdict. No priest was present, and Frederick was forced to place a crown on his own head while one of the Teutonic Knights read the ceremony. Leaving agents to take over the administration, he hastily returned to Europe and at San Germano made his peace with the Pope (July 23, 1230). Thereafter, his legal position was secure, and the Pope ordered the Patriarch to lift the interdict.

What followed in Jerusalem and Cyprus, however, was not orderly government by the Emperor's agents but civil war. For Frederick's imperial concept of government was totally opposed to the now well-established pre-eminence of the Jerusalem baronage. The barons of both Jerusalem and Cyprus, in alliance with the Genoese and a commune formed at Acre that elected John of Ibelin as mayor, resisted the imperial deputies, who had the support of the Pisans, the Teutonic Knights, Bohemond of

Antioch, and a few nobles. The clergy, the other military orders, and the Venetians stood aloof.

The barons were able to win out in Cyprus, and in 1233 Henry I was recognized as king. Even after John of Ibelin, the "Old Lord of Beirut," died in 1236, the resistance continued. In 1243 a parliament at Acre refused homage to Frederick's son Conrad, unless he appeared in person, and named Alice, queen dowager of Cyprus, as regent.

Triumph of baronial rule

Thus it was that baronial rule triumphed over imperial administration in the Levant. But the victory of the barons brought to the kingdom not strength but continued division, the more serious as new forces were appearing in the Muslim world. The Khwārezmian Turks, pushed south and west by the Mongols, had upset the power balance and gained the support of Egypt. After the ten years' peace had expired in 1239, a poorly organized Crusade under Thibaut IV of Champagne and Richard of Cornwall accomplished little. In 1244 an alliance of Jerusalem and Damascus failed to prevent the capture and sack of Jerusalem by Khwārezmians with Egyptian aid. All the diplomatic gains of the preceding years were lost.

The Crusade of Louis IX of France. In June 1245, a year following the final loss of Jerusalem, Pope Innocent IV opened a great ecclesiastical council at Lyons. Although urgent appeals for help had come from the East, it is unlikely that the Crusade was uppermost in the Pope's mind for a combination of crises confronted the church: massive complaints of clerical abuses, increasing troubles with the emperor Frederick II in Italy, and the advance of the Mongols into eastern Europe. Nevertheless, when King Louis IX of France announced his intention to lead a new Crusade, the Pope gave it his support and authorized the customary levy on clerical incomes.

Character of Louis IX

As a crusader, the saintly Louis was the antithesis of his predecessor, Frederick. Possessed of a rare combination of religious devotion with firmness as a ruler and bravery as a warrior, he seemed the very ideal of the crusader. He was beloved by his subjects and respected abroad. He ardently believed the Crusade to be God's work, and he was far from sympathetic with the Pope's using Crusade propaganda against the Emperor.

It was three years before Louis was ready to embark. Peace had to be arranged with England, transport had to be provided by Genoa and Marseilles, and funds had to be raised from the King's domain and from the towns. When the King embarked in August 1248 he was accompanied by his queen; his brothers Robert of Artois and Charles of Anjou; many distinguished French nobles, including John of Joinville, author of the *Vie de St. Louis;* and a small contingent of English. His army was a formidable one, numbering perhaps 15,000. France was left in the experienced hands of the queen mother, Blanche of Castile.

The Crusade arrived at Cyprus in September, and it was again decided to attack Egypt. Since a winter campaign was not feasible and Louis rejected the suggestion that he attempt negotiations—though he did exchange envoys with the Mongols—it was not until May 1249 that an expedition of some 120 large and many smaller vessels got under way. Fortune favoured them at first, and Damietta was again in Christian hands by June. Shortly afterward, the army was strengthened by the arrival of Louis's third brother, Alphonse of Poitiers. Sultan aṣ-Ṣāliḥ-Ayyūb's death was followed by confusion in Cairo, which, after some argument, had become the crusaders' objective. In February 1250 Robert of Artois led a surprise attack on the Egyptian camp two miles (three kilometres) from al-Manṣūrah and, rejecting the advice of more experienced campaigners, impetuously moved ahead, only to be trapped within the city. Many knights lost their lives. Louis soon arrived with the main army and won another victory, albeit a costly one, near al-Manṣūrah. It was the last crusader success.

Meanwhile, Tūrān-Shāh, the Sultan's son, arrived and temporarily dominated dissident factions in Cairo. Frankish supply ships from Damietta were intercepted, and before long the crusaders were suffering from famine and disease. Louis, reluctant to abandon a work to which he had dedicated his very kingdom, perhaps delayed too long before ordering a retreat. But he refused the pleas of others to protect himself by moving out ahead. Instead, he led his soldiers and along with many of them was captured as the Muslim forces closed in.

The capture of Louis

The King and nobles were held for ransom, but many non-noble captives were killed. The Queen, who had just given birth to a son sorrowfully named John Tristan, managed with great courage to secure sufficient food and to persuade the Genoese and Pisans not to evacuate Damietta until it could be ceded formally by treaty and the King's ransom arranged. On May 6, 1250, the King was released and Damietta surrendered.

Despite the pleadings of his advisers, Louis did not return home immediately. He felt bound in conscience to negotiate the release of as many prisoners as possible, and he was able before he left in April 1254, four years later, to improve the defenses of the kingdom by strengthening a number of fortifications. Thus, he atoned in some small measure for the failure of the Crusade.

During these same years the southern wing of the Mongols under Hülegü overran Mesopotamia and in 1258 took Baghdad, thus ending the venerable ʿAbbāsid caliphate. But two years later (1260) the Mamlūks of Egypt, a new dynasty that had arisen from the leaders of former slave bodyguards of the sultan, defeated the Mongols at ʿAin Jālūd in Syria and halted their southward advance. The Muslim states of Syria were caught in the middle, and the Latin states were in grave danger. King Hethum (Hayton) of Armenia threw in his lot with the Mongols, and his son-in-law Bohemond VI of Antioch-Tripoli followed suit. But the barons at Acre were still more disposed to deal with the Muslims, whom they knew, than with the terrifying and unknown Mongols.

The accession of Baybars as Mamlūk sultan

In 1260, after murdering his predecessor, Baybars became sultan of Egypt. Though this famous Mamlūk sultan did not live to see the final fall of the Latin states, before his death in 1277 he had reduced them to a few coastal outposts. Baybars was ruthless and totally devoid of the generous chivalry that the crusaders had admired in Saladin. Most of his conquests were followed by massacre of the inhabitants, often including the native Christians, especially when they had been in league with the Mongols. In 1265 he took Caesarea, Haifa, and Arsuf. The following year he conquered Galilee and devastated Cilician Armenia. In 1268 Antioch was taken and all the inhabitants slaughtered. The great Hospitaller fortress of Krak des Chevaliers fell three years later.

These disasters again brought pleas for aid from the West. King Louis, who felt a heavy responsibility for the collapse of his first effort, began to make new preparations in 1267. But the French king's second venture, the Eighth Crusade, never reached the East. For reasons that are not entirely clear but probably owed something to the influence of Louis's brother Charles of Anjou, the expedition went instead to Tunis. Charles had recently been named by the papacy as the successor to the Hohenstaufens in Sicily. In 1268 he defeated Conradin, the last of the Hohenstaufen line, and was soon involved in grandiose Mediterranean projects, which ultimately included even Byzantium.

There was little support in Europe for Louis's Crusade, which embarked from southern France in July 1270. Moreover, soon after the French landed in North Africa, disease struck the troops and claimed the lives of both Louis and his son John Tristan. Charles arrived with the Sicilian fleet in time to bargain for an indemnity to evacuate the remnants of the army. Thus, the Crusade ended in tragedy and brought no help to the East. Moreover, except for the expedition of Prince Edward of England (1271–72), who had arrived in North Africa too late to be of assistance, Louis's Crusade was the last.

The final loss of the crusader states. Although Europe was aware of the gravity of the Eastern situation, it was both unwilling and unable to give substantial aid, as Pope Gregory X discovered following his pleas at the Council of Lyons in 1274. The sense of common endeavour that had produced the First Crusade was no

longer to be found. Not only had political tensions increased, but the reforms of the Fourth Lateran Council of 1215 had not produced the results hoped for. Criticism of ecclesiastical policies was more outspoken.

Crusading had also become more expensive. The time had passed when a Crusade army was made up of knights serving under a lord and paying their own way. Economic pressures were causing many nobles to seek royal service. Royal armies, therefore, tended to become more professional, and many knights as well as foot soldiers served for pay. King Louis's crusades had necessitated complicated financing, causing his successors serious financial trouble.

Inevitably, Outremer was affected by the general malaise of the West, and the chronic divisions that were a major cause in Outremer's downfall paralleled those of Europe. From the time of Frederick II, the kingdom had been governed by absentee rulers, in theory the Hohenstaufens represented in the East by agents, followed after 1243 by regents of the Jerusalem dynasty chosen by the High Court of barons. Finally, in 1268, on the death of the last Hohenstaufen, the crown was given to Hugh III of Cyprus. But in 1276 he returned to Cyprus, thoroughly frustrated. Then in 1277 Charles of Anjou, with papal approval, bought the rights of the nearest claimant and sent his representative. But Charles's principal interests at the time lay in Byzantium. Finally, after Charles's death in 1285, the barons once again chose a native ruler, Henry II of Cyprus.

Successive regents had failed to dominate the Jerusalem baronage, ultimately resulting in the disintegration of the entire structure of Outremer into separate parts. Antioch-Tripoli before its fall had been increasingly aloof and through intermarriage closely tied to Armenia. In Acre, the seat of government of the kingdom, there was a commune of barons and bourgeois. Immigration had ceased, and the barons were now reduced in numbers as old families had died out. Some resided in Cyprus, and others were nominal lords in Palestine of fiefs of lands actually under Muslim control. The military orders, habitually in conflict, were virtually distinct entities with extensive connections in Europe. The bourgeois population had also considerably altered in composition during the 13th century. Many criminals and other undesirables had found their way to Acre. More important, the earlier homogeneity of a French character had given way to an Italian predominance. But the Italians of Outremer were as divided as they were in Italy. The Genoese–Venetian rivalry extended to the Levant and occasionally, as in Acre in 1256, resulted in outright war.

Papal concern for Outremer was not confined to efforts to enlist military aid. Its financial support was continuous. Popes took an active interest in diplomacy and exchanged envoys with Eastern rulers, both Muslim and Mongol. Further, the 13th-century patriarchs of Jerusalem, commonly named by the pope, were also papal legates. But neither absentee king nor pope nor patriarch-legate could bring to the Latin East the unity necessary for its survival.

Thus, to the crusaders, divided among themselves and isolated, the death of Baybars in 1277 brought only temporary respite. In 1280 they again failed to join the Mongols, whom Sultan Qalāʾūn defeated in 1281. The ineffectiveness of the Jerusalem administration was becoming apparent even to Easterners, and the Il-Khan Abagha sent his deputy Rabban Sauma to the kings of Europe and the Pope to seek an alliance. The effort was fruitless.

The fall of Tripoli and Acre

Tripoli fell in 1289, and Acre, the last crusader stronghold on the mainland, was besieged in 1291. After a desperate and heroic defense, the city was taken by the Mamlūks, and the inhabitants who survived the massacre were enslaved. Acre and all the castles along the coast were systematically destroyed.

Perhaps because they sensed their greater isolation, the Franks of the 13th century seem not to have developed further the distinctive culture of their predecessors. The remarkable palace of the Ibelins at Beirut, built early in the century, did boast Byzantine mosaics. But, no doubt partly because of King Louis's four-year stay in the kingdom, existing remains of churches and castles indicate a close following of contemporary French Gothic, and manuscripts are more strongly French in style. Literary tastes were also distinctly French. At the coronation festivities for Henry II in 1286, in total disregard—or perhaps in chivalrous defiance—of the ruin surrounding them, the nobles amused themselves by acting out the romances of Lancelot and Tristan.

The greatest cultural achievement of the Second Kingdom was the famous collection of legal treatises the *Assises de Jérusalem*. Those parts that were compiled in the middle years of the century and, therefore, in the atmosphere of the wars against the agents of Frederick II constitute a veritable charter of baronial rights. In fact, two of the authors were members of the Ibelin family, and a third, Philip of Novara, was a close associate. These sections indicate a shift from the earlier *Livre au roi*, which more nearly reflects the attitudes of the 12th century. Nevertheless, the *Assises* belong to medieval Europe's legal renaissance.

In many respects, therefore, the Franks of the last days of Outremer were what in a modern setting might be called colonials, but they were entirely on their own, for there was no mother country to come to their rescue.

The later Crusades and the Kingdom of Cyprus. Europe was dismayed by the disaster of 1291 but not surprised and shocked as it had been in 1187, for the end had been foreseen. Pope Nicholas IV had tried to organize aid before 1291, and he and his successors continued to do so afterward, but without success. France, which had always been the main bulwark of the Crusades, was in serious conflict with England, which eventually led to the Hundred Years' War in 1337. Moreover, although it could scarcely have been understood at the time, western Europe around 1300 was experiencing the first impact of a population decline and what proved to be a prolonged economic depression.

The fate of the military orders

In the East, the military orders could no longer offer a standing nucleus of troops. In 1308 the Hospitallers took Rhodes and established their headquarters there. In 1344, with some assistance, they occupied Smyrna, which they held until 1402. Meanwhile, the Teutonic Knights had moved their operations to the Baltic area. The Templars were less fortunate. In 1308 the French Templars were arrested by Philip IV, and in 1312 the order was suppressed by Pope Clement V.

It is not surprising, therefore, that such response as did follow papal urgings was largely in the form of Crusade theories. For some years after 1291 various projects were elaborated, all designed to avoid previous mistakes and explore new tactics. The Franciscan missionary Ramon Llull (died 1315), for example, in his *Liber de fine*, suggested a campaign of informed preaching as well as military force. Pierre Dubois (*c.* 1305–07) submitted a detailed scheme for a Crusade to be directed by Philip IV of France, and in 1321 Marino Sanudo in his *Secreta fidelium crucis* produced an elaborate plan for an economic blockade of Egypt. But none of these or any other such schemes was put into effect.

King Peter I of Cyprus finally organized an expedition that in 1365 succeeded in a temporary occupation of Alexandria. After a horrible sack and massacre, the unruly crusaders returned to Cyprus with immense booty. Peter planned to return, but no European aid was forthcoming, and after his murder in 1369 a treaty of peace was signed. No further crusades set out with Jerusalem as the objective. What followed were not really crusades in the old sense but campaigns such as the crusades of Nicopolis in 1396 and Varna in 1444, whose purpose was to defend Europe against the Ottoman Turks, a new power in the East.

With the failure of all attempts to regain a foothold on the mainland, Cyprus remained the sole crusader outpost, and after 1291 the island kingdom was faced with a serious refugee problem. It was in Cyprus that many of the institutions established by the Franks survived. For, although Jerusalem and Cyprus normally had separate governments, through intermarriage and the exigencies of diplomacy the histories of the two had become inter-

woven. Regents of one were often chosen from among relatives in the other. It has been noted that many Jerusalem barons resided in Cyprus. With suitable modifications, the *Assises de Jérusalem* applied on the island. As on the mainland, the French character of the Cypriot Latins is evident in the remains of Gothic structures.

In one respect Cyprus did differ from the mainland. Whereas the First Kingdom had established a reasonable modus vivendi with its native population, such was not the case in the island kingdom. Many Greek landholders had fled, and those who remained apparently suffered a loss of status. All Greeks resisted the Latinizing efforts of the early-13th-century popes and their representatives. Innocent IV was more flexible, but tension persisted until the Turkish conquest in the 16th century.

THE RESULTS OF THE CRUSADES

The entire structure of European society changed during the 12th and 13th centuries, and there was a time when this change was attributed largely to the Crusades. Historians now, however, tend to view the Crusades as only one, albeit a significant, factor in Europe's development. It is probable, for example, that the departure of unruly elements on the First Crusade aided the King of France and the great feudatories in preserving order and extending their authority. It is also likely that the disappearance of old families and the appearance of new ones can be traced in part to the Crusades, but generalizations must be made with caution. It must, moreover, be remembered that, while some crusaders sold or mortgaged their property, usually to ecclesiastical foundations, others made it over to relatives. The loss of life was without doubt considerable; many, however, did return to their homes.

Economic and social impact on Europe

The sectors acquired by burgeoning Italian cities in the crusader states enabled them to extend their trade with the Muslim world and led to the establishment of trade depots beyond the Crusade frontiers, some of which lasted well beyond 1291. The transportation they provided was significant in the development of shipbuilding techniques. Italian banking facilities became indispensable to popes and kings. Catalans and Provençals also profited and, indirectly, so did all of Europe. Moreover, returning crusaders brought new tastes or increased the demand for spices, Oriental textiles, and the like. But such demands can also be attributed to changing lifestyles and commercial growth in Europe itself.

Religious effects

The papacy launched the First Crusade, and crusades remained a constant concern. While initial success undoubtedly enhanced papal prestige, later failures had the opposite effect, and the diversion of Crusade propaganda to war against heretics or the emperor aroused criticism. So also did the levies on clerical incomes, though it must be noted that papal crusade financing played an important role in Europe's economic development.

The establishment of the Franciscan and Dominican friars in the East during the 13th century made possible the promotion of missions within the Crusade area and beyond. Papal bulls granted special facilities to missionary friars, and popes sent letters to Oriental rulers soliciting permission for the friars to carry on their work. Often the friars accompanied or followed Italian merchants, and, since the Mongols were generally tolerant of religious propaganda, missions were established in Iran, Inner Asia, and even China. But, since Islāmic law rigidly prohibited propaganda and punished apostasy with death, conversions from Islām were few. The Dominican William of Tripoli had some success, presumably within the crusaders' area; he and his colleague, Ricoldus of Montecroce, both wrote perceptive treatises on Islāmic faith and law. Other missionaries usually failed, and many suffered martyrdom. In the 14th century the Franciscans were finally permitted to reside in Palestine as caretakers for the holy places, but not as missionaries.

The Crusades, especially the Fourth, so embittered the Greeks that any real reunion of the Eastern and Western churches was out of the question. But certain groups of Eastern Christians came to recognize the authority of the pope and were usually permitted to retain their native liturgies. Though most of the missions that grew out of the Crusades collapsed with the advance of the Ottoman Turks in the Middle East in the mid-14th century, some of the contacts which the Western Church had made with its Eastern brethren remained.

The association of missions with Crusades posed a moral theological problem that troubled medieval thinkers. Thirteenth-century theologians held that conversion could not be forced, but most agreed that force could legitimately be used to preserve a situation in which peaceful propaganda was possible, and they continued to support the Crusade. Furthermore, such was Europe's fear of Muslim power that the Crusade idea persisted well into the 17th century, and the conviction that, in certain circumstances, war might be just became more deeply enrooted in the conscience of the West. Along with the now generally accepted use of the word crusade to denote any common endeavour in a worthy cause, this is one of the most enduring results of the movement.

The justification of warfare

Unlike Sicily and Spain, the Latin East did not, it seems, provide an avenue for the transmission of Arabic science and philosophy to the West. But the Crusades did have a marked impact on the development of Western historical literature. From the beginning there was a proliferation of chronicles, eyewitness accounts, and later more ambitious histories, in verse and prose, in the vernacular as well as in Latin.

It must be emphasized that the Crusades should not be viewed too exclusively in terms of cause and effect. If they were sometimes a factor contributing to changes in the West, so also they were affected by those changes. Rather, the Crusades and the Latin Kingdom should be considered as an integral part of the diversified culture of Europe in the Middle Ages.

BIBLIOGRAPHY. The standard bibliographical work on the Crusades, including sources, secondary studies, and journal articles, is H.E. MAYER, *Bibliographie zur Geschichte der Kreuzzüge* (1960). A similar study in English, somewhat briefer but well organized, is A.S. ATIYA, *The Crusade: Historiography and Bibliography* (1962). The best full-scale treatments of the Crusades in English are S. RUNCIMAN, *A History of the Crusades*, 3 vol. (1951–54); and K.M. SETTON (ed.), *A History of the Crusades*, 2nd ed., vol. 1–2 (1969), a cooperative work by a number of historians. Additional volumes are in preparation. Briefer studies include: A.S. ATIYA, *Crusade, Commerce, and Culture* (1962), which places the Crusade movement in the context of contemporary developments, Eastern as well as Western; E. BARKER, *The Crusades* (1923); and R.A. NEWHALL, *The Crusades* (rev. ed., 1963). Two recent, significant studies are F. COGNASSO, *Storia delle crociate*, vol. 1 (1967, second volume in preparation); and H.E. MAYER, *Geschichte der Kreuzzüge* (1965).

On the institutions of the Crusader states, the most important recent work is J. PRAWER, *Histoire du Royaume latin de Jérusalem* (1969, French trans. of vol. 1 of the two-volume Hebrew work of 1963). Still indispensable are J.L. LAMONTE, *Feudal Monarchy in the Latin Kingdom of Jerusalem, 1100–1291* (1932); and J. RICHARD, *Le Royaume latin de Jérusalem* (1953). D.C. MUNRO, *The Kingdom of the Crusaders* (1935), ed. by A.C. KREY, contains the Lowell lectures given in 1924 by the dean of American Crusade historians.

J.A. BRUNDAGE (ed.), *The Crusades: Motives and Achievements* (1964), presents the points of view of different scholars. T.S.R. BOASE, *Kingdoms and Strongholds of the Crusades* (1971), is a recent study of Crusade architecture with illustrations. R.C. SMAIL, *Crusading Warfare, 1097–1193* (1956), is a study of the campaigns of the 12th century. E. SIVAN, *L'Islam et la Croisade, idéologie et propagande dans les réactions musulmanes aux Croisades* (1968), is an important analysis of the Muslim reaction to the Crusades.

Useful selections of sources in English translation are: J.A. BRUNDAGE, *The Crusades: A Documentary Survey* (1962); F. GABRIELI (comp.), *Storici arabi delle Crociate* (1963; Eng. trans., *Arab Historians of the Crusades*, 1969); R. PERNOUD, *Les Croisades* (1960; Eng. trans., *The Crusades*, 1963).

(M.W.B.)

Crustacea

Crustacea is the name given to a large class of invertebrate animals of the phylum Arthropoda. They include the decapods (crabs, lobsters, shrimps, crayfish, prawns); amphipods (beach, or sand, hoppers); euphausiids (krill); isopods (pill bugs, wood lice, or sow bugs); cirripedes (barnacles); cladocerans (water fleas); anostracans (brine

shrimps, fairy shrimps); and a vast multitude of less familiar forms that are not distinguished by any popular names. Crustaceans are distinguished from the members of other arthropod classes by being generally of aquatic habit; by breathing through either gills or the general surface of the body; and by having two pairs of antenna-like appendages in front of the mouth and at least three pairs of postoral (*i.e.*, posterior to, or behind, the mouth) limbs, which act as jaws. There is so much diversity, both of structure and of habits, within the class that it is all but impossible to give a brief definition that would apply to all members. (W.T.Ca./W.L.Sc./Ed.)

By courtesy of (*Branchipus*) R.C. Moore (ed.), *Treatise on Invertebrate Paleontology*, U.S. Geological Society of America and The University of Kansas; after (*Anaspides*) Siewing and (*Thermosbaena*) Monod in A. Kaestner, *Invertebrate Zoology* (1970), John Wiley & Sons, Inc.; from (*Mesocypris terrestris*) *Memoires of the Institut Francais d' Afrique Noire* (1966); (others) *Invertebrate Zoology* by Paul A. Meglitsch, Copyright © 1967 by Oxford University Press, Inc., reprinted by permission

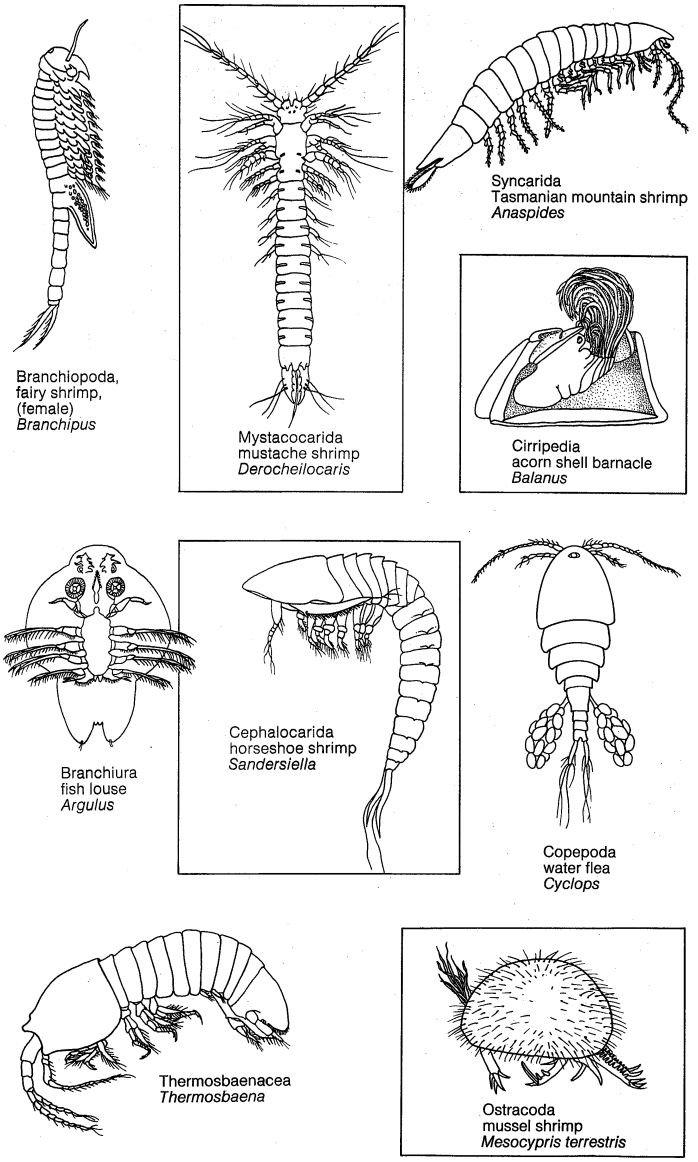


Figure 1: Body plans of representative crustaceans (branchiopod, mystacocarid, syncarid, cirripede, branchiuran, cephalocarid, copepod, thermosbaenid, and ostracod).

GENERAL FEATURES

Distribution and abundance. The majority of crustaceans are aquatic animals, either marine or freshwater. Present in greatest diversity, often in vast numbers, in all oceans from the tropics to polar seas, they are most numerous in coastal waters and in the surface and upper layers of the open sea, where copepods commonly constitute 70 percent of the animal life. They decrease in number with increasing depths. No decapods live below 5,500 metres (18,000 feet), but some copepods, isopods, and amphipods occupy deep isolated trenches at 7,000 to 10,000 metres (23,000 to 33,000 feet). Many species live in freshwater lakes, ponds, transient pools, and rivers throughout the world, some at altitudes of over 5,000 metres (16,000 feet). Large geographically isolated lakes and seas may have a rich endemic fauna. Nearly 300 species of amphipods occur in Lake Baikal in the Soviet Union, for example, and *Hyalella*, a genus of amphipods, shows the same sort of adaptive radiation in Lake Titicaca in South America, as do other species in other areas.

On tropical beaches some crabs are amphibious, emerging at ebb tide to feed; various land crabs and hermit crabs are more or less terrestrial, although their larval stages are marine. Best adapted of the terrestrial forms are the isopods; although the majority live in moist habitats, some can tolerate desert conditions.

Size range and diversity of structure. Crustaceans include some of the smallest and largest living arthropods. The adult water flea *Alonella* is less than 0.25 millimetre (about 0.01 inch) in length. The giant Japanese spider crab *Macrocheira*, on the other hand, sometimes exceeds 45 centimetres (18 inches) in shell width; its limbs extend 3.6 metres (12 feet) or more. The edible giant crab of Tasmania (*Pseudocarcinus*), with a shell width of 40 centimetres (16 inches), weighs as much as 13.6 kilograms (30 pounds), and the European edible crab (*Cancer pagurus*) may occasionally weigh 5.4 kilograms (12 pounds). The majority of crustaceans, however, are of small size, with an occasional large species in each order —*e.g.*, the deep-sea woodlouse, *Bathynomus giganteus*, 30 to 35 centimetres (12 to 14 inches) long; and *Pennella*, a copepod parasite of whales, 30 centimetres long.

The great diversity of crustaceans is reflected in the present classification, which includes eight subclasses and 36 orders (see below *Annotated classification*). Within many of the larger orders, marked diversity of form also occurs, depending on the different modes of life of the species. Certain parasitic copepods, cirripedes, and isopods are so extremely modified that the adults have lost most or all crustacean, and even arthropod, structure; the larvae, however, reveal the relationships.

Importance to man. The crustaceans of most obvious importance to man are the larger species, chiefly decapods. Important fisheries exist in many parts of the world for shrimps, prawns, spiny lobsters, and for the king crab (*Paralithodes*) of the north Pacific and its southern counterpart, the centolla, fished off the coast of Chile. Many species of true crabs—such as the blue crab, Dungeness crab, and the stone crab, all in North America, and the edible crab of Europe—are valuable sources of food. The most highly prized decapod is probably the true lobster (*Homarus* species), although overfishing early in the 20th century has greatly diminished the catches of both the North American and the European species. Freshwater crustaceans are of less economic importance. Included among them are crayfish and some river prawns and river crabs. Many species have only local market value. The mitten crab (*Eriocheir*), for example, although greatly esteemed as food in China, has been utilized primarily as animal fodder in Europe, where it was accidentally introduced and subsequently became a pest. Of the terrestrial decapods, some land crabs are eaten locally, and the robber crab, or coconut crab (*Birgus*), is valued on some Indo-Pacific islands. On other islands its use is forbidden on religious grounds. It is probable that no crustaceans are poisonous unless they have been feeding on the leaves and fruits of poisonous plants.

Edible crustaceans

The large acorn shell (*Balanus psittacus*), a barnacle measuring up to 27 centimetres (11 inches) in length, is regarded as a delicacy in South America, and a stalked barnacle (*Mitella pollicipes*) is eaten in parts of France and Spain. In Japan, barnacles are allowed to settle and grow on bamboo stakes, later to be scraped off and crushed for use as fertilizer. Planktonic (*i.e.*, drifting) copepods such as *Calanus* and euphausiids, or krill, may be present in such great numbers that they discolour large areas of the open sea, thus indicating to fishermen where shoals of herring and mackerel are likely to be found. The water flea (*Daphnia magna*) and the brine

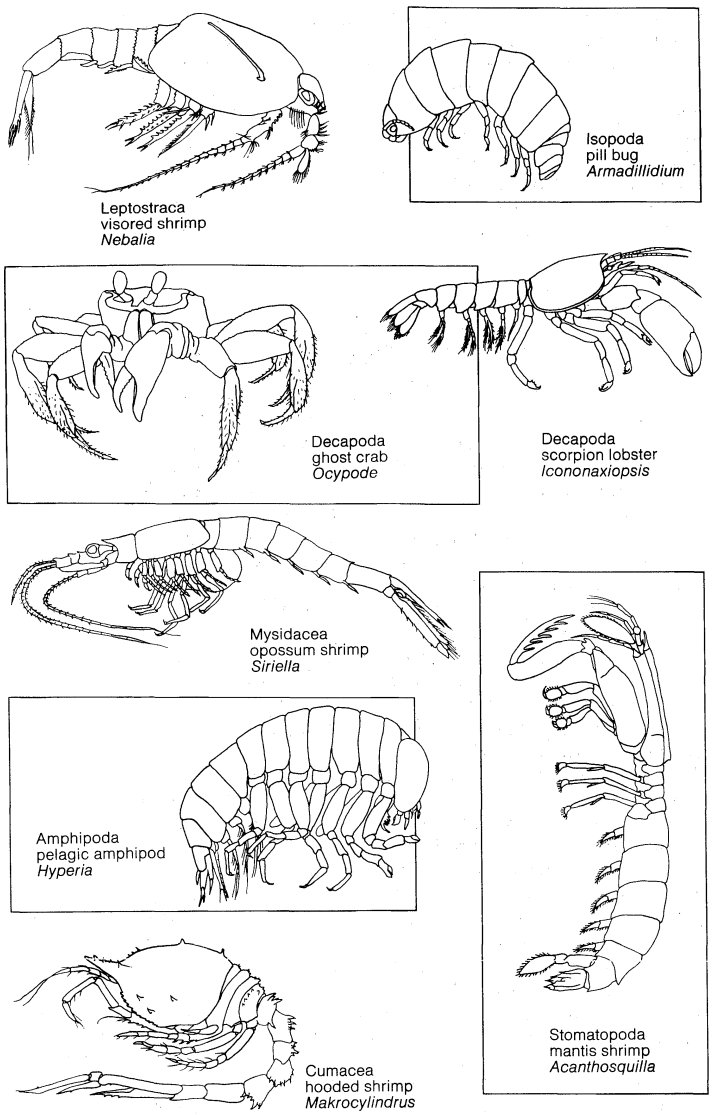


Figure 2: Body plans of representative crustaceans (leptostracan, Isopod, amphipod, stomatopod, mysidacean, cumacean, and decapods).

From (*Nebalia*) *Invertebrate Zoology* by Paul A. Meglitsch, Copyright © 1967 by Oxford University Press, Inc., Reprinted by permission; (*Armadillidium*) L.B. Holthuis, *Fauna Van Nederland;* (*Ocypode*) adapted from photograph by William J. Jahoda from National Audubon Society; (*Icononaxiopsis*) W.T. Calman, *A Treatise on Zoology* (1909), ed. Ray Lankester (copyright A. & C. Black Ltd., London; reprint Asher, Amsterdam, 1964); (*Siriella*) W.M. Tattersall and O.S. Tattersall, *The British Mysidacea* (1951), The Ray Society Publications; (*Hyperia*) F. Dahl, *Die Tierwelt Deutschlands,* Gustav Fischer Verlag, in A. Kaestner, *Invertebrate Zoology,* vol. III (1970), John Wiley & Sons, Inc.; (*Acanthosquilla*) *Invertebrate Identification Manual* by Richard A. Pimental © 1967 by Litton Educational Publishing, Inc., Reprinted by permission of Van Nostrand Reinhold Company; (*Makrocylindrus*) L. Fage, *Faune de France*

shrimp (*Artemia salina*) are used as fish food in aquariums and fish ponds, and the larvae of the latter are widely used as food during rearing in captivity of the larvae of larger crustaceans. Ostracods, of which numerous fossil and subfossil species are known, are of importance to geologists and oil prospectors.

Harmful and destructive crustaceans

Much damage may be done to rice paddies by burrowing crabs of various species and by the mud-eating, shrimplike *Thalassina* of Malaya. By undermining paddy embankments, they allow water to drain away, thus exposing the roots of the plants to the sun; if near the coast, salt water may thus be allowed to seep into the paddies. Tadpole shrimps (*Triops*) are often numerous in rice fields, where they stir up the fine silt in search of food, killing many of the plants. Land crabs and crayfish may damage tomato and cotton crops.

Barnacles play a dominant role in the fouling of ships' hulls. The speed of vessels with fouled hulls may be reduced by as much as 50 percent. The damage done to the submerged timbers of piers and jetties by the gribble (*Limnoria*), a small wood-boring isopod, is sometimes considerable. Even clay stone and reinforced cement may be extensively damaged by species of the isopod genus *Sphaeroma* if certain types of rock or calcium-containing materials are used in the cement. (I.G.)

NATURAL HISTORY

Reproduction and life cycle. *Sexual features.* In most crustaceans the sexes are separate; in the cirripedes and in some parasitic isopods, however, hermaphroditism is the rule, and isolated instances occur in other groups, especially among decapods. Parthenogenesis (*i.e.*, development of individuals from unfertilized eggs) is common in branchiopods and ostracods and occurs in at least one genus of terrestrial isopods.

In species in which the sexes are separate, sexual dimorphism is often striking. The males often have clasping (prehensile) organs for holding the female; they may be formed by modification of almost any appendage—antennules, antennae, thoracic limbs, or even some of the mouthparts. Some of the appendages in the region of the genital openings may be modified for the purpose of transferring the sperm to the female, as, for instance, the first and second abdominal appendages in the decapods.

In the higher decapods the male is often larger than the female, but in other groups the reverse is more frequently the case. In some parasitic copepods and isopods the minute male is attached, like a parasite, to the enormously larger female.

Hermaphroditism in cirripedes

In the Cirripedia some aberrant types of sexual relationship exist. Most cirripedes are hermaphrodites and are capable of both cross- and self-fertilization, but, in certain species, minute degenerate males exist and are attached to individuals of ordinary size. Since these dwarf males do not pair with females but with hermaphrodites, they have been called complemental males. In other species large individuals have become purely female by atrophy of the male organs and are entirely dependent on the dwarf males for fertilization.

Eggs. Most Crustacea carry the eggs after extrusion. They are retained between the valves of the carapace in some branchiopods and ostracods or within the mantle cavity in cirripedes. Among the malacostracans the Peracarida have a brood pouch, which is formed by overlapping plates attached to the bases of some of the thoracic legs. In most decapods the eggs are carried on the abdominal appendages of the female. In a few cases (Cladocera, terrestrial Isopoda) the developing embryos are nourished by a special secretion while in the brood chambers.

The majority of crustaceans hatch from the egg in a form differing more or less from that of the adult and pass through a series of free-swimming larval stages. In many instances, however, the newly hatched young resemble the parent in general structure.

Larvae. In Crustacea with the most complete series of larval stages, the earliest stage is the nauplius, which has only three pairs of appendages. The nauplius usually has an oval nonsegmented body. The three pairs of limbs correspond to the antennules, antennae, and mandibles (jaws) of the adult. The antennules are simple, the other limbs each have two branches, and all three pairs are used in swimming. The antennae and mandibles have a spinelike projection at the base and seize food and push it into the mouth. The mouth is overhung by a large upper lip, or labrum. The paired eyes are not yet evident, but the unpaired eye is usually conspicuous.

A nauplius larva differing only in details from that just described is found in most of the Branchiopoda, Cephalocarida, Mystacocarida, Copepoda, Cirripedia, and, in a more modified form, in some Ostracoda. Among the Malacostraca the nauplius is found in the Euphausiacea and some of the most primitive Decapoda. Many crustaceans that hatch at a later stage exhibit evidence of a nauplius stage in the embryonic development. It seems certain, therefore, that the possession of a nauplius larva must be regarded as a very primitive character.

As development proceeds, the body of the nauplius elongates, and its posterior part becomes segmented, new somites, or segments, being added at successive molts (periodic shedding of skin) from a zone in front of the telsonic, or tail, region. The appendages, which appear as

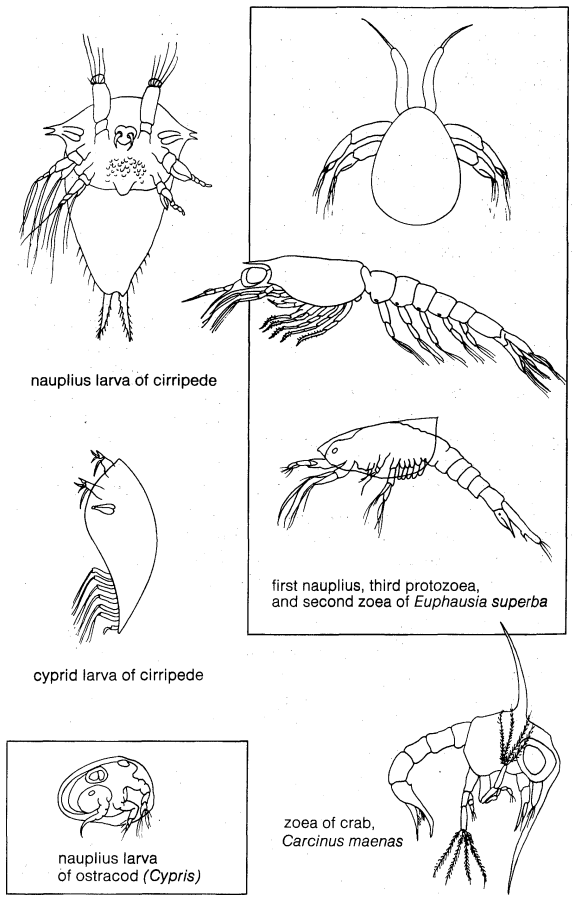


Figure 3: Crustacean larval types.
From (*Euphausia superba*) F.C. Fraser, "On the Development and Distribution of the Young Stages of Krill (Euphasia superba)," *Discovery Reports XIV* (1936), Cambridge University Press; (others) W.T. Calman, *A Treatise on Zoology* (1909), ed. Ray Lankester (copyright A. & C. Black Ltd., London; reprint Asher, Amsterdam, 1964)

buds on the ventral (belly) surface of the somites, become differentiated, like the somites that bear them, in regular order from front to rear. With the elongation of the body, its dorsal covering begins to project behind as a shell fold, the beginning of the carapace. The paired eyes appear under the cuticle at the sides of the head but become stalked at a comparatively late stage.

Because the course of development outlined above (in which the somites and appendages appear in regular order) is in good agreement with that in the Annelida (*e.g.*, earthworms), it is regarded as the most primitive.

In crustaceans other than some branchiopods and copepods the primitive scheme is modified. The earlier stages, for example, may be passed within the egg; thus the larva, on hatching, has reached a stage more advanced than the nauplius. In other modifications the gradual appearance of somites and appendages may be accelerated so that comparatively great advances take place at a single molt, or individual somites or pairs of appendages may appear in advance of their neighbours, disturbing the regular order of succession. The last type of modification is especially evident in the Malacostraca, in which it leads

Zoea

to the very peculiar larva known as the zoea. In the usual zoea, found in the true crabs, the posterior five or six thoracic somites are delayed in development, being represented by a short unsegmented region of the body at a stage when the abdominal somites behind them are fully formed and may even carry appendages.

Most larval forms swim freely at the surface of the sea and show adaptations to this pelagic habit of life. Various spines and other projections from the surface of the body often develop; they are probably less important as defensive organs than as aids to flotation. Also considered as aids to flotation are the greatly developed carapace of stomatopod larvae and the extreme flattening of the body in the membranous larvae of the spiny lobsters and their allies.

Complete suppression of metamorphosis (transformation between larva and adult) is found in the freshwater crayfish and the river crabs but is by no means universal among freshwater Crustacea. On the other hand, a few marine crabs are known to be hatched in a form differing little from that of the adult. (W.T.Ca./W.L.Sc./Ed.)

Ecology. Many crustaceans play important ecological roles. Amphipods and mysids, for example, form a large part of the food of commercial fishes. Even more important are the crustaceans adapted to graze on microscopic organisms such as diatoms and dinoflagellates. The copepod *Calanus* and the krill, or euphausiid shrimps, provide food for many kinds of fishes, seabirds, and even for the great baleen whales. There is also a host of copepod fish parasites, and many copepods, water fleas, and ostracods may serve as intermediate hosts for helminth parasites of vertebrates, including man. In areas where the lung fluke (*Paragonimus*) is prevalent, freshwater crayfish, prawns, and crabs, if eaten raw or imperfectly cooked, can transmit the parasite to man. In the tropics, a freshwater copepod (*Cyclops*) is an intermediate host of the guinea worm (*Filaria* or *Dracunculus*), and man can be infected by drinking unboiled water. Crayfish prey upon the molluscan carriers of *Bilharzia*, the blood fluke parasite of man.

FORM AND FUNCTION

General characteristics. Because crustaceans are essentially aquatic animals, the same appendages can perform several functions simultaneously. In the fairy shrimps (Anostraca), for example, in which no shell fold, or carapace, exists, many similar foliaceous (*i.e.*, flattened, leaflike) limbs function in locomotion and the gathering of food. The specialization of certain limbs for particular functions allows for adaptations such as reduction in their number or a shortening of the body. These modifications have occurred independently in most subclasses and in more than one way. The formation of a carapace, which protects and overhangs at least the anterior part of the body, facilitates the establishment of efficient respiratory currents. Although some groups (Cephalocarida, Syncarida) never possessed a carapace, others (amphipods, isopods) seem to have discarded it as incompatible with their mode of life.

Somites and appendages. As in all arthropods, the crustacean body consists of a series of segments, the somites or metameres, which may be free or more or less coalesced; each may bear a pair of jointed appendages. Body and limbs are encased in a continuous sheet of chitinous (tough, horny) cuticle, or integument, also known as the exoskeleton. Free somites are separated from each other by thinner areas of chitin, forming movable joints. Only in some small animals, such as *Bathynella* and harpacticid copepods, are the somites almost circular; usually they are more or less depressed or compressed. Typical arthropod structures such as the dorsal plate, or tergum, and the ventral plate, or sternum, may or may not be distinguishable; between them are the lateral areas, or pleura.

In addition to the true somites is a purely embryonic

Modified from Huxley and Siewing in A. Kaestner, *Invertebrate Zoology* (1970); John Wiley & Sons, Inc.

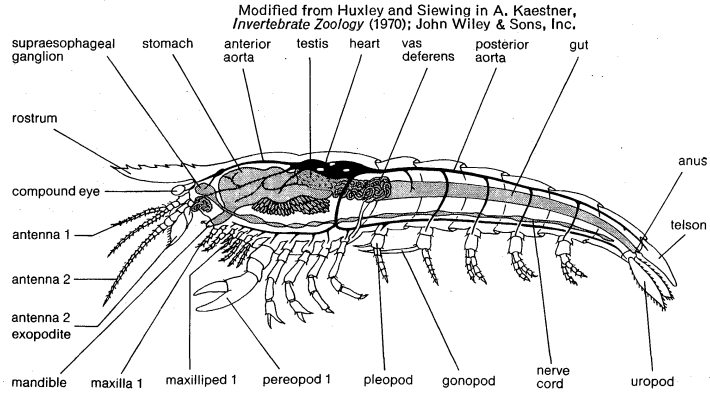


Figure 4: External and internal anatomy of a crustacean.

anterior presegmental part, the acron, on which are found the eyes (and perhaps the antennules). The posterior part, on which the anus opens, is not a true somite; rather, it is the posterior portion that remains after all the body somites have been budded off during embryonic development from the formative zone just in front of it; this telson, or tail, has neither nerves nor appendages; it often ends in two prongs, the caudal furca.

Formation of head

In no crustacean are all the body segments distinct and separate. At least four fuse with the acron to form the head, or cephalon. Of these, two segments are preoral (*i.e.*, are in front of the mouth) and carry feelers (antennules and antennae); the three segments bearing the mandibles, maxillules, and maxillae (all feeding or chewing appendages) are postoral. In species in which additional somites coalesce with the head to form a cephalothorax, the appendages (maxillipeds) usually show features of transition to those behind them. Since the shell fold, which forms the carapace, arises from the maxillary somite, only the first five are regarded as cephalic. The Mystacocarida are exceptional in that distinct traces of a more primitive cephalic segmentation are retained in both larva and adult. In adult mantis shrimps (Stomatopoda) two small movable pieces occur in front of the carapace; they bear the eyestalks and the antennules and are not present in the larva.

Throughout the subclass Malacostraca, the trunk is divided into two regions, or tagmata: a thorax of eight somites and an abdomen of six, rarely seven, somites. The female gonopores (openings that discharge eggs or sperm) are always on the sixth, the male ones on the eighth, thoracic somite. Often the sixth abdominal somite fuses with the telson to form a pleotelson; in some isopods other somites are included in the pleotelson. The trunk varies in the other subclasses from a few to more than 40 somites, or rings, and the limbs from one to as many as 71 pairs. The position of the gonopores also varies. The abdomen may be reduced to a mere stump, as in barnacles and the whale louse (*Cyamus*).

Carapace. The carapace, when present, varies greatly in size and shape. In primitive forms such as *Triops* (formerly *Apus*) and *Nebalia*, the shell fold remains free from the trunk, which it envelops to a large extent. It may form a bivalve shell enclosing body and limbs entirely, as in ostracods and clam shrimps (Conchostraca), or leaving the head free, as in water fleas (Cladocera). In the barnacles it is an enveloping fleshy "mantle" usually strengthened by shelly plates. In Malacostraca the carapace may fuse dorsally with some or all of the thoracic tergites, overhanging at the sides to protect the gill chambers. In this way the shell, or cephalothorax, of a lobster or crab is formed. Anteriorly, the carapace may form a beak, or rostrum.

External features. *Appendages.* Although the paired eyes of crustaceans are often borne on stalks, which may be movable and divided into two or three segments, they are not true limbs. The great diversity of forms assumed by crustacean appendages results from modifications of a fundamental type consisting of a peduncle, or stalk, the protopodite, bearing two branches, the exopodite and the endopodite. This simple form occurs in the cirri of barnacles and in the abdominal swimmerets of lobsters. The protopodite, of one, two, or rarely three segments, may have on its outer and inner margins additional lobes known as exites and endites. Thin-walled exites (epipodites) often function as simple or branched gills; the endites assist in transporting or chewing food material and are always present in the mouthparts. The cylindrical walking legs of many Malacostraca have become secondarily one-branched (uniramous) through loss of the exopodite in the adult; the series of segments belongs to both protopodite and endopodite. Leaflike thoracic limbs are found only in Leptostraca and Branchiopoda. The two-branched (biramous) limb occurs in the Leptostraca. In the fairy shrimp the protopodite and its well-developed endites and exites form most of the limb.

The protopodite

It has been suggested that all crustacean limbs can be derived from the generalized biramous limb of the Cephalocarida.

Specializations of the appendages. The antennules (first antennae) are uniramous in the nauplius larva and in the adults of all subclasses except the Malacostraca, in which they may have two or even three branches. Moreover, they are truly preoral in origin, being innervated from the brain or a preoral ganglion. In most cases they are sensory in function, but they may be used for swimming or, by the male, as claspers; in the cirripedes they form the organs of attachment to any solid surface.

The biramous antennae (second antennae) in the nauplius larva are situated at the sides of the mouth and may assist the mandibles in pushing food into the mouth. They become preoral in the course of development and normally retain both rami in the adult. They often serve as organs of locomotion, as claspers in the male, or as organs of attachment in parasites. In the Malacostraca the antennae are chiefly sensory; the endopodite is often a long lash, or flagellum; the exopodite either may be a flattened scale or may disappear.

The mandibles are biramous locomotor limbs in the nauplius of many crustaceans and may have masticatory, or chewing, lobes (gnathobases) on the protopodites. This form of mandible is retained in the adult mystacocarid and in some copepods and ostracods. In most cases, however, the exopodite is lost, and the endopodite, together with the distal part of the protopodite, forms a palp. The coxa, or first segment, of the protopodite with its gnathobase (endite) becomes a powerful jaw, often with distinct cutting and grinding processes. In bloodsucking parasites the mandibles are often piercing structures (stylets) enclosed in a tubular proboscis formed by fusion of the upper and lower lips.

The maxillules and the maxillae (first and second maxillae) tend to be flattened leaflike appendages. The first pair usually function in passing food to the mouth but may serve other needs; in males of the minute fossil *Lepidocaris* they formed claspers. The second pair perform various functions in connection with feeding and respiration.

Differentiation of limbs

The limbs behind the head region show little differentiation among themselves in the cephalocarids, most branchiopods, cirripedes, and many copepods. The ostracods, however, provide an excellent example of the plasticity and adaptability of crustacean limbs; the trunk is very short, and each of the five to seven pairs of limbs (including those of the head) may be specially adapted for specific functions in various groups. In the Malacostraca, the two series of eight thoracic and six abdominal appendages may be subdivided. The first one, two, or three pairs of thoracic appendages may be modified as additional mouthparts (maxillipeds). The endopodites of the unmodified pairs are often walking legs; the exopodites, sometimes natatory (used for swimming) or respiratory, often are absent. One or more pairs following the maxillipeds may end in pincers, either chelae or subchelae (in which the terminal segment folds back against the next one); in raptorial (hunting) animals at least one pair may be a large and powerful weapon. The abdominal limbs are primarily biramous swimming organs. When the last pair differ from the others, they are called uropods. In shrimps and lobsters the uropods form, together with the telson, a tail fan. In amphipods the first three pairs are natatory pleopods, and the last three pairs are uropods. Some or all pleopods may be absent or reduced, unless modified in the male for sperm transfer.

Integument. The exoskeleton, which protects the body and provides attachment for muscles, is secreted over the whole body surface by a single layer of cells, the epidermis (hypodermis). It varies from a relatively thin, flexible, undifferentiated sac, as in parasitic copepods and cirripedes, to a rigid and massive shell, as in crabs and lobsters. The exoskeleton is composed of two parts: an epicuticle without chitin (a tough, horny material) or wax; and a chitinous cuticle. The lining of the foregut and the hindgut and the membranes separating the segments of the body and limbs undergo little modification during growth. Elsewhere, the integument is often stiffened by the deposition of calcium salts. Hardening by tanning (sclerotization) may occur at the tips of walking

legs or the pincers. In Cephalocarida and Branchiopoda, in which little or no calcification or hardening occurs, the shape of body and limbs is maintained by internal fluid pressure; muscular movements take place against this pressure.

In decapods with a thick limy shell, the structure of the integument is very complex. The laminated cuticle consists of three layers; an outer pigmented layer, a median heavily calcified layer that may also have pigment, and an inner membranous layer. The integument contains ducts of dermal glands and extensions from the epidermis that fill hollow sensory structures called setae. Other modifications of the integument include corneal lenses of the eyes and, in some Malacostraca, the lenses of luminous (light producing) organs.

Internal skeleton

An internal skeleton is usually present in the form of ingrowths of the integument, known as apodemes, which serve as points for the insertion of muscles. In decapods, especially crabs, the apodemes unite to form an elaborate framework, the endophragmal skeleton.

All integumentary structures are cast off at each molt. Only the Conchostraca and the thoracican cirripedes are able to retain and build up portions of the exoskeleton of the carapace with calcium carbonate, while frequently molting the chitinous exoskeleton of the rest of the body.

Respiratory organs. In many of the smaller crustaceans, such as copepods, and *Lucifer* among the decapods, in which no special respiratory organs occur, gas exchange takes place through the entire thin integument. The inner wall of the carapace, facing the trunk, is often highly vascular and may in many groups be the only respiratory organ. Gills, when present, are formed by modifications of parts of appendages, most often the epipodites. These thin-walled, lamellate structures are present on some or all of the thoracic appendages in cephalocarids, fairy shrimps, and many Malacostraca. In mantis shrimps (Stomatopoda), gills are found on the exopodites of the pleopods. In euphausiids the single series of branched epipodial gills are fully exposed. In decapods the gills, protected by the overhanging carapace, are arranged in three series at or near the limb bases. As an adaptation to aerial respiration, the branchial chambers are greatly enlarged in certain land crabs and serve as lungs, the inner membrane being richly supplied with blood vessels. In isopods the respiratory function has been taken over by the abdominal appendages; either both rami or the endopodite become thin and flattened. Most sow bugs and pill bugs have, in addition, trachea-like infoldings in some of the exopodites.

Internal features. *Alimentary system.* As a rule the gut runs straight through the body, except anteriorly where it descends to the mouth. In a few cases it is slightly sinuous, or twisted; only rarely is it actually coiled upon itself. An upper lip (labrum) covers the mouth anteriorly; at its posterior border there may be a lower lip (labium), which may have a pair of lobes (paragnaths). Both foregut and hindgut have a chitinous lining—a continuation of the exoskeleton. The midgut varies greatly in length; its functions of digestion and absorption are increased by a varying number of tubular outgrowths (ceca). Commonly there is one anterior pair of ceca, which may be more or less ramified and which forms a massive digestive gland.

The foregut is provided with the strongest muscles. Teeth and hairs may be present in animals that eat large particles. Malacostraca, especially the decapods, have a more complex "stomach," with a gastric mill for grinding the food and a filtering apparatus consisting of bristles that strain particles. The mill and the filter are often found in separate anterior and posterior chambers. In some extremely specialized parasites, which live entirely at the expense of their hosts (Rhizocephala and monstrillid copepods), the gut is either vestigial or absent.

Circulatory system. As in other arthropods, the circulatory system of crustaceans is largely lacunar; *i.e.*, the blood flows in sinuses, or channels, without definite walls. Cirripedes and many ostracods and copepods have no heart, the blood being kept in motion by either a blood

Heart

pump or rhythmic movements of the body, gut, or appendages. When present, the heart lies in a blood sinus, or pericardium, with which it communicates by paired valvular openings, or ostia. In the more primitive crustaceans such as fairy shrimps or stomatopods, the heart is a long tube, with spiral muscles in its wall, and it extends almost the entire length of the trunk; there is a pair of ostia in each somite except the last. In more advanced crustaceans, however, shortening of the heart occurs, and the number of ostia may be reduced to three pairs or less. The position of the heart depends on that of the respiratory organs; it usually lies in the thorax or cephalothorax but is mainly in the abdomen of isopods. Malacostraca have a well-developed system of elastic-walled arteries including an anterior and usually a posterior aorta.

The red pigment hemoglobin has been observed in the blood of the Branchiopoda and in other subclasses except Malacostraca. Hemocyanin is the respiratory pigment (*i.e.*, the pigment that carries the oxygen) in decapods and stomatopods.

Excretory system. The principal excretory organs—two pairs of glands in the antennal and maxillary somites—open on or near the bases of the appendages. Rarely are both pairs present and functional in adults (lophogastrid mysids), although one may replace the other in the course of development. In Branchiopoda the antennal gland is functional in the larva, the maxillary one in the adult. In decapods the reverse may be true. Each gland has an end sac, representing a vestige of the coelom, and a convoluted duct, which may expand into a bladder before opening to the exterior. The antennal, or green, gland of decapods is usually very complicated, and prolongations from the bladder may ramify through the body.

The glands are not concerned particularly with nitrogen excretion or osmoregulation—*i.e.*, the balance of salts—but are instrumental in maintaining the ionic balance of the body fluids; conserving calcium, potassium, and glucose; and selectively eliminating magnesium and sulfates. Waste products such as ammonia and urea are eliminated through the gill epithelium, and other end products such as urea are stored in special cells (nephrocytes) in the gills and leg bases.

Nervous system. The nervous system varies from the simple ladderlike arrangement found in some annelids to great complexity. In the primitive Branchiopoda a simple "brain," or dorsal, preoral ganglionic mass, has nerve connections with the eyes and the antennules. The antennal ganglia are postoral, and the antennal nerves are situated on the connective ring around the gullet. The two halves of a ventral nerve chain are widely separated, and a pair of ganglia, united by double transverse bands, occurs in each limb-bearing somite. In higher groups the antennal ganglia are coalesced with the brain, and in many decapods additional centres develop and the brain is very complex. A system of nerves and ganglia associated with the heart and alimentary canal is well developed in decapods. Giant fibres are especially well developed in the central nervous system of shrimps and crayfish, which, by a sudden flexure of the abdomen with tail fan expanded, can travel backward rapidly. Many crustaceans exhibit this escape mechanism.

Functional features. *Sense organs.* The sense organs of crustaceans include visual, tactile, olfactory, and balancing organs. The eyes are of two kinds: an unpaired median, or nauplius, eye and paired compound eyes. Only the median eye is present in the earliest larva; it consists of a small group of three or four simple eyes (ocelli) innervated from the forebrain. Sometimes it persists as the only visual organ in the adult and may occasionally have conspicuous lenses as in *Sapphirina* and related copepods. Compound eyes, when present, appear at a later stage in development, at which time the median eye may degenerate or disappear. The compound eyes, which may be stalked or not stalked (sessile), are very similar in detailed structure to those of insects. The visual elements, or ommatidia, which vary in number from a few to 14,000, are separated by pigment sheaths. Their so-called crystalline cones are covered by a transparent cornea, which is usually divided into lenslike facets corresponding to the

Types of eyes

underlying ommatidia. Visual elements are reduced or lacking in cave-dwelling crustaceans, although the eyestalks may persist, and many small crustaceans, such as cephalocarids and mystacocarids, have no eyes.

Other sense organs are formed by modifications of the hairs (setae) that occur on the body and appendages. Rather stiff hairs respond to mechanical movement and are tactile. More complicated structures, such as the hair peg organs of decapods, sense the direction and speed of water movements over the body; still other structures detect distant disturbances in the water (*e.g.*, the approach of another animal). Special olfactory hairs, or esthetases, are found chiefly on the antennules, sometimes on the antennae, and may be better developed in the male than in the female. Other setae on mouthparts, pincers, and legs can detect food. Statocysts, found only in some Malacostraca, are for orientation with respect to gravity. These paired organs are situated at the base of the antennules in many decapods, on the inner branches of the uropods in most mysids, or on the dorsal surface of the telson in a few isopods. Each is an open or closed pit, with sensory hairs on the inner surface and with one or more statoliths, either sand grains introduced from the exterior after each molt or limy concretions formed within the pit. Various crustaceans can produce sound by rubbing together body parts (stridulation), but no definite auditory organs have as yet been detected.

Dermal glands. Secretions from dermal glands on various parts of the body serve a variety of purposes: to entrap food particles before they are swallowed, as in some branchiopods; to form a gelatinous capsule around the body to resist desiccation, as in some copepods; to form during oviposition, or egg laying, a protective chamber around the eggs until all are firmly glued to setae on the abdominal pleopods, as in crayfish; to cement the animal to a surface, as in cirripedes; and to cement together sand grains or algal filaments in the construction of burrows, as in some tanaids and amphipods. Cytherid ostracods are assisted in climbing over smooth surfaces by a thread, along which they can return, spider-fashion, if necessary. The silk, which is secreted by shell glands within the carapace, passes through the modified exopods of the antennae.

Luminous organs. Minute ostracods and copepods are responsible for phosphorescence of the sea—*i.e.*, a phenomenon in which the sea seems to give off light. Among the Malacostraca, most euphausiids and a few mysids and decapod shrimps also become luminous under excitation. The light-producing organs vary greatly in number, arrangement, and structure. In the open type, secretions that become luminous upon contact with seawater are produced by special dermal glands that are scattered over the body surface (as in copepods) or localized near the mouth (as in ostracods, the mysid *Gnathophausia*, and some decapods). Other decapods and the euphausiids have light-producing organs of the closed type, called photophores. They are either complex innervated lights—with reflector, condensing lens, and muscles for directing the beam—or simple linear streaks without a lens. In sergestid shrimps some tubules of the midgut ceca (organs of Pesta) are luminous.

The biological significance of luminescence in crustaceans is unknown, but it might serve for species recognition; to attract prey; to escape from predators under cover of the luminous cloud; to act as countershading, in which case light would have to be produced at all times (see also BIOLUMINESCENCE).

Hormones and endocrine glands. Hormones are blood-borne chemical agents that act on cells at a point distant from the site of formation. Two categories of tissues are involved: specialized neurosecretory cells of the central nervous system and tissues of non-nervous origin. Of the former, the most important is the X-organ–sinus-gland complex, situated in the eyestalks of many decapods. Only three non-nervous tissues are known to produce hormones: the ovaries of the female, the androgenic glands of the male, and a pair of ventral glands, the Y-organs, located in the maxillary somite in decapods. The hormones of these glands act directly on other tissues of the body, but their activities are controlled by some of the neurohormones produced by the X-organ–sinus-gland complex. This complex also regulates maturation of the gonads, dispersal of pigments and colour change, and certain metabolic processes.

Hormonal control of maturation of the gonads and development of the secondary sexual characters has been studied only in some decapods, amphipods, and isopods. It is probable that the sex-determining genes determine only the formation or suppression of the androgenic glands. If the glands develop, the young animal becomes a male; if the androgenic glands are suppressed, the animal becomes a female.

Molting. A firm exoskeleton places severe restrictions on growth; hence, it must be cast off periodically. Unlike most insects, many crustaceans continue to grow long after sexual maturity is reached; the cyclical nature of crustacean molting is apparent and gives crustacean physiology its unusual and distinctive features. Studies on young and adult decapods have revealed that, far from being a brief interruption of the animal's normal existence, molting profoundly affects most or all of its life. Although molting—exuviation or ecdysis—is of very short duration, tissue growth and changes in the integument occur during 70 percent of the intermolt period; during the other 30 percent the animal stores metabolic reserves in preparation for the next molt.

Stages of the molting cycle

The four stages of the molting cycle are: premolt (pre-ecdysis), characterized by thinning and softening of the old integument, with resorption of most of its calcium salts when present, and formation of some layers of the new exoskeleton underneath; molt (ecdysis), with rapid intake of water and shedding of skin; postmolt (metecdysis), a period of recovery when the new integument is soft but rapidly hardening; and intermolt, either diecdysis in species that molt throughout the year or anecdysis when molting is seasonal. Sometimes, as in the spider crab *Maia*, the molt of puberty is the terminal one, and anecdysis is permanent. Throughout most of the intermolt cycle, certain neurosecretory cells of the X-organs secrete a molt-inhibiting hormone that is stored in, and released by, the sinus glands. A fall in the concentration of this hormone stimulates the Y-organs to form and secrete their molting hormone. The hormone initiates and integrates the metabolic events leading to ecdysis; once pre-ecdysis has been initiated, the cycle must proceed. When terminal anecdysis is reached, either the Y-organs atrophy (*Maia*) or a high concentration of molt-inhibiting hormone must be maintained. Only in the latter case will removal of the eyestalks (*Carcinus*) result in continued molting and growth. A few crustaceans such as the lobster or the edible crab (*Cancer pagurus*) continue to molt throughout life, though at longer intervals; thus they may attain very large sizes. (I.G.)

PALEONTOLOGY AND CLASSIFICATION

Paleontology. Although fossil remains of Crustacea are abundant in rock strata belonging to all the main divisions of the geological time scale from the most ancient up to the most recent, they disclose little regarding the phylogeny of the class. This results in part from the fact that many important forms must have escaped fossilization altogether because of their small size and delicate structure. Many preserved forms are known only from the carapace, or shell, the limbs being absent or represented only by fragments. The accident that preserved the minute branchiopod *Lepidocaris* in the Old Red Sandstone of northwestern Europe is not likely to have been often repeated. *Lepidocaris*, however, is of recent date as compared with the varied fauna of Crustacea discovered in the Middle Cambrian (about 535,000,000 years ago) layers of the Canadian Rockies.

There is reason to believe that many of the chief groups were already differentiated before the beginning of the geological record. Shrimplike forms that can be definitely referred to the Malacostraca begin to appear in the Upper Devonian (370,000,000 to 345,000,000 years ago). Syncarida and Hoplocarida can be recognized in the Carboniferous (345,000,000 to 280,000,000 years ago), and

true decapods appear in the Permian (280,000,00 to 225,000,000 years ago) and the Triassic (225,000,000 to 190,000,000 years ago).

In the dearth of evidence from paleontology, data afforded by comparative anatomy and embryology must be employed to reconstruct the course of evolution within the class. Conclusions reached in this way must of course remain more or less speculative so long as they cannot be checked by the results of paleontology.

Earlier attempts to reconstruct the genealogical history of the Crustacea were based on the theory of recapitulation, in this instance the assumption that the successive stages of the larval history, especially the nauplius and zoea, reproduced the actual structure of ancestral types. It is now generally agreed that this theory cannot be applied to the zoea, the characters of which must result from secondary modification. As regards the nauplius, however, the constancy of its general structure in the most diverse groups of Crustacea strongly suggests that it is a very ancient type, and the view has been strongly advocated that the Crustacea must have arisen from an unsegmented nauplius-like ancestor. Resemblances between the more primitive Crustacea and annelid worms, in such characters as the structure of the nervous system and the mode of growth of the somites, can hardly be ignored, however, and it is reasonable to suppose that the Crustacea originated from some stock that already possessed these characters.

Origin of crustaceans

The hypothetical ancestral crustacean has long been thought to have resembled, in general form, a branchiopod such as *Triops* (*Apus*), with an elongated body composed of numerous similar somites and ending in a caudal fork; with a carapace originating as a shell fold from the cephalic region; with eyes probably stalked; and with antennae and trunk limbs both biramous. More recently this claim has been made for the cephalocarids, of which the first described (1955) was *Hutchinsoniella*. Though similar in body form, cephalocarids are eyeless and have a much less well developed carapacial fold; in addition, the body ends in caudal rami instead of a caudal fork.

(W.T.Ca./W.L.Sc./Ed.)

Classification. *Distinguishing taxonomic features.* In classifying the Crustacea, taxonomists rely on a variety of characters: the form and extent of the carapace, if present; the number of trunk somites, or segments, and how many fuse with the head or with the telson; the number and degree of specialization of the trunk limbs; the presence or absence of paired eyes and of a caudal furca—*i.e.*, a forked-tail process; the position and kind of respiratory organs. The position of the genital openings, the mode of attachment of the eggs to the female, and the stage at which the first larva hatches may also be important. Parasitic and sedentary forms may differ markedly as adults from free-living species.

Annotated classification. The following classification of the Crustacea is that adopted in most modern textbooks —*e.g.*, those of T.H. Waterman (*The Physiology of Crustacea*, 1960–61) and A. Kaestner (*Invertebrate Zoology*, vol. 3, 1970). Categories represented only by fossils are inserted according to R.C. Moore (*Treatise on Invertebrate Paleontology*, 1969). Groups marked with a dagger (†) are extinct and known only from fossils.

CLASS CRUSTACEA

About 35,000 to 40,000 species known; mostly of aquatic habit, with 2 pairs of feelers in front of mouth; at least 3 pairs of appendages behind mouth, acting as jaws.

Subclass Cephalocarida

Recent; primitive blind shrimps, with head shield but no carapace; maxilla and all thoracic limbs alike; 11 abdominal somites, telson with furca; 2–4 mm; larva a nauplius; found on coasts of North America, Japan, and New Caledonia, intertidal to 300 m. Few known species.

Subclass Branchiopoda

Lower Devonian (395,000,000 to 375,000,000 years ago) to present; predominantly freshwater, all continents. Compound eyes usually sessile (*i.e.*, unstalked) and close together or coalesced; trunk limbs usually foliaceous (*i.e.*, leaflike); larva usually a nauplius; approximately 900 species. See BRANCHIOPODA.

Order Anostraca (fairy shrimp, brine shrimp). Lower Devonian to present. Carapace absent, eyes stalked; trunk limbs 11 to 19 pairs, all anterior to gonopores; antennae modified as claspers in male; furcal rami (branches) short. 5–100 mm.

†*Order Lipostraca.* Middle Devonian; Scotland, freshwater. One species, *Lepidocaris*, a minute blind form related to Anostraca.

Order Notostraca (tadpole shrimps). Carboniferous (345,000,000 to 280,000,000 years ago) to present. Carapace a broad shield above trunk; 35 to 71 trunk limbs of which 29 to 52 are posterior to the gonopores; furcal rami long and multi-articulate. 70–100 mm long. Two genera.

†*Order Kazacharthra.* Lias, Kazakh S.S.R.

†*Order Acerostraca.* Lower Devonian, Germany.

Order Conchostraca (clam shrimps). Lower Devonian to present. Bivalve carapace with hinge and adductor muscle, enclosing body; 10 to 32 trunk limbs, 1 to 16 postgenital; furcal rami clawlike. 6–17 mm long.

Order Cladocera (water fleas). Oligocene (38,000,000 to 26,000,000 years ago) to present. Carapace usually enclosing trunk, leaving head free; 4 to 6 pairs of trunk limbs; furcal rami clawlike; development direct, except in *Leptodora.*

Subclass Ostracoda (mussel, or seed, shrimps)

Lower Cambrian (570,000,000 to 550,000,000 years ago) to present; worldwide, marine, freshwater, and a few terrestrial species. Bivalve shell enclosing body and limbs; body short, unsegmented, never more than 2 pairs of trunk limbs; furca present or absent. Larva a nauplius with precociously formed shell; mostly 0.14–8 mm long; more than 2,000 living species.

Order Myodocopa. Marine. Antennal notch in shell; 2 pairs of trunk limbs; 1.8–8 mm long.

Order Cladocopa. Marine. No antennal notch in shell; no trunk limbs; 0.14–0.75 mm long.

Order Podocopa. Marine, freshwater, and terrestrial. Second trunk limb for walking or grooming; furca long and pointed or absent; 0.23–3 mm long.

Order Platycopa. Marine. One pair trunk limbs; furca leaflike; 0.5–1 mm long.

†*Order Archaeocopida.* Cambrian and possibly Early Ordovician (about 450,000,000 to 500,000,000 years ago). North America and Europe.

†*Order Leperditicopida.* Ordovician to Devonian. Europe and North America.

†*Order Palaeocopida.* Ordovician to Middle Permian (about 250,000,000 years ago). North America and Europe.

Subclass Mystacocarida

North and South America, Mediterranean, South Africa; intertidal beaches to 25 m. Blind interstitial forms; head retaining traces of segmentation; trunk of 10 somites, telson with furca; 0.5 mm long; larva a nauplius. Few species known.

†**Subclass Euthycarcinoidea**

Triassic. Freshwater, France and Australia; 5–65 mm long.

Subclass Copepoda

Miocene (26,000,000 to 7,000,000 years ago) to present, worldwide distribution, marine and freshwater; 10,000 species. Free-living and parasitic; without carapace or compound eyes; 1 or more trunk segments fused to head; typically 6 pairs of thoracic limbs, none on abdomen; telson with furca. Eggs carried in 1 or 2 egg strings; larva usually a nauplius.

Order Calanoida. Head and thorax wide, abdomen narrow, with major articulation between them; antennules long, one often modified in male. Usually pelagic. 6 mm long or less.

Order Harpacticoida. Usually benthic (*i.e.*, living on the bottom). Body generally elongated, articulation between thoracic somites 5 and 6, the latter joined to genital somite; short antennules, both modified in male. Less than 2 mm long.

Order Cyclopoida. Mostly free-living. Pear-shaped body with articulation between thoracic somites 5 and 6; antennules often modified in male; 0.6–5 mm long.

Order Notodelphoida. Marine, associated with ascidians (*e.g.*, sea squirts). Body cyclopoid or wormlike; sometimes with dorsal brood pouch. Female 2–12 mm long.

Order Monstrilloida. Adults planktonic, immature stages in polychaetes, ophiuroids, or gastropods. No mouthparts or gut.

Order Caligoida. Parasitic on marine and freshwater fishes. Body flattened; mouthparts suctorial. Female usually 2.6–30 mm long.

Order Lernaeopodoida. Parasitic on fishes. Maxillae modified for attachment to host. Larva a first copepodid. Female to 32 mm long.

Subclass Branchiura

Order Arguloida (fish lice). Temporary fish parasites with wide flat carapace, unsegmented abdomen, compound eyes and 4 pairs of thoracic limbs. Eggs deposited on substrate. 10–30 mm long; 120 species.

Subclass Cirripedia (barnacles)

Upper Silurian (420,000,000 to 395,000,000 years ago) to present, worldwide, marine to brackish water. Sedentary or parasitic; head usually reduced and abdomen lost. Carapace usually an enveloping mantle; 6 pairs of trunk limbs (cirri); nauplius with a pair of lateral horns. Usually hermaphroditic; 900 species. See CIRRIPEDIA.

Order Ascothoracica. Cretaceous (65,000,000 to 136,000,000 years ago) to present, marine ectoparasites or endoparasites of corals and echinoderms. Some species have long 4-segmented abdomen, telson, and furca. Sexes separate. Mantle bivalved. Females 3–12 mm long; about 30 species.

Order Thoracica (acorn, wart, and goose barnacles). Upper Silurian to present. Six pairs of cirri arranged evenly along thorax; permanently attached by preoral region, which may form a long stalk; mantle usually covered by permanent limy plates; stalked forms, 1–80 cm long; about 650 species.

Order Acrothoracica. Carboniferous to present. Small cirripedes that bore into coral, limestone, or mollusk shells. First pair of cirri near mouth, remaining pairs at posterior end of thorax; no limy plates. Sexes separate. About 30 species.

Order Rhizocephala. Parasitic on decapod crustaceans chiefly. No gut or alimentary canal at any time. Larva typical cirripede nauplius. About 200 species.

Subclass Malacostraca

Lower Cambrian to present. Typically with compound eyes, stalked or sessile; a carapace covering the cephalothorax; 8 thoracic and 6 abdominal somites, each with paired appendages; about 19,000 species.

Superorder Phyllocarida

Lower Cambrian to present.

†*Order Hymenostraca.* Cambrian to Ordovician.

†*Order Archaeostraca.* Ordovician to Triassic. To 75 cm in length.

Order Leptostraca. Upper Permian to present, marine. Small shrimplike forms with bivalve carapace, movable rostrum; a 7th abdominal somite without appendages; and furcal rami; 6–40 mm; about 10 species.

†*Superorder Eocarida*

Mid-Devonian to Permian; 2 orders.

Superorder Syncarida

Carboniferous to present. Primitive, without a carapace; eyes stalked, sessile (without a stalk), or absent; 5 or more pairs of thoracic limbs; more than 70 species.

†*Order Palaeocaridacea.* Carboniferous to Permian.

Order Anaspidacea. Triassic to present, freshwater, Tasmania and Australia. Eyes stalked or sessile; furca absent; first 2 pairs of pleopods, or swimmerets, of male copulatory, seminal receptacle in female; 8–50 mm long.

Order Bathynellacea. Freshwater, all continents except North America. Minute, blind, wormlike forms with 6th abdominal somite fused to telson, with furca; pleopods reduced or absent; 0.55–5.4 mm long. More than 60 species.

Order Stygocaridacea. Permian to present, freshwater, South America and New Zealand. Minute, blind, wormlike; head fused with first thoracic somite; 1.6–4.2 mm long.

Superorder Hoplocarida

Carboniferous to present.

Order Stomatopoda (mantis shrimps). Jurassic to present, marine, warmer coastal waters. Short, shallow carapace; abdomen well developed, 1.2–35 cm long. About 250 species.

†*Order Palaeostomatopoda.* Carboniferous. Subchelate limbs equal; telson with median spine and furca. *Perimecturus.*

†*Order Aeschronectidae.* Upper Carboniferous.

Superorder Peracarida

Permian to present. Carapace, when present, rarely fuses with more than 4 thoracic somites; a lacinia mobilis on mandible of adult; typically oostegites in female. Development direct. About 10,300 species.

Order Thermosbaenacea. Recent. Minute interstitial or subterranean forms, Tunisia, Italy, Dalmatia, Dead Sea, Texas. Carapace forms a temporary dorsal brood pouch; 4 mm long. About 6 species.

Order Mysidacea (opossum shrimps). Triassic to present; worldwide; marine, brackish, and freshwater. Shrimplike, with large carapace, stalked eyes, antennal scale, natatory thoracic expodites. 3–30 mm long, rarely to 180 mm long. More than 625 species.

Order Cumacea. Permian to present, mainly marine, all oceans, littoral to great depths. Easily recognized by the inflated head and thorax, long slender abdomen and uropods. 1.1–35 mm. About 770 species.

Order Spelaeogriphacea. Present. A small, blind shrimp from stream in Bats Cave, Table Mountain, South Africa. Carapace short, telson free, uropods long. To 7.5 mm long; 1 species, *Spelaeogriphus.*

Order Tanaidacea. Permian to present, mainly marine, coastal to 8,200 m. Carapace short, fused to first 2 thoracic somites, abdomen short; 2nd thoracic limb chelate; 2–25 mm long. More than 300 species.

Order Isopoda (pill bugs, sow bugs). Triassic to present, worldwide, marine (free or parasitic), freshwater and terrestrial. Body usually flattened dorsoventrally; carapace absent; eyes sessile; thoracic limbs 2 to 8 cylindrical legs; abdomen short, 1 or more somites fused with telson; pleopods specialized for respiration. Usually 0.7 to 35 mm long. More than 4,000 species.

Order Amphipoda (beach hoppers, scuds, well shrimps). Upper Eocene to present. Worldwide, marine, freshwater and semiterrestrial. Typical members recognized by laterally compressed body; lack of carapace; sessile eyes; long thoracic limbs, some with platelike gills; 1–50 mm, rarely to 140 mm; about 4,600 species.

Superorder Eucarida

Carapace large, fused dorsally to all thoracic tergites; eyes stalked; development rarely direct.

Order Euphausiacea (krill). Marine, all oceans, pelagic. Shrimplike, carapace shallow, leaving the single row of branched gills exposed. Eggs usually shed freely, 1st larva a nauplius; 6–81 mm long. About 90 species.

Order Decapoda (Shrimps, lobsters, crabs). Permian or Triassic to present. Cosmopolitan, mainly marine, some freshwater and semiterrestrial. Carapace deep, enclosing the gill chambers. Abdomen varies from long and extended to short and tucked under cephalothorax. Includes the largest living arthropods; about 8,500 species. See DECAPODA.

Critical appraisal. The relationships of the various subclasses to each other are still largely conjectural. The close affinities of the Ascothoracica to the cirripedes as a whole are generally recognized, and there seems no valid reason for separating them as a subclass Ascothoracida. They retain the bivalve shell of the cyprid larva, and the positions of male and female gonopores are identical with those of Cirripedia. Some ascothoracicans have a thorax of six, and an abdomen of four, segments plus a telson and furca. This body plan they share with the Copepoda, and from it can be derived the other Cirripedia, the Mystacocarida, and the Branchiura. The term Maxillopoda has recently been proposed for these four subclasses; and perhaps the Ostracoda diverged from this line. The four recent orders of Ostracoda listed above may be regarded as suborders, arranged in the orders Myodocopida and Podocopida.

Opinions differ markedly with regard to the recent orders of the subclass Branchiopoda. Sometimes the Anostraca are separated from the other three, which are placed in the superorder Phyllopoda, a term that is sometimes used for the three non-cladoceran orders. Or again, the Conchostraca and Cladocera may be regarded as suborders of the order Diplostraca. The superorder Pancarida has been proposed for the order Thermosbaenacea, but some specialists regard them as aberrant Peracarida.

BIBLIOGRAPHY. W.T. CALMAN, "Crustacea," in RAY LANKESTER (ed.), *A Treatise on Zoology*, pt. 7 (1909, reprinted 1964), a classic English textbook, with earlier references; A. KAESTNER, *Invertebrate Zoology*, vol. 3 (1970), an up-to-date English adaptation of vol. 2 of the *Lehrbuch der Speziellen Zoologie*, 2nd ed. (1967), an excellent survey of morphology physiology, embryology, and ecology—systematic part extended to world fauna, with special emphasis on North American and European species; J. GREEN, *A Biology of Crustacea* (1961); A.P.M. LOCKWOOD, *Aspects of the Physiology of Crustacea* (1967), for undergraduates and as background reading for postgraduates; R.C. MOORE (ed.), *Treatise on Invertebrate Palaeontology*, pt. Q, *Arthropoda 3, Crustacea, Ostracoda*

(1961), and pt. R, *Arthropoda 4, Crustacea Exclusive of Ostracoda, Myriapoda, Hexapoda* (1969), includes information on recent crustaceans and on the evolution of the Arthropoda of interest to neontologists, although the emphasis is on the systematics of fossil crustaceans; W.L. SCHMITT, *Crustaceans* (1931, reprinted 1965), an excellent and most readable introduction for amateurs and students, by a specialist with first-hand field experience; T.H. WATERMAN (ed.), *The Physiology of Crustacea*, 2 vol. (1960–61), authoritative reviews of various aspects of physiology, each with an extensive bibliography; H.B. WHITTINGTON and W.D.I. ROLFE (eds.) *Phylogeny and Evolution of Crustacea* (1963).

(I.G.)

Cryogenics, Applications of

Cryogenics deals with the production of low temperatures and the utilization of low-temperature phenomena. The name is derived from the Greek *kryo*, meaning "icy cold" or "frost."

The cryogenic temperature range, in the sense of the term most often used, has been arbitrarily defined as extending from an approximate upper limit of −150° C (−238° F) down to absolute zero (−273° C, 0° Kelvin).

Cryogenic engineering deals with the many processes and devices utilizing these low temperatures. Examples of these include electrically superconductive devices and procedures for handling and processing of the fluids that have normal boiling points in this region.

Cryogenic temperatures customarily are listed on the absolute scale, in degrees Kelvin (°K). On this scale the absolute zero of temperature is at 0° K, corresponding to −460° F, or −273° C. Conversion from the Kelvin scale (°K) to the Celsius scale (°C) is accomplished by subtracting 273 from the Kelvin value.

The related study of low-temperature physics commonly involves physical research conducted at temperatures below the boiling point of oxygen; *i.e.*, below about 90° K.

Cryogenic temperatures are considerably below those encountered in common refrigeration processes. At these temperatures such physical and chemical properties of materials as strength, ductility, electrical resistance, and thermal conductivity are greatly altered from ambient conditions generally regarded as about 300° K. In finding applications for these changed properties, cryogenic engineering has become both a special discipline and a fast-growing industry.

Although cryogenic engineers have played major roles in the life-support and propulsion systems of missile and space programs, the utilization of cryogenics in the 1970s extends far beyond aerospace technology.

HISTORY

Cryogenics is generally considered to have been born in 1877 with the announcement that oxygen had been liquefied in small quantities at a boiling point of −183° C.

Other milestones in the development of cryogenics include the invention of the vacuum-jacketed glass flask in 1888, the liquefaction and separation of air into its components in 1895, the liquefaction of helium in 1908, and the discovery of superconductivity in 1911. In the 1920s and 1930s techniques were developed that led to the attainment of a temperature very close to absolute zero.

Although Goddard successfully used liquid oxygen to power rocket flight in 1926, it was Germany's use of liquid oxygen to propel V-2 rockets during World War II that awakened the world to the vast possibilities of cryogenics for space programs.

By 1947 the first commercially produced system for liquefying helium, at 4.2° K, was developed, and by 1960 cryogenic engineering had progressed to the point at which a temperature of $\frac{1}{1,000,000}$° K was achieved.

Cryogenic engineering has broadened considerably, extending also to medicine and food preservation and many other fields during the past 10 years.

THE PROPERTIES, PRODUCTION, AND STORAGE OF CRYOGENS

Liquid oxygen reacts violently under certain conditions with aluminum, titanium, and steel. Liquid fluorine is even more reactive than oxygen. Liquid nitrogen and liquid helium are clear, nonflammable fluids; helium is inert and nitrogen relatively inert; liquid hydrogen is colourless, has about $\frac{1}{14}$ the density of water, and can be violently reactive. The helium isotope helium-4 is most unusual; it is a normal fluid down to 2.17° K, the lambda point, but below that temperature it becomes the superfluid helium II, with properties exhibited by no other fluid.

Cryogenic fluids

At cryogenic temperatures the magnetic, thermal, and electric properties of most substances also are greatly altered. Some metals become perfect conductors of electricity; that is, they become superconducting. The behaviour at cryogenic temperatures of stainless steels, aluminum, copper, and nickel generally includes increases in tensile strength, yield strength, fatigue strength, and hardness, while impact strength decreases. These and other still-unexplored changes in materials promise many further applications of cryogenics in the fields of chemistry, physics, space technology, and medicine.

PRODUCTION AND STORAGE OF CRYOGENIC FLUIDS

Gases used in cryogenic engineering are cooled to their boiling—or liquefying—points by three basic methods: liquid expansion, Joule-Thomson expansion, and expansion in an engine (see REFRIGERATION EQUIPMENT). Cooling below 1° K is generally required only for basic research on the properties of matter; a particle accelerator at Stanford University, California, uses refrigerators working at 1.8° K. Such systems are very large and have extremely high power requirements.

After production cryogenic liquids generally are stored in specially designed tanks using superinsulation or in Dewar vessels; *i.e.*, flasks with double walls of silvered glass or highly polished metal having an evacuated space between them, very similar except for size to the common thermos bottle. Liquid air, oxygen, nitrogen, and even hydrogen can be kept for several hours in such vessels without further thermal protection; liquid helium has such a low heat of vaporization that it can be kept for any length of time only if the Dewar vessel is in turn surrounded by a similar, larger flask containing liquid nitrogen or liquid air. It can be stored for long periods in superinsulated vessels.

Dewar vessels

Liquefying natural gas reduces its volume more than 600 times, offering obvious advantages in transportation and storage. Liquid natural gas (LNG) is chiefly methane and boils normally at about −260° F. Pipelines for liquid natural gas have been proposed in North America, running from the far northern production fields south to the markets in the U.S. and Canada where the fuel is needed. Well-insulated pipelines, with recooling and pumping stations at intervals of perhaps 80.5 kilometres (50 miles), would be required.

In the early 1970s a small but increasing number of automobiles in the U.S. were running on liquid natural gas, which is stored in specially designed vehicle fuel tanks. The liquid is vaporized in a coil of copper tubing before it enters the carburetor of an automobile engine; ignition timing must be adjusted, but the conversion of autos to liquid natural gas is rather simple, unlike the major redesign necessary for jet aircraft power plants. Storage of liquid natural gas in service stations posed early problems. An advantage of natural gas is that its burning causes far less air pollution than does the combustion of gasoline.

By 1970 liquid natural gas was being moved not only by truck, tank car, and pipeline, but by ship from Alaska to Japan, and from North America and Africa to Europe. The first such oceangoing vessel, the "Methane Pioneer," made its initial voyage in 1959. Plans were also being developed for the large-scale transportation of liquid natural gas from South America to the U.S. and to Europe and Asia.

Storage tanks of up to 25,000,000 gallons were made part of the liquid natural gas energy-source network. The very large-scale production and subsequent large-scale storage, shipment, and utilization of this substance is un-

questionably the biggest facet of cryogenics' growth today. Tomorrow we may have a parallel growth in liquid hydrogen.

CRYOGENIC APPLICATIONS

Food preservation and shipment. Since 1955 liquid nitrogen has to some extent replaced conventional mechanical refrigerators for preserving food during shipment. The liquid nitrogen refrigeration system is simple and without moving parts. The storage tank is filled with liquid nitrogen, which when released from the spray headers, vaporizes immediately, refrigerating the truck's interior. Freezing temperatures are readily maintained, and temperatures as low as that of the liquid nitrogen are possible. Inert nitrogen gas in the storage compartments prevents oxidation loss and rodent activity. Another advantage of the system is that constant humidity is maintained so that wilting from dehydration is thus prevented. Even such perishables as lettuce and strawberries are delivered in good condition, at full weight, and with no change in appearance. Ships equipped with liquid nitrogen systems have carried produce across the Pacific with no losses. Precooling of the compartment to be refrigerated is not necessary. Other advantages of liquid nitrogen systems are simplicity of design, installation, and operation, with very accurate temperature control assured.

Quick-freezing tunnels for preparing frozen foods possess advantages that parallel those of the liquid nitrogen refrigerated vehicles (see Figure 1).

By courtesy of Airco Industrial Gases, Division of Airco, Inc.

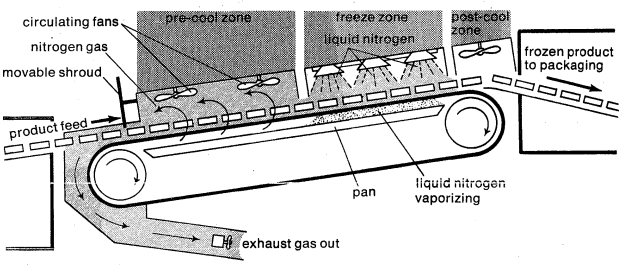


Figure 1: Tunnel freezer.

Cryomedicine. This discipline covers two distinct yet closely related categories: cryobiology, which is the study of the effect of low temperatures on biological materials, and cryosurgery, performed with a freezing probe instead of a cutting edge. Although the techniques employed in cryosurgery are quite modern, cold has been used in surgery for almost 100 years.

Freezing an aqueous solution, such as the content of a living cell, is a dehydration process. As water is converted to ice, the solutes present become more concentrated, until precipitation occurs. In a cell or tissue the dissolved colloids are subjected to increasing concentrations of electrolytes such as salt. Base-acid changes may occur, with the concentration of components contributing to the concentration of the hydrogen ion. The normal spatial arrangements of subcellular components may be disrupted by the volume change during expansion. All these and other phenomena introduce irreversible changes that can destroy the viability of a cell.

Selective tissue destruction

Even though very low temperature can destroy life, when selective tissue destruction is the end sought, as in surgery, cryogenics becomes a useful tool in well-being and in the prolongation of life.

Openchowski of Russia in 1883 developed an instrument utilizing evaporation of ether by warm air to produce cold for studies on the cerebral cortex of dogs. He applied his cold probe for periods of one to three minutes, freezing the area till solid. He noted that his treatment prevented hemorrhage and widespread tissue damage, which have become typical observations of cryogenic surgical treatment.

Cryogenic surgery, using a probe similar to that shown in Figure 2, has been performed in more than 5,000 cases of Parkinson's disease since April 1961. Relief from trembling can almost be assured. Ninety-three percent of the operations resulted in abolition of tremor and rigidity without any untoward side effects. The mortality rate was only one percent. These results are even more impressive when we consider the fact that some of the patients were of advanced age.

Many applications of cryogenics in ophthalmology date from the early 1930s when work was thus done to correct retinal separation. Cryoextraction of cataracts is accomplished by establishing an ice bond between the probe and the lens. Corneal buttons are being preserved cryogenically for grafting.

In otolaryngology, tonsillar tissue was the first to be cryogenically treated because of the ease of application. Cryotonsillectomy is now commonplace. The treatment, which is now relatively simple, can be done in a doctor's office, and the patient's pain and discomfort are greatly reduced as compared with conventional operational procedures.

By courtesy of Frigitronics Inc.

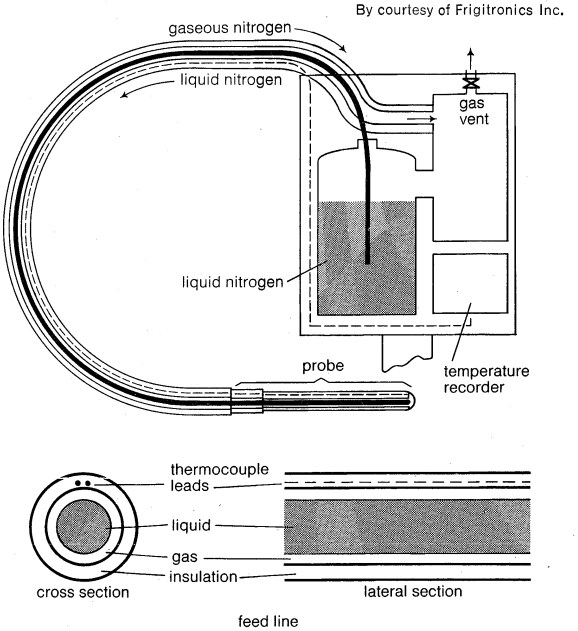


Figure 2: Principles of typical cryosurgery equipment.

Prostate malignancies have been treated cryogenically since 1964. The bladder is emptied, examined, and inflated with air, and the cryoprobe inserted into the urethra. The major advantages, as compared to conventional therapy, are the absence of hemorrhage and the need for less anesthesia. The disadvantage is the retention of dead tissue that can, in some cases, cause obstruction and infection.

Space applications. The principal space applications of cryogenic technology are in propulsion systems, fuel cells, life-support systems, system refrigerators, and space-simulation chambers.

Rocket propellants. Some cryogenic rocket propellants, such as liquid oxygen and liquid hydrogen used in combination, have a higher specific impulse (*i.e*, provide thrust more efficiently) than do the solid propellants available in 1970. A nuclear-powered rocket that uses hydrogen as the working fluid would achieve a still higher specific impulse; the hydrogen would be carried in its liquid form because a smaller and lighter fuel tank is possible. Consequently, some current and proposed propulsion systems are, and will be, cryogenic. In some future systems the propellant tankage must also be reusable, and very large quantities of propellants will be required.

Designs for cryogenic storage on the ground, as in the case of liquid natural gas service stations, present no insoluble problems, but long-term storage in space will depend on better insulation and the ability to expel and transfer the fluid. Solutions for important problem areas, such as protection from meteoroids, leakage control, and

reliability in the weightless space environment, are necessary before long-term storage in space can be considered feasible.

For large-volume requirements in space, weight savings in tanks would accrue from storing cryogens at less than critical pressure, and by utilizing insulation offering the lowest thermal conductivity-density factor. The construction materials must have high strength-to-weight ratios. Heat leaking into the cryogen must be minimized by control of any leakage along supports and piping, by evacuation of the system to prevent gas conduction losses, and by selecting the optimum radiation shielding to minimize radiation effects. Due to the degree of nonuniformity of superinsulation materials and installations, however, experimental measurements of a tank design are necessary to confirm predictions of thermal performance.

A small hydrogen vessel designed as part of a fuel cell is shown in Figure 3. The inner vessel can be titanium

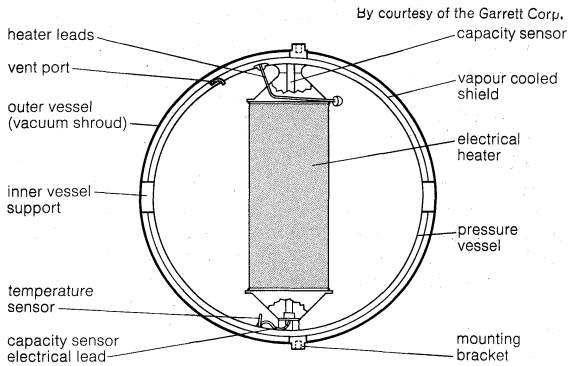


Figure 3: Elements of a liquid hydrogen vessel.

for hydrogen storage but not for oxygen. The outer shell can be aluminum.

Life-support systems. Cryogens in life-support systems are used to supply oxygen for high-altitude aircraft and for space cabin atmospheres. Liquid nitrogen and liquid helium are used as a source of pressurizing gas or to supply a two-gas atmosphere. Liquid hydrogen may be used in a process for recovering oxygen from expired carbon dioxide.

Space refrigeration techniques. Early in the space programs it became desirable to increase the sensitivity, reliability, and accuracy of some communication, mapping, and navigational devices by cooling certain sensors and optical components. Spaceborne refrigerators having one to two watt capacity at 77° K or lower were therefore necessary.

Cooling sensors

Significant accomplishments in adapting miniaturization techniques to the standard cycles, however, have been few. These systems must have high reliability, operate for very long periods of time, and must not introduce mechanical or electrical disturbances. Power requirements must be low. The Vuilleumier refrigerator, based on a 1918 U.S. patent, may meet these requirements. Although this type of thermal compression device was ignored for many years, in 1970 it was being used in aircraft and proposed for space. Basically, it employs two sets of low-speed piston displacers, a heat source, regenerators that connect the cylinders, and a drive mechanism. In most cases the electric motor is driven at less than 600 revolutions per minute to reduce wear.

Because of the critical power and reliability requirements of mechanical systems, solid cryogens are especially attractive for short-term cooling and with small cooling loads. An example is the use of solid hydrogen. The reasonable life of a solid hydrogen icebox is from three to 12 months with a cooling load of less than one watt at 10° K.

Cryogenic vacuum systems. For several industrial and research purposes, including space simulation, cryogenic pumping is used to attain hard vacuums beyond the reach of mechanical pumps. Atmospheric gases will condense on a surface if the temperature is low enough, much as moisture condenses on a cold windowpane. In one system an adsorbent, such as silica gel, is bonded to the surface of a cryopanel. The pumping speeds of cryogenically cooled adsorbents at very low pressures are sensitive to the amount adsorbed, but independent of the depth of the adsorbing material. The capacity of the material to adsorb increases rapidly with decreasing temperatures. By cooling the adsorbent to 77° K, all gases except hydrogen, helium, and neon can be efficiently trapped.

By cooling from 11° to 20° K, hydrogen can be trapped; and at the liquid helium temperature of 4.2° K gaseous helium can be adsorbed. Further studies of these cooling processes are needed to optimize cryosorption and cryopanel configurations.

Ultrahigh vacuums are used to study effects on materials, whereas chambers operating at high vacuums are used to test components and systems.

Superconductive devices. Kamerlingh Onnes, the Dutch physicist who investigated the dependence upon temperature of the electrical resistance of metals in 1911, observed that the resistance of mercury fell suddenly to zero when the metal was cooled to within a few degrees of absolute zero. Onnes coined the term superconductivity for this phenomenon. The temperature below which a material becomes superconductive has come to be designated the transition temperature, or critical temperature (T_c).

Since the transition temperatures of all known superconductors are in the cryogenic range, cryogenic engineering is required for applications of the phenomenon. In 1970 the chief application was the construction of superconductive electromagnets, primarily for research in physics, as in magneto-hydrodynamic power generation experiments and bubble chamber systems. Superconductive power transmission was being studied and appeared promising for the future, although the costs for refrigeration of long cables would be high. Other potential applications are in bearings, motors, and transformers. Superconductive magnets can be used with infrared detectors, lasers, and masers. The use of superconductive magnets in medical research has also proved successful.

Cryogenic bolometer

The bolometer, an instrument for measuring radiation, particularly in the infrared region, is more sensitive at low temperatures and most sensitive when it can be made superconductive. The first cryogenic bolometers used tantalum and required liquid helium for cooling. Later work on niobium nitride, which becomes superconducting at 14.3° K, appears to have produced a more practical device capable of operating on liquid hydrogen.

Future applications. There are at least four principal areas in which major advancements can be foreseen as the first century of cryogenics draws to a close: sources of energy, superconductivity, medicine, and pollution control.

It appears that worldwide petroleum production may peak shortly after the year 2000, with reserves decreasing in the United States in the 1980s. Hydrogen, which can be made available in almost limitless quantities by the electrolysis of water, offers an alternative chemical fuel, a major energy source for land, airborne, and space transportation systems. The necessary electrical energy for large-scale electrolysis can be supplied from nuclear power plants. The use of liquid natural gas in vehicles now seems to foreshadow such development.

Superconductive cables appear likely to play a role in electric power transmission and distribution systems, particularly if suitable materials can be obtained with transition temperatures higher than those currently available. In the early 1970s it was believed that hyperconductive cables of high purity aluminum cooled with liquid hydrogen could be utilized with existing technology dependent upon two factors: enough liquid hydrogen must be available and the costs of replacing existing power networks must not be excessive.

The magnetically suspended train—not quite a flying carpet, but nevertheless a unique form of high-speed

ground transportation—is a possible result of the future development of high-current and high-field superconductors. The train could be propelled magnetically by track currents.

Cryogenics technology also is being expanded in medicine; *e.g.*, in cancer therapy. Leucocytes, grown outside the human body and stored in liquid nitrogen, may someday prove an effective treatment. Cryogenic biological archives may be established for certain organs and tissues. Kidneys, endocrine glands, corneal tissue, and even hearts will perhaps be successfully frozen with viability retained. With all these successes, however, the prolonged viable preservation of whole animals, especially man, is now considered impossible, due to the complex and varied nature of the body. Cryogenic storage of spermatozoa, already used in cattle breeding, will be greatly increased for domestic animals, and has been applied to man.

Pollution control

The "no smog" characteristics of cryogenic fuels offer hope that the environment can be made more livable. Through cryogenic applications pollution of air and water can be drastically reduced, and landscapes beautified with the installation of cryogenically cooled underground power cables.

Cryogenics will have ever greater effect on man's life as developments from today's research become essential commodities.

BIBLIOGRAPHY. R.W. VANCE (ed.) and H. WEINSTOCK (co-editor of vol. 1), *Applications of Cryogenic Technology*, 2 vol. (1969–70), a thorough review of the applications of cryogenics emphasizing LNG and including an analysis of the thermodynamics of refrigeration; R.W. VANCE (ed.), *Cryogenic Technology* (1963), a summary of the theory and applications of cryogenics, including comprehensive reviews of cryobiology and superconductivity, and co-edited with W.M. DUKE, *Applied Cryogenic Engineering* (1962), a review of the early applications of cryogenics to the U.S. missile and space programs; R.B. SCOTT, *Cryogenic Engineering* (1959), the first text on cryogenic engineering, analyzing liquefaction technology and serving as an introduction to cryogenics; K.D. TIMMERHAUS (ed.), *Advances in Cryogenic Engineering* (annual), experimental data for all disciplines covered by cryogenics; *Proceedings of the First International Cryogenic Engineering Conference, Kyoto, Japan* (1967), a useful summary of applications as provided by engineers and scientists on a worldwide basis; theories and applications of superconductivity were emphasized along with possible future developments.

(R.W.V.)

Cryptology

Cryptology (from the Greek *kryptos*, "hidden," and *logos*, "word"), the science dealing with disguised or secret communications, is concerned with the methods and devices employed to camouflage communications (or even their existence), and to penetrate such communications contrived by others. Cryptology today plays a major role in the governmental and military affairs of nations and a lesser role in banking, industry, and commerce. Cryptology has also played a part in the deciphering of lost languages, and even in attempts to establish authorship of works, such as Shakespeare's plays. Although cryptology was concerned initially with secrecy of written messages, its principles apply not only to enciphered or encoded messages (cryptograms) but also to enciphered speech, enciphered facsimile, and enciphered television transmissions.

Cryptology embraces the twin or complementary sciences of signal security and signal intelligence. Signal security is concerned with all the methods of protecting one's own signals against interception and reading or utilization by unauthorized persons (generally called "the enemy"). Signal intelligence comprises all the methods employed in acquiring information or intelligence by intercepting and solving the enemy's cryptosignals or nullifying his signal security so the signals or information derived from them can be used against him.

SIGNAL SECURITY

There are three important aspects to signal security. Physical and personnel security assures that physical arrangements, facilities and procedures for safeguarding cryptomaterials (codes, ciphers, key lists, cipher machines) are adequate and that the personnel using codes and ciphers are trustworthy. Transmission security embraces the means, methods, and procedures for assuring that no information useful to the enemy is inadvertently disclosed either by indiscretions of operators or by equipment malfunctions. Cryptosecurity deals with the technical adequacy of the cryptography or cryptosystems employed.

Cryptography. Cryptography (from the Greek *kryptos*, "hidden," and *graphein*, "to write") deals with the methods involved in preparing cryptograms—messages or writings intended to be incomprehensible to all except those who legitimately possess the means to reproduce the original plaintext. Conversion of a plaintext message into a cryptogram is called encrypting (or, more specifically, enciphering or encoding); reconverting the cryptogram back into its intelligible form, when done by a legitimate or authorized communicator, is called decrypting (specifically, deciphering or decoding). Although in theory no sharp line of demarcation can be drawn between code systems and cipher systems, in modern practice the technical differences between them are sufficiently marked to warrant their being treated separately. Cipher systems will be discussed first, then code systems; only a very limited number suitable for serious usage can here be outlined.

Cipher systems. Generally speaking, cipher systems involve a cryptographic treatment of textual units of fixed length, usually one, two, or three letters. These textual units are treated as symbols without reference to their identities as component parts of words, phrases, and sentences. Every practical cipher system must combine two elements: a set of rules, processes, or steps constituting the basic cryptographic method of treatment or procedure, called the general system, agreed upon in advance by the communicators and constant in character; and a specific key that is variable in character. The specific key may consist of a number or a series of numbers, a word, a phrase, or a sentence.

General system, specific key

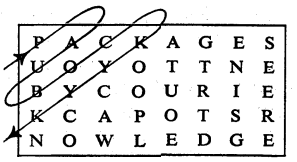
In encipherment, the key controls the steps under the general system and determines the specific nature or exact composition of the cipher message produced; in decipherment, the key similarly controls the steps and determines what the deciphered text will be. When all of these operations are performed correctly, the two plaintexts (before and after the cryptography) should be identical or nearly so. The general system should be so constructed that, even if it is known to the enemy, no properly enciphered message can be read by him unless he also knows the specific key or keys applicable to that message.

Despite great diversity in the external appearance and internal constitution of ciphers, there are only two basic classes of systems. These two classes are called transposition and substitution. A transposition cipher involves a rearrangement or change in the sequence of the letters of the plaintext message without any change in their identity; while a substitution cipher involves a replacement of the plaintext letters by other letters (or other symbols) without any change in their sequence. Transposition and substitution may be combined in a single cryptosystem.

Transposition ciphers. Almost all transposition ciphers involve the use of a geometrical figure or design in which the elements of the plaintext (usually single letters) are first inscribed according to a previously agreed-upon direction or method and then transcribed by another (and different) agreed-upon method. In nearly all cases the specific key controls (1) the use of designs of a specific nature and dimensions, and (2) variation in the manner of inscription or transcription or both.

Route cipher

One of the simplest forms of transposition ciphers is a route cipher. In the example given, the message plaintext, PACKAGE SENT TO YOU BY COURIER STOP ACKNOWLEDGE, is inscribed by an alternate horizontal route on cross-section paper into a rectangle of eight columns, and the letters are transcribed, as indicated, in



an alternate diagonal route starting at the upper left-hand corner. The final ciphertext, grouped in five-letter groups as a letter-check and for convenience in transmission, then reads: PAUBO CKYYK NCCOA GTOAO WPUTE SNROL ETIEE SDGRE. In decipherment the cipher letters are inscribed in an alternate diagonal route following that originally used for the transcription. Using this method, the plaintext will reappear in the alternate horizontal route originally used for the inscription.

Other routes that may be employed are simple horizontal or vertical, alternate vertical, simple diagonal, and clockwise or counterclockwise spirals from a corner or the centre of the diagram. In the very common keyed columnar transposition cipher, the plaintext letters are written into a geometrical design, most often a rectangle, in a simple horizontal route from left to right and top to bottom, and the letters are transcribed from the columns in the sequence determined by a numerical key. In the example below, involving an incompletely filled rectangle, the numerical key is derived from the key word CIGARETTE by numbering the letters in the sequence in which they appear in the normal alphabet:

C	I	G	A	R	E	T	T	E
2	6	5	1	7	3	8	9	4
P	A	C	K	A	G	E	S	E
N	T	T	O	Y	O	U	B	Y
C	O	U	R	I	E	R	S	T
O	P	A	C	K	N	O	W	L
E	D	G	E					

thus, A is 1, C is 2, the two E's are 3 and 4, etc. The final ciphertext is KORCE PNCOE GOENE YTLCT UAGAT OPDAY IKEUR OSBSW. In decipherment, a rectangle with the proper number of cells (determined by the length of the message and the length of the key) must first be prepared. The ciphertext is then inscribed vertically down the columns according to the numerical key. Thus, the first three cipher groups—KORCE PNCOE GOENE—when inserted into the rectangle, would look as follows:

C	I	G	A	R	E	T	T	E
2	6	5	1	7	3	8	9	4
P			K		G			E
N			O		O			
C			R		E			
O			C		N			
E			E					

When all of the columns have been inscribed, the entire plaintext of the cryptogram will reappear.

Double transposition ciphers

The foregoing examples involved single transposition. In double transposition ciphers, the ciphertext resulting from the first transposition is put through a second transposition in the same or a different key. In the case of the last example given above, if the second transposition used the same key as the first, the primary ciphertext KORCE PNCOE . . . would be inscribed in simple horizontals under the key, and then the columns taken off in

2	6	5	1	7	3	8	9	4
K	O	R	C	E	P	N	C	O
E	G	O	E	N	E	Y	T	L
C	T	U	A	G	A	T	O	P
D	A	Y	I	K	E	U	R	O
S	B	S	W					

numerical-key order. The final ciphertext would then be CEAIW KECDS PEAEO LPORO UYSOG TABEN GKNYT UCTOR.

In decipherment the order of steps is reversed: the two rectangles of the proper dimensions are prepared, and the cryptogram is inscribed in the columns of the second rectangle under the digits of the key according to their numerical order. When the second rectangle is completely filled, the rows of this rectangle are transcribed and written down the columns of the first rectangle according to its numerical key, and the plaintext will reappear in the rows.

Transposition ciphers exist in many modifications. In the diagonal-columnar variation example below, a key of 18 elements has been expanded from a shorter key phrase, and a variation in transcription has been employed:

C	H	I	N	E	S	E	L	A	U	N	D	R	Y	C	H	I	N
2	7	9	12	5	16	6	11	1	17	13	4	15	18	3	8	10	14
A	M	M	U	N	I	T	I	O	N	T	R	A	I	N	L	E	A
V	I	N	G	G	R	E	E	N	V	I	L	L	E	A	T	O	N
E	F	O	U	R	T	H	R	E	E	Z	E	R	O	T	O	D	A
Y	S	T	O	P	T	R	O	O	P	T	R	A	I	N	A	R	R
I	V	I	N	G	N	E	W	T	O	N	A	T	Z	E	R	O	S
I	X	F	O	U	R	F	I	V	E								

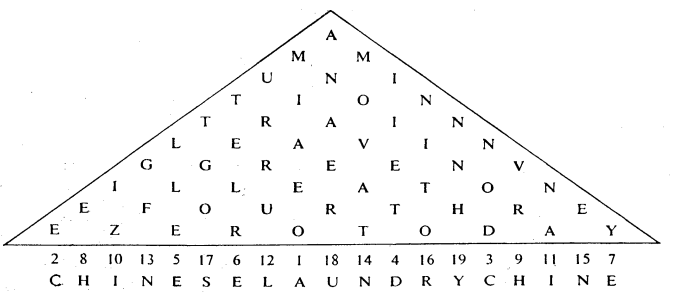
The ciphertext is formed by taking off diagonals, alternating in direction, of the letters under the key numbers 1, 2, 3, and 4 as shown, following which the remaining letters are taken off in columns according to the key, not duplicating the letters that already have been taken off in the diagonals.

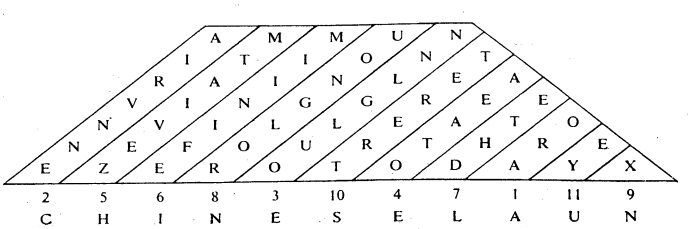
The final ciphertext would then be OEHTG OAIOO RNERR . . . ATIRT NNVEP OEIIZ.

In another variation, which is known as the interrupted-key columnar transposition cipher, the numerical key serves two purposes: it determines the cut-off point (and therefore the number of letters) in each row of the diagram; and determines the order of transcription of the columns.

C	H	I	N	E	S	E	L	A	U	N	D	R	Y
2	6	7	9	4	12	5	8	1	13	10	3	11	14
A	M	M	U	N	I	T	I	O	·	·	·	·	·
N	·	·	·	·	·	·	·	·	·	·	·	·	·
T	R	A	I	N	L	E	A	V	I	N	G	·	·
N	E	W	T	O	·	·	·	·	·	·	·	·	·
N	A	T	O	N	E	S	·	·	·	·	·	·	·
I	X	·	·	·	·	·	·	·	·	·	·	·	·
T	H	R	·	·	·	·	·	·	·	·	·	·	·
E	E	Z	E	R	O	T	O	·	·	·	·	·	·
D	A	Y	X	·	·	·	·	·	·	·	·	·	·

The final ciphertext here is OVANT NNITE . . . ITOEX NILEO I. (Note, in the diagram, the final null X added to complete the row properly.) Other geometrical designs, such as triangles or trapezoids, have been used, as in the following examples:





In the first example, the columns are taken off from their tops according to the key, so that the ciphertext would read MIAEO ENODI . . . . In the second example, the diagonals are read from the bottom, so that the final ciphertext reads AROEN NVRIA. . . .

Grilles

A type of transposition occasionally encountered involves the use of a square sheet of paper or cardboard called a grille, in which small square cells have been cut out in definite but irregular positions; the letters of the plaintext are inscribed on a sheet underneath the perforated design. Usually the grille is revolved 90 degrees in four successive operations so that the resulting square of letters inscribed beneath the grille is completely filled, and then the letters are taken out in groups of five, reading horizontally or otherwise, according to agreement. The perforations must, of course, be correctly disposed on the grille so every space on the sheet over which it is placed in inscribing the letters is filled after the four turns have been completed.

Monoalphabetic substitution. These systems make use of a single cipher alphabet consisting of a plain component, in which are found the plaintext letters to be enciphered; and a cipher component, in which are found the ciphertext equivalents of the plaintext letters. Cipher alphabets may be direct standard, as in the example below

Plain: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Cipher: L M N O P Q R S T U V W X Y Z A B C D E F G H I J K

(one of the 25 possible slides or juxtapositions of the two sequences for monoalphabetic substitution), with which ARTILLERY would be enciphered as LCETWWPCJ; or they may be reversed standard, as in the example below

Plain: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Cipher: C B A Z Y X W V U T S R Q P O N M L K J I H G F E D

(one of the 26 possible slides of these sequences), with which ARTILLERY would be enciphered as CLJURRYLE. Note that a sequence running against itself in reverse must always give rise to reciprocal substitution: here R plain is L cipher, and L plain is R cipher (or, in cryptologic notation, $R_p = L_c$, and $L_p = R_c$), and so on for the 12 other reciprocal pairs. Cipher alphabets may also be systematically mixed in some fashion, such as in the example below

Plain: A T H E N S B C D F G I J K L M O P Q R U V W X Y Z
Cipher: L M N O Q T V W X Y Z P I R A E U S B C D F G H J K

in which the plain and cipher components are two different keyword-mixed sequences (at the juxtaposition $A_p = L_c$), or in the example below

Plain: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Cipher: D A K S Y E B M T Z L C N U P F O V H G Q W I J R X

in which the cipher component is a transposition-mixed sequence based on the key word DELPHI and derived from the following diagram:

D	E	L	P	H	I
A	B	C	F	G	J
K	M	N	O	Q	R
S	T	U	V	W	X
Y	Z				

Multiliteral systems. The monoalphabetic systems just described are uniliteral systems; *i.e.*, the cipher equivalents for the plaintext elements consist of single letters. Multiliteral systems are those in which the cipher equivalents consist of two or more characters, typically pairs of letters or digits produced from biliteral or bipartite matrices or diagrams such as the following:

	B	L	A	C	K
W	A	B	C	D	E
H	F	G	H	IJ	K
I	L	M	N	O	P
T	Q	R	S	T	U
E	V	W	X	Y	Z

	0	1	2	3	4	5	6	7	8	9
0	C	I	G	A	R	E	T	B	D	F
1	H	J	K	L	M	N	O	P	Q	S
2	U	V	W	X	Y	Z	.	,	(	)

Encipherment is accomplished by designating the row and column coordinates at which the plaintext element appears; thus, in the diagrams above, plaintext E is enciphered as WK in the first diagram or as 05 in the second. (Note that in the first matrix the coordinates are systematically mixed whereas in the second matrix it is the interior that is mixed, based on the key word CIGARETTE—with repeated letters of course suppressed after their initial occurrence.) Multiliteral systems may also employ mixed-length units in the ciphertext, as in the following monome–dinome matrix (wherein the internal key is based on CIGARETTES, and the sequence for the column coordinates is the numerical key derived from the letters at the top of the columns, CIGARETSNO):

	2	5	4	1	8	3	0	9	6	7
—	C	I	G	A	R	E	T	S		
6	B	D	F	H	J	K	L	M	N	O
7	P	Q	U	V	W	X	Y	Z	.	,

In this system the letters in the top row are enciphered as single digits (monomes), whereas the other letters are enciphered by pairs of digits (dinomes). Thus, the word HEADQUARTERS would be enciphered as

61 3 1 65 75 74 1 8 0 3 8 9,

which would appear as 61316 57574 18038 9 . . . when grouped into five-digit groups for transmission. There is no ambiguity in decipherment, since the row-coordinate digits (here 6 and 7) cannot represent monomes. Multiliteral systems may be modified to yield two or more cipher values, called variants, for the plaintext elements, as shown in the example below:

Variant systems

			U	V	W	X	Z
			P	Q	R	S	T
A	F	K	A	B	C	D	E
B	G	L	F	G	H	IJ	K
C	H	M	L	M	N	O	P
D	I	N	Q	R	S	T	U
E	J	O	V	W	X	Y	Z

Here plaintext E has six variants, any one of which may be used at the whim of the encipherer: AT, AZ, FT, FZ, KT, and KZ. Furthermore, since this is a commutative matrix (*i.e.*, the row coordinates are distinct from the column coordinates), the cipher equivalents may also be taken in column-row order, providing a total of 12 variants for each plaintext letter. Another system for variant encipherment is the following table, which provides four dinome equivalents for each plaintext letter:

A	24	38	68	89	H	06	45	75	96	O	12	26	56	77	U	18	32	62	83
B	25	39	69	90	IJ	07	46	51	97	P	13	27	57	78	V	19	33	63	84
C	01	40	70	91	K	08	47	52	98	Q	14	28	58	79	W	20	34	64	85
D	02	41	71	92	L	09	48	53	99	R	15	29	59	80	X	21	35	65	86
E	03	42	72	93	M	10	49	54	00	S	16	30	60	81	Y	22	36	66	87
F	04	43	73	94	N	11	50	55	76	T	17	31	61	82	Z	23	37	67	88
G	05	44	74	95															

Here the plaintext sequence consists of 25 letters (combining I and J), and the cipher sequences consist of slides of the dinomes 01–25, 26–50, 51–75, and 76–00; the starting points of the four cipher sequences are set according to the specific key, which in this case is 01 = C, 26 = O, 51 = I, and 76 = N, spelling the key word COIN.

Digraphic systems. All of the preceding monoalphabetic substitution ciphers are monographic systems: the plaintext elements treated are single characters. In digraphic systems the plaintext being enciphered consists of a pair of letters forming an indivisible compound in the enciphering process. The encipherment may be accomplished through the use of a 26 × 26 square, such as the following one, which is shown in fragmentary form, in which the letters of the plaintext pair are found in the side and top sequences, and the digraphic cipher equivalent is the pair of letters in the cell at the junction of the row and column coordinates. This table has been constructed so as to yield reciprocal substitution (*e.g.*, AB in plaintext is CH in cipher; and CH in plaintext is AB in

first letter \ second letter	A	B	C	D	E	F	G	H	I	J	K	X	Y	Z
A	FX	CH	XE	YY	ZA	YG	FB	CD	EF	XJ	ZX	EA	DJ	FH
B	NY	DC	NB	ZI	XX	DX								
C				AH				AB					ND	
D			BB		YA					AY		BF		
E	AX					AI								
F		AG			NZ			AZ				AA		
N		BC		CY									BA	FE
X					AC					AJ		BE		
Y	DE						AF						AD	
Z	AE								BD			AK		

cipher), so a single table will suffice for both encipherment and decipherment, avoiding construction of a separate deciphering table in which the ciphertext letters would have been found in the row and column coordinates and the plaintext equivalent in the interior of the table. The best known system of digraphic encipherment is the Playfair cipher, invented by Sir Charles Wheatstone, but sponsored at the British Foreign Office by Lord Lyon Playfair. This cipher uses a 5×5 square, such as the following, in which I and J have been combined into a single cell:

Playfair cipher

C	IJ	G	A	R
E	T	B	D	F
H	K	L	M	N
O	P	Q	S	U
V	W	X	Y	Z

In encipherment, if the two letters of the pair are in the same row, their cipher equivalents are the letters immediately to their right, so that CA in plaintext is IR in cipher; CR in plaintext is IC in cipher; and RA in plaintext is CR in cipher. If the two letters are in the same column, their cipher equivalents are the letters immediately below, so that SA in plaintext is YD in cipher; HO in plaintext is OV in cipher; and VE in plaintext is CH in cipher. If the letters are found in opposite corners of an imaginary rectangle as defined by the diagonal, the cipher equivalents are the letters of the opposite diagonal, so that ES in plaintext is DO in cipher; PA in plaintext is SI in cipher; and RY in plaintext is AZ in cipher. If doubled letters should be within a pair, they are customarily separated by an X; and an X is also added to the plaintext if the final digraph is incomplete, as may be seen in the following example (using the square given above):

Plain: IN TE RX RO GA TI ON NO WI NP RO GR ES SX

Cipher: RK BT GZ CU AR KT UH HU IT KU CU AC DO QY

The cipher letters would then be grouped into 5's for transmission, so that the cryptogram would read RKBTG ZCUAR. . . .

Polyalphabetic substitution. These systems involve a multiplicity of cipher alphabets, used in accordance with definite rules and procedures; the cipher equivalents of plaintext elements are therefore not at the whim of the encipherer, as is the case with systems involving monoalphabetic substitution with variants. The cipher alphabets employed may be interrelated (*e.g.*, the 26 different alphabets produced by the interaction of a pair of sliding components, such as those discussed under monoalphabetic substitution), or they may be any number of entirely unrelated alphabets. The best known type of polyalphabetic substitution employs a Vigenère table illustrated in Figure 1, named after the French cryptographer Blaise de Vigenère (1523–1596), who popularized it though he did not invent it. In using this table, the plaintext letters are found in the sequence above the square, the key letters in the column at the side of the square, and the cipher letters at the junction of the key-letter row and the plaintext-letter column. Thus, plaintext D in the key of F is cipher I; or, in cryptologic notation, $D_p(F_k) = I_c$. Instead of direct standard alphabets, as in the original Vigenère table, the alphabets may be reversed standard, or they may even involve a mixed plain component against the 26 slides of a mixed cipher component. In periodic polyalphabetic substitution the key, usually a plaintext word or phrase, is used cyclically, as in the following example employing the repeating key FLUTES in conjunction with the Vigenère table given below:

Key:	F	L	U	T	E	S	F	L	U	T	E	S	F	L	U	T	E	S	F	
Plain:	D	I	S	R	E	G	A	R	D	M	Y	L	A	S	T	M	E	S	S	A
Cipher:	I	T	M	K	I	Y	F	C	X	F	C	D	F	D	N	F	I	K	X	L

	G	E	X	X	A	X	B	P	C											

(Note the enciphered X's at the end, added to complete the final group of five letters.) In actual practice, it is customary to encipher all the letters belonging to alphabet 1 (*i.e.*, the "F" alphabet in this example), then those of alphabet 2, etc., thereby speeding up the process and minimizing the chance of error. Polyalphabetic encipherment (including modifications involving mixed components) may also be accomplished by means of a pair of sliding strips on which are written the double-length sequences of the plain and cipher components, or by means of concentric disks, such as the World War I device shown in Figure 2 (employing reversed standard alphabets); by convention, the successive key letters (located in the cipher component) are set against an index letter in the plain component, usually either letter A in plaintext (A_p) or the initial letter of the plain component. Thus, the letters of the key word or phrase serve to indicate the number, identity, and sequence of the cipher alphabets

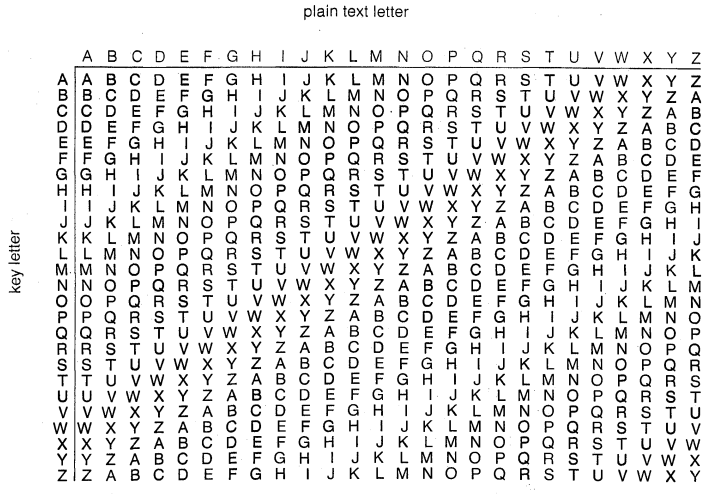


Figure 1: *The Vigenère table.*
The cipher equivalent of a given plaintext letter enciphered by a given key letter is that letter that stands at the intersection of the vertical column headed by the former and the horizontal row begun by the latter. Thus, plaintext letter E enciphered by key letter M yields cipher letter Q, or, in brief notation, $E_p(M_k) = Q_c$; $E_p(X_k) = B_c$, etc.

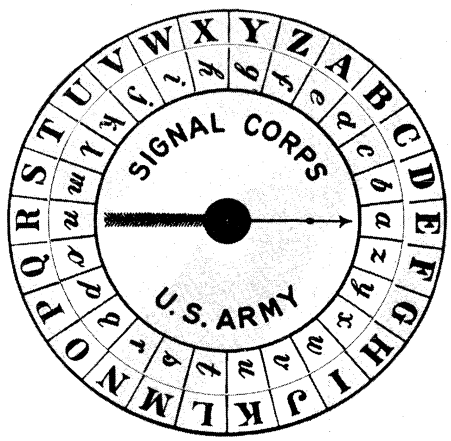


Figure 2: United States Army cipher disk.

Gronsfeld cipher

to be used in the operations of encipherment and decipherment. The system known as the Gronsfeld cipher is identical with a Vigenère system with direct standard alphabets, except that only the first ten alphabets are used in conjunction with a numerical key. The key digit indicates how much displacement the plaintext letters should have (as measured on the normal sequence); hence, the limitation of only the first ten rows of the Vigenère table. The example below shows an encipherment with the repeating 12-digit keying sequence 743189602512:

Key:	7	4	3	1	8	9	6	0	2	5	1	2	7	4	3	1	8	9	6	0	2	5	1	2	7
Plain:	D	I	S	R	E	G	A	R	D	M	Y	L	A	S	T	M	E	S	S	A	G	E	X	X	X
Cipher:	K	M	V	S	M	P	G	R	F	R	Z	N	H	W	W	N	M	B	Y	A	I	J	Y	Z	E

The advantages of the Gronsfeld cipher are that encipherment and decipherment may be accomplished without the use of a table and that easily available long numerical keys (*e.g.*, the numbers in a telephone directory) may be used by the correspondents.

Aperiodic substitution

Aperiodic polyalphabetic substitution systems employ either long keys that do not repeat (as, for example, the text from a book) or keys used in such a way as to disturb or eliminate periodicity, even if the basic key is relatively short and therefore cyclic in nature. In a system involving a plaintext interruptor, for example, the occurrence of a predetermined plaintext letter might cause a skip of one position in the basic key, as in the following Vigenère encipherment, in which the letter A in plaintext (A_p) is the interruptor letter and FLUTES the basic key:

Key:	F	L	U	T	E	S	F	U	T	E	S	F	L	T	E	S	F	L	U	T	S	F	L	U	T
Plain:	D	I	S	R	E	G	Ⓐ	R	D	M	Y	L	Ⓐ	S	T	M	E	S	S	Ⓐ	G	E	X	X	X
Cipher:	I	T	M	K	I	Y	F	L	W	Q	Q	Q	L	L	X	E	J	D	M	T	Y	J	I	R	Q

The interruptor letter may also be a designated cipher letter; furthermore, the interruptor letter may influence the key in a number of ways—*e.g.*, it may cause the immediately preceding key element to be repeated, or it may cause the key to reset to its initial position. Autokey systems are those in which the key is automatically generated from the plaintext or the ciphertext, after an initial or priming key of one or more letters. In the plaintext autokey cipher below, for example, the first letter of the message

Key:	X	D	I	S	R	E	G	A	R	D	M	Y	L	A	S	T	M	E	S	S	A	G	E	X	X
Plain:	D	I	S	R	E	G	A	R	D	M	Y	L	A	S	T	M	E	S	S	A	G	E	X	X	X
Cipher:	A	L	A	J	V	K	G	R	U	P	K	J	L	S	L	F	Q	W	K	S	G	K	B	U	U

is enciphered in the key of X, after which the plaintext itself serves as the key for enciphering the rest of the message. In the following example of a ciphertext autokey cipher, the beginning of the message is enciphered

Key:	F	L	U	T	E	S	I	T	M	K	I	Y	I	K	P	W	G	J	I	C	I	I	K	B	A
Plain:	D	I	S	R	E	G	A	R	D	M	Y	L	A	S	T	M	E	S	S	A	G	E	X	X	X
Cipher:	I	T	M	K	I	Y	I	K	P	W	G	J	I	C	I	I	K	B	A	C	O	M	H	Y	X

with the initial key FLUTES, after which the ciphertext being produced serves as the key. In the running-key cipher the key text is taken from a book in possession of the correspondents; it is necessary either to agree to the starting point of the key or to designate this by means of an indicator that might specify the page, paragraph, line or word number in the particular book being used.

Fractionation systems. These systems employ a multiliteral substitution as the first step, transforming the plaintext into biliteral (or sometimes triliteral) equivalents, that are then subjected to a further encipherment, usually transposition. The following bipartite square and the encipherment of a message in vertical dinomes under the plaintext letters provide an example:

	1	2	3	4	5
1	F	L	U	T	E
2	S	A	B	C	D
3	G	H	I	K	M
4	N	O	P	Q	R
5	V	W	X	Y	Z

A	L	L		Q	U	I	E	T		I	N		T	H	I	S		S	E	C	T	O	R
2	1	1		4	1	3	1	1		3	4		1	3	3	2		2	1	2	1	4	4
2	2	2		4	3	3	5	4		3	1		4	2	3	1		1	5	4	4	2	5

The ciphertext is formed by taking off all the digits in the top row, and then in the second row, as dinomes (21 14 13 11 34 . . . 42 31 15 44 25) and converting the dinomes thus formed into single letters using the same square; the final ciphertext would then read STUFK . . . OGEQD. The foregoing example is that of an aperiodic fractionation system, since the keying is dependent upon the message as a whole. In periodic fractionation the plaintext is divided into regular-length groupings, and the treatment is applied to the individual groupings in turn. Note the following example with a grouping-interval of 5:

Periodic fractionation

A	L	L	Q	U		I	E	T	I	N		T	H	I	S	S		E	C	T	O	R
2	1	1	4	1		3	1	1	3	4		1	3	3	2	2		1	2	1	4	4
2	2	2	4	3		3	5	4	3	1		4	2	3	1	1		5	4	4	2	5

The dinomes taken off here are 21 14 12 22 43 . . . 12 14 45 44 25, which, when converted back into letters, become STLAP . . . LTRQD. A fractionation cipher may also make use of a tripartite cipher alphabet, such as that shown below:

F	111	A	131	I	221	O	233	W	322
L	112	B	132	J	222	P	311	X	323
U	113	C	133	K	223	Q	312	Y	331
T	121	D	211	M	231	R	313	Z	332
E	122	G	212	N	232	V	321	*	333
S	123	H	213						

Using this alphabet in a fractionation cipher with a period of five produces the encipherment shown below, wherein the equivalents of the plaintext letters are the vertical trinomes (*i.e.*, three-digit groups) beneath them:

A	L	L	Q	U		I	E	T	I	N		T	H	I	S	S		E	C	T	O	R
1	1	1	3	1		2	1	1	2	2		1	2	2	1	1		1	1	1	2	3
3	1	1	1	1		2	2	2	2	3		2	1	2	2	2		2	3	2	3	1
1	2	2	2	3		1	2	1	1	2		1	3	1	3	3		2	3	1	3	3

The trinomes are now recombined horizontally under each grouping of five letters as shown by the parentheses, 111 313 111 112 223 . . . 111 232 323 123 133; the trinomes are then converted back to letters, yielding the final ciphertext FRFLK . . . FNXSC. Should the grouping 333 occur in the recombination, its substitute will be either the 27th character in the alphabet (*) or the result of a special convention; *e.g.*, wherein the grouping 332 will always be represented by ZA as an inseparable compound, and the grouping 333 represented by ZB. The deciphering clerk, on encountering a Z, will take the next letter into consideration; therefore, no cryptographic ambiguity will be present.

Germany's ADFGVX cipher

A famous case of fractionation in combination with columnar transposition was a cipher used by the German Army high command in World War I, a system known to the Allies as the ADFGVX cipher because only those six letters—with easily distinguishable Morse equivalents—appeared in the ciphertexts. This system used a 6 × 6 bipartite square containing the 26 letters and the 10 digits, with coordinates A, D, F, G, V, and X. The biliteral text produced in the first step was inscribed into a transposition rectangle, and the columns were taken off according to the numerical key to produce the final cipher text, as in the following example:

	A	D	F	G	V	X
A	G	7	E	5	R	M
D	A	1	N	Y	B	2
F	C	3	D	4	F	6
G	H	8	I	9	J	Ø
V	K	L	O	P	Q	S
X	T	U	V	W	X	Z

A	L	L	Q	U	I	E	T	I	N
DA	VD	VD	VV	XD	GF	AF	XA	GF	DF
T	H	I	S	S	E	C	T	O	R
XA	GA	GF	VX	VX	AF	FA	XA	VF	AV

G	E	R	M	A	N	Y
3	2	6	4	1	5	7
D	A	V	D	V	D	V
V	X	D	G	F	A	F
X	A	G	F	D	F	X
A	G	A	G	F	V	X
V	X	A	F	F	A	X
A	V	F	A	V		

final cipher text:

VFDFF VAXAG XVDVX AVADG
FGFAD AFVAV DGAAF VFXXX

Code systems. A code system is a specialized form of substitution in which entire words, long phrases, or even sentences of the plaintext are replaced by arbitrarily selected equivalents—which may be other words, groups of letters, groups of figures, or combinations thereof. It is only exceptionally that the substitution process is applied to elements smaller than whole words, and when this is done the elements are single letters, groups of letters, or syllables. The transition between cipher and code is not too clear, as may be seen from the example shown below of a syllabary square, which contains the 26 letters, 10 digits, and 64 digraphs and trigraphs chosen on the basis both of frequency considerations

Syllabary squares

	1	2	3	4	5	6	7	8	9	0
1	A	I	AL	AN	AND	AR	ARE	AS	AT	ATE
2	ATI	B	2	BE	C	3	CA	CE	CO	COM
3	D	4	DA	DE	E	5	EA	ED	EN	ENT
4	ER	ERE	ERS	ES	EST	F	6	G	7	H
5	8	HAS	HE	I	9	IN	ING	ION	IS	IT
6	IVE	J	Ø	K	L	LA	LE	M	ME	N
7	ND	NE	NT	O	OF	ON	OR	OU	P	Q
8	R	RA	RE	RED	RES	RI	RO	S	SE	SH
9	ST	STO	T	TE	TED	TER	TH	THE	THI	THR
0	TI	TO	U	V	VE	W	WE	X	Y	Z

and the combinative potentialities of the particular polygraphs; the word ESTIMATED might here be encrypted by dividing it up into one of the possible sets of groupings EST-I-M-ATE-D and replacing them by their dinome equivalents 45 54 68 10 31. Other squares or matrices, of larger dimensions, may contain not only a larger number of different plaintext elements within the matrix but may also duplicate some of the more frequent plaintext elements, providing plaintext variants within the matrix. Furthermore, when letters are used as coordinates, variant cipher equivalents may be incorporated into the system. When a matrix also incorporates words (or even short phrases) among the plaintext elements, it is called a code chart.

Code books

True code systems make use of a code book in which the words, phrases, and sentences of the vocabulary are listed in a systematic manner and accompanied by their equivalent code groups; correspondents must possess identical copies of the document in order to communicate with each other. There are various types of codes each suited to particular types of correspondence. Some are large books for general business or social correspondence; others are intended for particular industries and contain highly specialized vocabularies. Most large commercial firms have their own private codes, constructed especially for their use. Code is used in commercial communications principally for economy in communication; secrecy is of secondary importance. In modern military signal communications, code is used to effect secrecy, brevity, and speed, especially in front-line communications. The condensing power of a code book varies with the extensiveness of its vocabulary, since a small book can list only the most common words and a few phrases and sentences, whereas a large book can include practically all the words likely to be used in telegraphic communication, together with many common phrases and sentences.

In encoding a message, a code clerk merely replaces the plaintext words, phrases, sentences, and numbers by their code equivalents. The code text is built up from code units, each representing the longest possible plaintext unit the code book affords. If the sentence ENEMY TROOP CONCENTRATION WEST OF NEWTON is to be encoded, for example, and the code book lists the phrase ENEMY TROOP CONCENTRATION, the code group representing this phrase would be used in preference to the separate code groups representing the individual words; and the place name NEWTON, if not in the code vocabulary, could be encoded by using the two groups for NEW and TON. (Many codes contain a small block of code groups designated as supplement groups, the meanings of which are defined in special lists that are effective for specific dates or geographical areas.)

Modern codes usually contain code groups composed of three, four, or five letters or four or five figures. The length of the code groups used depends upon the size of the code. A code may contain two or more parallel sets of code groups of different types. In many commercial codes and in some military and naval codes, for example, there is one series of literal code groups and another series of digital groups, both applying to the same series of vocabulary elements. Code groups of modern codes are constructed by the use of special tables, called permutation tables, from which are generated code groups having a two-letter differential, which guarantees that each code group differs from all the others by at least two letters; *e.g.*, BACOF and BACUG, or BACOF and BOCAF, a differential that presents a positional difference of two letters. The permutation table helps to correct errors in transmission or reception.

As regards their construction, codes are generally one-part (alphabetical) codes or two-part (randomized) codes. In one-part codes the plaintext groups, as well as the code groups, are arranged in alphabetical or numerical order. Such a code serves for decoding and encoding. In two-part codes the plaintext groups are arranged in alphabetical order accompanied by their code groups in a nonsystematic order. The code groups are assigned to the plaintext groups at random by drawing the code groups out of a box in which they have been thoroughly mixed, or by some other manner involving the element of chance. Such a list can serve only for encoding. For decoding, another list is required in which the code groups are arranged alphabetically or numerically, accompanied by their meanings as given in the encoding section. The following brief extracts from typical one-part and two-part codes illustrate the difference:

one-part code		two-part code			
		encoding section		decoding section	
ABABD	a	GAJVY	a	ABABD	obstructed
ABACF	abaft	TOGTY	abaft	ABACF	term
ABAHK	abandon	FEHIL	abandon	ABAHK	zero
ABAJL	—it	BAYLT	—it	ABAJL	stalemate
ABALN	abandoned	ZYZYZ	abandoned	ABALN	franc(s)
ABAMP	—by	NYSYZ	—by	ABAMP	acceding
ABAWZ	abandoning	IFWUZ	abandoning	ABAWZ	building
ABBAD	abandonment	RUMGO	abandonment	ABBAD	do not send
...	...	...	...	...	...
ZYZYZ	zero	ABAHK	zero	ZYZYZ	abandoned

Two-part codes are used by many governments for secret diplomatic, military, and naval communications for their greater security, because the recovery of one code group by an enemy cryptanalyst will give no clues as to any other group; and greater accuracy because lists of numbers, dates, and similar categories have equivalents not related, thus facilitating the detection of errors.

Enciphered code

Sometimes the code groups of a code message undergo a further process of encipherment; either of the two general classes of cipher methods, transposition and substitution, may be used in enciphering code. Enciphered code is used to give security to messages encoded with a commercial or other nonsecret code; and to give addi-

tional security for highly classified communications. Transposition methods involve either transposing the elements of each code group according to a fixed permutation or transposing the text of a code message by one of the transposition methods already described. Substitution methods may involve monoalphabetic, digraphic, or polyalphabetic encipherments, such as those discussed above. A favourite method with two-part codes is that in which the code group standing at a prearranged interval in the encoding section before or after the code group representing the actual word or phrase intended to be conveyed is substituted; the interval may remain fixed within a single message, or it may vary according to some predetermined key. The most important of the various methods of enciphering code are arithmetical methods. If the code groups are numerical, the addition (usually noncarrying or mod 10, so that, for example, $8 + 7 = 5$) of an arbitrarily selected number (called an additive) to each code group in the code message constitutes a simple form of encipherment. In decipherment the additive is merely removed from the received enciphered code groups by subtraction, leaving the plain code groups, which can then be decoded with the code book. Instead of a fixed additive, the additive may be a sequence of digits used in the manner of a repeating key, the sequence being either agreed upon in advance or derivable from a literal key or by some other method. Additives may also be compiled in specially prepared key tables or key books, in which case an indicator would be necessary to designate the starting point of the key used.

Cipher devices and cipher machines. These range in complexity from simple rotating concentric disks such as that illustrated in Figure 2 to large machines and electrically operated cipher teleprinters. The Wheatstone cipher device, shown in Figure 3, incorporates a very

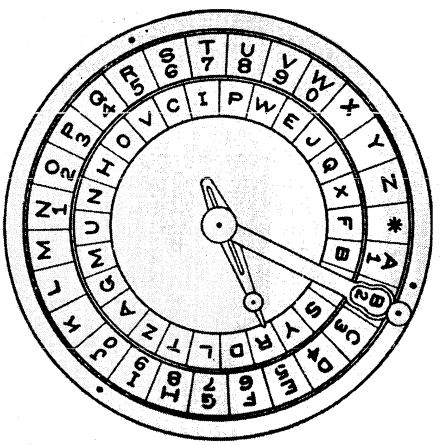


Figure 3: Wheatstone cipher device.

Geared disk devices

interesting principle. Although invented by the English physicist Sir Charles Wheatstone in 1867, a device identical in principle was constructed by an American, Decius Wadsworth, in 1817. Wheatstone's device incorporates two concentric circles of letters, a 27-character plain component including a word spacer, and a 26-character cipher component. The two hands, which may be set independently, are geared together so that, when the long hand (indicating the plaintext) makes a complete revolution, the short hand makes 1 1/26 revolutions, causing the short hand to point one letter in advance of where it pointed at the end of the preceding revolution of the long hand. In encipherment, the two hands are first set to prearranged initial positions and the long hand moved clockwise to the first plaintext letter; the short hand will then point to the cipher equivalent. The long hand is now moved clockwise to the second plaintext letter, and so on for each letter of the message; a word spacer is enciphered at the end of every word. An added convention is that, whenever a doubled letter is to be enciphered, a rare letter such as X_p or Q_p is substituted for the second letter, so that the word ARTILLERY, for example, would be spelled ARTILQERY. In decipherment, after the hands are set to their proper initial positions, the long hand is rotated clockwise so that the short hand points to each succeeding cipher letter, the long hand in each case indicating the plaintext equivalent. (The principle of the Wheatstone device can be duplicated with a pair of sliding alphabet strips—a 27-character plain component and a double-length cipher component. After the two strips are set to their initial prearranged juxtaposition, the cipher equivalent is found for the first plaintext letter. The cipher component strip is moved one position to the left for succeeding encipherments whenever a letter on the plain component is located to the left of the immediately preceding plaintext letter of the text.) The Wheatstone device is particularly interesting in that the motion of the cipher component is highly irregular and unpredictable, depending as it does upon the particular sequence of the letters of the plaintext and upon the particular arrangement of the letters in the plain-component sequence.

In 1891 the French cryptologist Étienne Bazeries invented his "cylindrical cryptograph," shown in Figure 4, based on principles enumerated earlier by Thomas Jefferson of the U.S. The Bazeries device consists of 20 numbered rotatable disks, each bearing a different alpha-

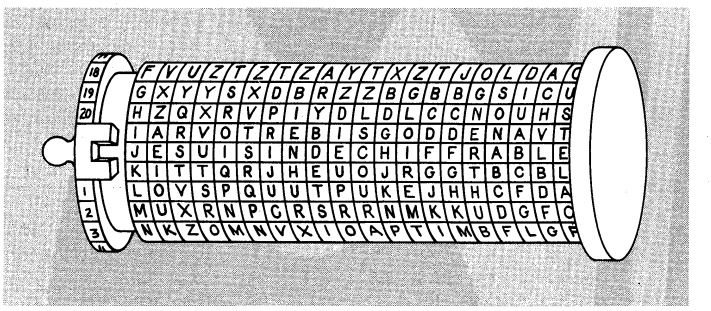


Figure 4: The Bazeries cylinder.

bet engraved on its periphery. The disks are arranged in an agreed-upon order on a shaft and rotated so that the first 20 letters of the message plaintext appear in a row; the ciphertext is then formed by arbitrarily taking off any other row. The remaining letters of the message are treated in the same fashion, 20 letters at a time. In decipherment, the disks are first arranged on the shaft according to the agreed-upon order, the first 20 cipher letters are set up across one row, and the other rows searched for the one and only row that contains plaintext all the way across; and so on for the rest of the decipherment. The Bazeries principle was embodied in a 25-disk cipher device, Type M-94, widely employed in the United States military services from 1923 until 1942.

Mechanical cipher machines

In 1924 Alexander von Kryha of Germany invented a spring-operated polyalphabetic cipher machine having as its principle the irregular displacement of two concentric disks representing the plain- and cipher-components, easily changeable by means of tab inserts for the letters. The displacement of the alphabets occurring after each encipherment is accomplished through a selector wheel having on its periphery 17 toothed sectors consisting of from one to six teeth, each the sectors being designated by the numbers 1–17. These teeth serve to displace the components a distance equivalent to the number of teeth in the sector; because of the manner of spacing between the toothed portions of the wheel, however, an additional displacement of four positions is added at each operation of the machine. The selector wheel has the following effective displacements or skips between its 17 numbered positions:

Pos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	1
Skips:	7	6	7	5	6	7	6	8	6	10	5	6	5	7	6	5	9	

If the components were direct standard alphabets, for example, set initially so that A in plaintext is A in cipher (and therefore a key of A) and the selector wheel is set to position 1, the second key would be A + 7 or the letter H; the third H + 6 or the letter N; the fourth N + 7

or the letter U; and so on. A subsequent model of the Kryha machine incorporates a selector wheel with 52 adjustable screws, or stops, each screw effecting a new displacement of the alphabets.

A compact, tape-printing, mechanical cipher machine invented by the Swedish engineer Boris Hagelin in the 1930s employs a complex mechanism to generate a long running key that is used in conjunction with reversed standard alphabets. In operation, the machine in effect subtracts each plaintext letter from the key to yield the cipher letter, and subtracts each cipher letter from the key to yield the plaintext letter; because of this, the machine has sometimes been called a "letter-subtractor machine." The machine has six wheels of identical diameters, with individual periods of 26, 25, 23, 21, 19, and 17. Equidistant around the peripheries of the wheels are engraved the following sequences:

"26 wheel": ABCDEFGHIJKLMNOPQRSTUVWXYZ
"25 wheel": ABCDEFGHIJKLMNOPQRSTUVXYZ
"23 wheel": ABCDEFGHIJKLMNOPQRSTUVX
"21 wheel": ABCDEFGHIJKLMNOPQRSTU
"19 wheel": ABCDEFGHIJKLMNOPQRS
"17 wheel": ABCDEFGHIJKLMNOPQ

At each lettered position near the edge of the wheel is a small pin that may be set in an active or inactive position. Each operation moves the six wheels one step; if they are initially aligned at AAAAAA, the second alignment will be BBBBBB, the 18th RRRRRA, and the 27th ABDFHJ. Since the wheels are relatively prime (*i.e.*, none of the numbers have a factor in common) to each other, the cycle of the machine is the product ($26 \times 25 \times 23 \times 21 \times 19 \times 17$) or 101,405,850; in other words, the wheels will not return to their initial position until after this number of letters has been enciphered.

Just behind the six wheels is a revolving drum, something like a squirrel cage, composed of two circular retaining plates holding 27 horizontal bars; on each are two lugs, one or both of which may be set at two of six effective positions (corresponding to the six wheels) on the bar or to neutral positions. When in the active position, the pins on a wheel engage those lugs that have been set opposite that wheel, causing the particular bars to be displaced; the displaced bars act as teeth of a gear wheel, shifting the reversed standard alphabets a corresponding number of positions. The number of lugs in the path of a particular wheel is known as the kick of that wheel; the total key is the sum of all the kicks contributed at a given position of the six key wheels.

Figure 5 shows the machine opened and with the inner cover raised to expose the mechanism. This machine was used extensively during World War II by United States forces under the name of Converter M-209 in the Army and CSP 1500 in the Navy.

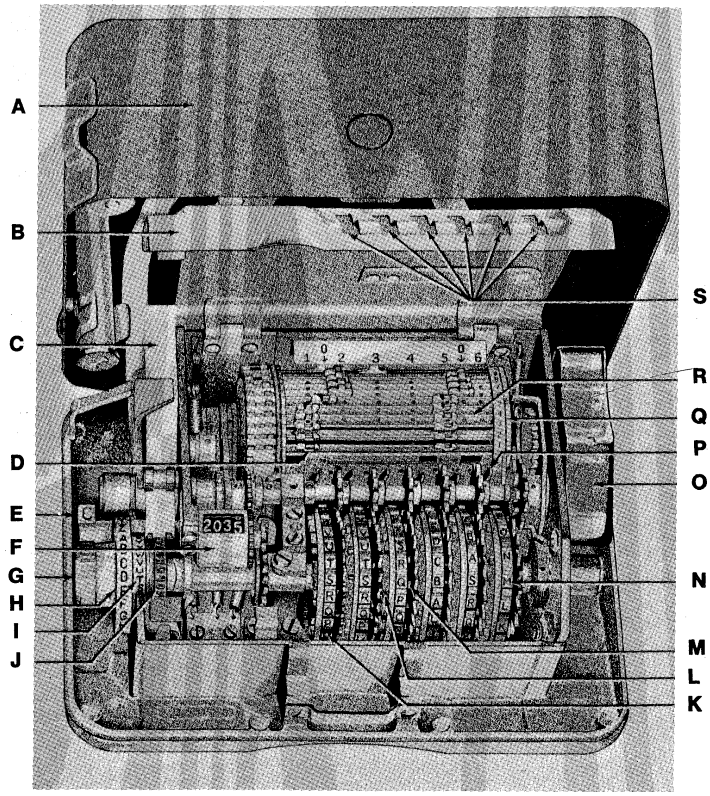


Figure 5: *Converter M-209.*
(A) Outer cover. (B) Inner cover. (C) Paper tape. (D) A lug. (E) Encipher-decipher knob. (F) Letter counter. (G) Setting knob, which simultaneously moves H, I, and J. (H) Indicating disk. (I) Reproducing disk. (J) Type wheel. (K) One of the key wheels. (L) Pin set to "inactive" position. (M) Pin set to "active" position. (N) Row of letters that will be at the bench mark when the inner cover is closed. (O) Drive knob. (P) A key-wheel lever, in the rear position where it will contact the lugs set opposite its corresponding wheel. (Q) Drum. (R) One of the drum lugs. (S) Key-wheel windows, through which the alignment of the key wheels at the bench mark may be seen when the inner cover is closed.

Electrical cipher machines

Electrical cipher machines often employ components called wired rotors as a means of generating a multiplicity of alphabets. A typical rotor may be visualized as a disk approximately three inches, or 75 millimetres, in diameter and perhaps one-half inch, or 13 millimetres, thick, made of molded plastic or other insulating material. On each of the two faces of the rotor are 26 small metal studs, or contacts, arranged in a circle near its edge. The studs on one face are connected by wires in mixed order to the studs on the opposite face, thus producing a mixed alphabet by means of electrical wiring. The rotors may have the normal A–Z sequence engraved on their peripheries as a means of aligning them at specific positions against a bench mark. Several of these rotors are used side by side in a cipher machine, so the studs of each rotor come into contact with the studs of those adjacent; the entire set of rotors in the machine is contained between two endplates each with a circle of 26 contacts that touch the studs of the rotors placed next to them. In encipherment, pressing a letter of a typewriter keyboard sends a current to one of the points on the input endplate from which it goes through the maze of rotors (each of which contributes the effect of one monoalphabetic substitution) to a point on the output endplate, which is connected either to one of 26 lamps or to a printing mechanism, thus producing the cipher equivalent of the plaintext letter. A three-rotor machine in schematic form is illustrated in Figure 6.

If the rotors never moved, the encipherment would be nothing more than simple monoalphabetic substitution with a mixed alphabet. After encipherment of each letter, however, one or more of the rotors will move one step, in effect generating a succession of different alphabets. The total possible number of alphabets is equal to 26 raised to the *n*th power, *n* being the number of rotors in the maze. Thus, for a three-rotor machine 26^3, or 17,576, different alphabets are possible, and for a five-rotor machine 26^5, or 11,881,376, different alphabets are possible. If the motion of the rotors is regular—*e.g.*, meter-like as in an ordinary counter or odometer—these numbers will

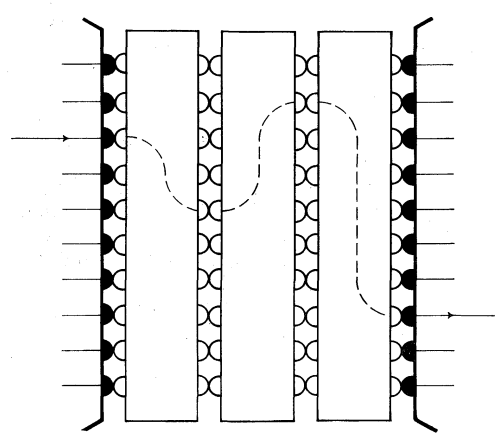


Figure 6: Three-rotor cipher machine (see text).

also represent the periods of the machines. (The rotor which moves one step with each operation is called the fast rotor; it controls the stepping of the medium rotor, which in turn controls the stepping of the slow rotor.) Figure 7A is a diagram of a three-rotor machine, in which the wiring of each rotor is represented by numbered contacts on the left- and right-hand faces; wires connect like-numbered contacts from one side of the rotor to the other. Figure 7B shows the machine after one letter has been enciphered and the fast rotor has moved one position; at this setting, the plaintext letters A, B, and C are enciphered as E, Q and T instead of as A, I and E in the previous setting. Machines for the automatic and simultaneous encipherment and decipherment of the Baudot telegraph code (see TELEGRAPH) used in teleprinter communications are usually constructed along the lines of key generators, and the key that is produced is applied to the plaintext Baudot characters according to the rule of Baudot addition (*e.g.*, where like impulses make a "mark" and unlike impulses a "space"). These machines often employ notched wheels to actuate relays that determine the substitution key for the plaintext Baudot character at each position of the message. One typical machine has as its cryptographic unit an assembly of ten notched wheels mounted on a common shaft so that each wheel steps one position with every operation of the keyboard of the machine. The first wheel has 96 positions, and each wheel thereafter has one more position than the preceding one, so that the tenth wheel has 105 positions. The depression of any key on the teleprinter keyboard actuates the mechanism, which moves the wheels one step each. (The period of this machine is the least common multiple of the wheel sizes, which is $2^5 \times 3^2 \times 5^2 \times 7^2 \times 11 \times 13 \times 97 \times 101 \times 103$, or approximately 8.65×10^{14}.) Behind each wheel is a cam switch, which is acted upon by an irregular arrangement of notchings on its periphery. At 26 of the positions of each wheel, spaced more or less regularly on the circumference, are the letters of the alphabet in normal order, permitting the wheels to be set at designated initial alignments against a bench mark. Each pair of wheels, 1–2, 3–4, . . . 9–10, is used in combination to produce a key stream for the five levels of a Baudot character. The key produced by the five pairs of wheels is applied by Baudot addition to the plaintext character to yield the enciphered character. At each operation of the keyboard, ten new notchings arrive at the sensing position to be combined into a new Baudot character.

Cipher teleprinters

Ciphony, cifax, and civision systems. Methods for the encryption of speech (ciphony), facsimile (cifax), and television (civision) signals employ the same general ideas of substitution and transposition which are to be found in literal cryptosystems, with one important difference. In literal cryptosystems the unit of encryption is usually a single character, whereas in ciphony, cifax, or civision, the corresponding unit is a timed portion of the continuously varying audio or image-scanning signal. Ciphony, cifax, and civision systems fall into two broad categories: privacy systems and security systems. The security afforded by a privacy system is minimal; it offers protection only against direct listening or direct viewing. Security systems, on the other hand, offer maximum—which in some cases might be absolute—protection against analysis.

Privacy and security systems

Speech privacy systems are those systems that operate directly on the speech itself, in either the frequency dimension or the time dimension, or both. The earliest speech privacy systems incorporated the principle of frequency inversion, such as is found in certain commercial transatlantic radiotelephone systems. This principle involves modulating the audio signal with a fixed high audio frequency, so that the resultant signal is the difference in frequency between the two component signals; this has the effect of changing high frequency sounds into lower frequency sounds, and vice versa. Systems were then developed in which several frequencies are used in a changing pattern for the modulation or in which the modulating signal is continuously changing within a frequency band. Also widely used in commercial ciphony equipments is band splitting: the entire speech band is split into several smaller bands, which are then shifted in the frequency spectrum. Both of the foregoing systems employ substitution; an example of transposition is time-division scrambling. In this system, speech can be recorded on a magnetic tape, and the transcription from the tape "read" in an irregular manner.

In speech security systems today the continuous speech wave is first converted into a series of binary impulses (similar to the impulses of the Baudot code) and these "binary digits" are then combined with a binary key stream (derived from a suitable key generator). Three principal methods by which speech is converted into binary form involve (1) pulse-code modulation, (2) delta modulation, or (3) a vocoder. Each converts the continuous speech wave into a digital approximation of the original wave.

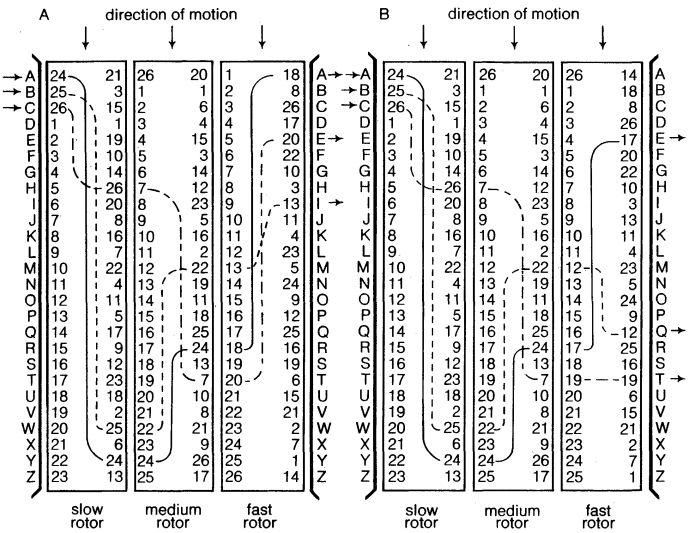


Figure 7: Three-rotor machine with wiring represented by numbered contacts. (A) Position as enciphering is begun. (B) Position after fast rotor has moved one position.

In pulse-code modulation, the amplitude of the speech wave is sampled at a rate equal to at least twice the highest frequency component in the wave; each sampled amplitude is then approximated by categorizing it as one of a number of discrete amplitude levels, represented by binary coding. In a typical system, the speech is sampled 6,000 times per second, and each sample is identified with one of 16 possible amplitude levels, represented by a four-element binary pattern; the output of this pulse-code modulation system is a stream of binary signals at the rate of 24,000 "bits" (*i.e.*, binary digits) per second.

In delta modulation, an approximating or synthetic wave is generated by comparing the amplitudes of the speech wave and the wave being generated. If the amplitude of the generated wave at the instant of sampling is less than that of the speech wave, the next fragment of the approximating wave is generated for a fixed interval (until the next sampling instant) with fixed constant positive slope; if the amplitude of the generated wave is more than that of the speech wave, the next fragment of the approximating wave is generated with a fixed constant negative slope—thus the generated wave being synthesized is constantly correcting itself, in zigzag fashion, so as to produce an approximation to the original speech wave. The positive and negative slope information is represented by binary digits; good quality speech can be obtained by the delta modulation process using approximately 25,000 comparisons per second.

The vocoder is a speech analyzer and synthesizer developed at Bell Telephone Laboratories Inc.; in a vocoder system the spectrum of the speech wave is separated into a number of nonoverlapping frequency bands, and the amplitude in each band is sampled and coded by the pulse-code modulation process. At the receiving end, a synthesizer creates artificial voice sounds from the in-

coming digital information (see TELEPHONE AND TELECOMMUNICATIONS SYSTEMS).

The earliest facsimile privacy system, patented in 1928, involved variations in the velocity of the synchronized sending and receiving drums of the facsimile equipment; exact synchronism of the two drums is essential for proper transmission. Other facsimile privacy systems include pretransmission scrambling (in which the plaintext copy is optically disarranged before transmission) and time-division scrambling similar to that employed for ciphony purposes. In facsimile security systems, the plaintext copy is optically scanned and converted into a stream of binary digital signals; each square inch of the image is converted into approximately 10,000 binary digits of black–white information. A binary key is added to the binary intermediate plaintext; at the receiving end, the same key is added to the incoming cipher signals to produce the original plaintext copy.

Television privacy systems may use adaptations of any of the methods of speech privacy systems. Television security systems likewise employ the same general techniques that are found in secure speech and facsimile systems; but because of the broadness of the frequency spectrum necessary to transmit the secure digitalized television signals and the very high speeds necessary in the key generation for encrypting them, major difficulties are encountered in the development of these systems.

SIGNAL INTELLIGENCE

The principal components of signal intelligence are: communication intelligence, derived from the interception and analysis of messages or communications between persons, and therefore termed communications signals; and electronic intelligence, derived from the interception and analysis of electromagnetic radiations considered to be noncommunications signals, such as radar, identification and recognition signals, navigational beacons, and telemetry signals. In practice, it is often difficult to distinguish between these two types of intelligence.

Communication intelligence. Communication intelligence is evaluated information concerning an enemy, derived from a study of his signal communications. The principal components of communication intelligence are: interception of signals or messages and forwarding raw traffic to communication intelligence centres for study; traffic analysis, including radio direction finding and position finding (evaluated information from this source is called traffic intelligence); cryptanalysis (and translation, when necessary) of message texts; and evaluation of data, that is, analysis of results obtained from the preceding steps and their correlation, collation, and comparison with information from other sources.

Traffic analysis. A great deal of information of military value can be obtained from signal communications even without solving encrypted traffic. In general terms, traffic analysis is the careful inspection and study of signal communications to penetrate camouflage superimposed upon the communication network for purposes of security. Specifically, traffic analysis reconstructs radio communication networks by the following: noting volume, direction, and routing of messages; correlating transmission frequencies and schedules used among and within the various networks; determining directions in which transmitters lie, by means of radio direction finding; locating transmitters geographically, by radio position finding; determining the system of assigning and changing radio call signs; and studying all items that constitute messages originated by operators and exchanged among themselves on a radio net.

By means of these procedures, traffic analysis is able not only to ascertain the location and disposition of military units (technically called "order of battle") and important troop movements but also to predict with a fair degree of reliability the areas and extent of immediately pending or future activities. Through traffic analysis a nation can obtain information of value concerning an enemy and can determine what information concerning his own forces is made available to the enemy through his own signal communications. Unit movements and preparations for military activity, for example, may be indicated by rising and falling traffic volumes and changes in the structure of the network; or the military function of a network may be revealed by the characteristic traffic pattern that results from transmissions incidental to planning, supply, and transportation; or a change of grouping, disposition of forces and fleets, and probable tactical developments may be manifested in the redeployment of the radio stations that serve military elements. These very important results are obtained without actually reading the texts of the intercepted messages. The cryptanalyst, on the other hand, is frequently able to make good use of bits of information disclosed by traffic analysis, such as faults noted in message routing and errors in cryptography that cause messages to be duplicated or cancelled. Cryptanalysis can provide important information for traffic analysis, since the solution of messages often yields data on such matters as impending changes in signal communication plans and operating frequencies and schedules. Cryptanalysis also yields data on specific channels, networks, or circuits that are most productive of intelligence.

Cryptanalysis. The solution of secret communications without any previous knowledge of the system or the key is the domain of cryptanalysis, requiring, in the case of modern cryptosystems, extensive theoretical study, unusual powers of observation, inductive and deductive reasoning, great concentration and perseverance, and a vivid imagination guided by good judgment, fused together by a special aptitude and intuition gained from long and varied practical experience. An isolated, short cryptogram may resist solution indefinitely, even if it is in a fairly simple system; but a large volume of material—even in a very complex cryptosystem—may well be solved with time and effort.

Qualifications of the cryptanalyst

In general, the art of cryptanalysis may be reduced to three basic steps: arrangement and rearrangement of data to disclose nonrandom characteristics or manifestations (*e.g.*, in frequency counts, repetitions, patterns, and symmetrical phenomena); recognition of the nonrandom characteristics or manifestations when disclosed; and explanation of the nonrandom characteristics when recognized. The requirements for the first step are experience or ingenuity and time—which may be appreciably reduced by the use of machine aids; for the second step, experience or statistics; and for the third step, experience or imagination, and intelligence.

The cryptanalyst studies all the messages in the cryptosystem under investigation and the possible relationship with other cryptosystems of the same enemy. All available collateral information concerning the correspondents and their past and present activities is gathered, together with data on the transmission and interception of the traffic. Indicators, if present in the message texts, are given particular attention; traffic is separated into homogeneous groups of messages; a search is made for repetitions within and between messages; the beginnings and endings of messages are scrutinized for possible stereotyped language; the traffic is examined for messages likely to contain the same underlying plaintext but having different cryptographic texts; and various statistical counts are made of single letters, digraphs, trigraphs, tetragraphs, and groups.

The cryptanalyst also arms himself with the salient cryptolinguistic data in the language of the enemy; he has available the relative frequencies of single letters, digraphs, trigraphs, and tetragraphs; and he compiles lists of frequent words, either based on past solutions or hypothesized to be present in the messages. If the language were English, for example, he would know the following relative frequencies in descending order of the plaintext letters (based on a count of 50,000 letters of governmental plaintext telegrams and reduced to 1,000 letters):

E 130	I 74	C 31	Y 19	X 5
T 92	S 61	F 28	G 16	Q 3
N 79	D 42	P 27	W 16	K 3
R 76	L 36	U 26	V 15	J 2
O 75	H 34	M 25	B 10	Z 1
A 74				

The relative frequencies in descending order of the 30 highest plaintext digraphs are the following (based on 50,000 letters and reduced to 5,000 digraphs):

EN 111	IN 75	NE 57	AT 47	CO 41
RE 98	TE 71	VE 57	TI 45	IO 41
ER 87	AN 64	ES 54	AR 44	TY 41
NT 82	OR 64	ND 52	EE 42	FO 40
TH 78	ST 63	TO 50	RT 41	FI 39
ON 77	ED 60	SE 49	AS 41	RA 39

The absolute frequencies of the 25 highest trigraphs in the sample of 50,000 letters are the following (note the prevalence of fragments of numbers, since the messages studied had a high incidence of spelled-out numbers in their plaintexts):

ENT 569	TIO 221	NIN 207	VEN 190	NTH 171
ION 260	FOR 218	STO 202	EVE 177	TWE 170
AND 228	OUR 211	EEN 196	EST 176	TWO 163
ING 226	THI 211	GHT 196	TEE 174	ATI 160
IVE 225	ONE 210	INE 192	TOP 174	THR 158

Similarly, the 25 highest frequency tetragraphs are shown below, accompanied by their absolute frequencies in the sample of 50,000 letters:

TION 218	WENT 153	IGHT 140	SEVE 121	REQU 98
EVEN 168	NINE 153	FIVE 135	ENTH 114	HIRT 97
TEEN 163	TWEN 152	HREE 134	MENT 111	COMM 93
ENTY 161	THRE 149	DASH 132	THIR 104	QUES 87
STOP 154	FOUR 144	EIGH 132	EENT 102	UEST 87

(Again note the effect of the prevalence of numbers in the plaintext, together with high-frequency words such as STOP, REQUEST, and COMMA.) In addition to letter-frequency data, the cryptanalyst has available lists of words arranged by word length in alphabetical and in rhyming order, lists of frequent word combinations, and lists of idiomorphs (word patterns). The idiomorphs are sorted by pattern, so that, for example, under the pattern *abaca* would be listed D*IVISI*ON and *LOCOMO*TIVE, and under the pattern *abba* would be listed B*ATTA*LION and SH*IPPI*NG.

Idiomorphs or word patterns

Transposition ciphers, recognized from the fact that the distribution of cipher letters will approximate that of plaintext, are solved by experimenting with matrices of various types, by partial anagramming of portions of the ciphertext, and by assuming probable words. Messages involving cryptographic errors and their subsequent correction are often exploitable and therefore are particularly valuable to the cryptanalyst.

In substitution ciphers, the characteristic relative frequencies of single letters, digraphs, and longer polygraphs serve as a basis for the assignment of plaintext equivalents to cipher values, but only when the cipher has been reduced to simplest terms, such as a monoalphabetic substitution cipher. In the final analysis, the solution of every cryptogram involving a form of substitution depends upon its reduction to monoalphabetic terms, if it is not originally in those terms. In general, the solution of periodic polyalphabetic substitution ciphers depends upon three main steps: determining the number of alphabets involved; allocating the letters of the ciphertext into the cipher alphabets to which they belong; and analyzing the individual monoalphabetic distributions to determine plaintext values of the cipher letters in each distribution or alphabet. In the case of aperiodic substitution systems or those employing very long keys, the first step above might be either impossible or very difficult; the only recourse might be to superimpose enough messages on the basis of repetitions so that the letters in the columns of the resulting superimposition diagram are monoalphabetically distributed and thus solvable. The solution of codes is primarily a linguistic problem in which an expert knowledge of the language is of paramount importance; a considerable number of messages might be necessary for solution, especially if a large two-part code is involved.

The principles and techniques of cryptanalysis apply in the decipherment or reconstruction of lost languages; for example, in the solution of the cuneiform script by Georg Friedrich Grotefend in 1802 and Sir Henry Creswicke Rawlinson in 1838, the decipherment of the Egyptian hieroglyphs by Jean-François Champollion in 1822, and in the solution of Minoan Linear B by Michael Ventris in 1952. The ultimate test may come when a cryptanalyst is faced with a communication from an extraterrestrial intelligence, and the world awaits his solution.

HISTORY

Because the official cryptologic work of all governments has always been kept secret, it is often impossible to ascertain the origin of cryptographic improvements or the development of cryptanalytic techniques, particularly those of modern times; these enter the public domain years after the fact, if at all. The classic cryptographic systems and elementary cryptanalytic techniques are described in the open literature, but the gap between the advent of a new cryptologic technique and its public revelation is an ever-widening one. It is incumbent upon cryptologic historians to trace developments as far back as open sources permit, knowing that in many cases the original attribution or date may be in error. Details of important and far-reaching cryptographic and cryptanalytic developments since 1940 are still shrouded in official secrecy and understandably so since the security of national communications and the continuance of successful intelligence efforts are adversely affected by public revelations.

The development of cryptography. The earliest cryptography was developed by the Egyptians, the ancient Hebrews, and the Indians, but the extent of its use is unclear. The first recorded use of cryptography for correspondence was by the Lacedaemonians, or Spartans, who at least as early as 400 BC employed a cipher device called the scytale for secret communications between military commanders. This device consisted of a tapered baton, around which was spirally wrapped a piece of parchment or leather, which was then inscribed with the message; when unwrapped, the parchment bore an incomprehensible set of letters, but when wrapped around another baton of identical proportions, the original text reappeared. Thus, the Greeks were inventors of the first transposition system. Aeneas Tacticus (4th century BC) wrote a work entitled *On the Defense of Fortifications*, one chapter of which was devoted to cryptography, making it the earliest treatise on the subject. The Romans used monoalphabetic substitution with direct standard alphabets: Julius Caesar used the setting in which plaintext A was enciphered as D, while Augustus Caesar used that in which plaintext A was enciphered as B.

The scytale

Modern cryptography and the first systematic cryptocorrespondence began around AD 1200 in the Papal States and the Italian city-states: the first ciphers involved vowel substitution (leaving consonants unchanged), and code symbols were introduced. The first manual on cryptography (*c.* 1379) was a compilation of ciphers by Gabriele de Lavinde of Parma, who served Clement VII; this manual, now in the Vatican archives, contains a set of keys for 24 correspondents and embraces symbols for letters, nulls, and several two-character code equivalents for words and names. The first brief code vocabularies (known as nomenclators) were gradually expanded into what were called repertories, and, later, codes. By 1400, variants for vowels were introduced, followed by a more extensive use of variants for all the plaintext letters, in addition to specific cipher equivalents for common plaintext digraphs, syllables, and words. By the middle of the 15th century, nomenclators were refined and became the mainstay for several centuries for diplomatic communications of nearly all European governments. Leon Battista Alberti published in 1470 his *Trattati in cifra* in which he described the first cipher disk, prescribing that the setting of the disk should be changed after enciphering three or four words, thus conceiving of the notion of polyalphabeticity; Alberti also used his disk to effect an encipherment of a small code of 336 groups, a concept four centuries ahead of his time. Further advances were made by the Abbot Johannes Trithemius (1462–1516) in his *Polygraphia* (the first printed book on cryptology), which contains the first description of a square table for polyalphabetic encipherment. Giovanni Battista della Porta published in 1563 his *De Furtivis Literarum Notis*, which contains a modified

First cipher disk

form of square table and the earliest example of a digraphic cipher. The *Traicté des chiffres* published in 1586 by Blaise de Vigenère contains the square table commonly attributed to him, as well as descriptions of the first plaintext and ciphertext autokey systems. In the 17th century, two-part codes were used for the first time by France under Louis XIII and Louis XIV, and at the same time Papal repertories introduced an innovation in which many of the cipher characters represented two or more different plaintext letters. Codes were expanded in the next three centuries, so that by 1860 large codes were used for diplomatic communications, and cipher systems were a rarity for this use. Cipher systems prevailed, however, for military communications except for high-command communications. In the early history of the United States, repertories were used; during the Civil War the Federal Army used small codes and a word-transposition cipher, while the Confederate Army used the Vigenère cipher with direct standard alphabets.

In the first two years of World War I, all the belligerents used cipher systems almost exclusively for their tactical communications; code systems were still used in the main for high-command and diplomatic communications, but by 1917 small codes became common for low-level communications. Nevertheless, certain complicated cipher systems were used for high-level communications through the end of the war—the first version of the famous ADFGVX cipher, for example, introduced by the Germans on March 1, 1918. In the postwar period, significant advances were made in cryptography, particularly in the realm of cipher machines. During World War II, cipher machines were used extensively, but not exclusively, by both the Allies and the Axis powers; a plethora of codes and manual cipher systems nevertheless prevailed on all sides. In the years since World War II, even more remarkable advances in cryptography have been made.

Advances in cryptanalysis. The technological advances in cryptography had their counterparts in cryptanalysis. The principles of cryptanalysis were known to the Arabs at the beginning of the 15th century: in 1412 Qalqashandi, a Persian living in Egypt, completed a 14-volume encyclopaedia in which appeared not only the first treatment of both substitution and transposition but also the first exposition of cryptanalysis in history. The use to which the Arabs put this information is, however, unknown. *Trattati in cifra,* mentioned above, contains cryptanalytic theories and procedures. The oldest treatise devoted entirely to cryptanalysis is by Sicco Simonetta, a cryptanalyst in the service of the Sforzas of Milan, who on July 4, 1474, wrote a brief paper setting forth rules of cryptanalytic procedure. Among early cryptanalysts whose documented successes have been preserved are Giovanni Soro (died 1544), Venice's principal cryptanalyst for almost 40 years; Giovanni Batista Argenti and his nephew Matteo Argenti, both of whom were in the papal service in the last half of the 16th century; the French mathematician François Viète, who served Henry IV at the end of the 16th century; Antoine Rossignol, France's greatest cryptanalyst of the 17th century, who served Louis XIII and Louis XIV; and the mathematician John Wallis (1616–1703), who for over five decades was England's greatest cryptanalyst. In 1863 the important work, *Die Geheimschriften und die Dechiffrir-Kunst* by Friedrich W. Kasiski, described for the first time the solution of the Vigenère cipher, hitherto regarded as impregnable, and indeed called "*le chiffre indéchiffrable.*" Nevertheless, in 1757, almost 100 years before Kasiski, a 13-alphabet polyalphabetic cipher was solved by Jacques Casanova, so the general method of solution must have been known in the "black chambers" of the period; it seems doubtful that there were breaches of security, since violations were often treated as capital offenses. In the United States, the first cryptanalysts were the Rev. Samuel West and Elbridge Gerry (later the fifth Vice President), who solved a cryptogram of a Tory spy for George Washington. During the Civil War, Federal Army cryptanalysts were said to have solved every message intercepted, while the Confederate Army had to publish some cipher messages in the newspapers with a request for solution.

First exposition of cryptanalysis

Poor cryptography—or good cryptanalysis—has often dramatically affected the course of history. In World War I, the Russians lost the Battle of Tannenberg, in August 1914, because of the failure of their cryptographic communications. The solution by the British of perhaps the most famous cryptogram in history, the Zimmermann telegram of January 16, 1917 (offering Mexico territorial gains if she would enter the war on the German side), was a major contributing factor to the entry of the United States into the war on April 6, six weeks after the revelation to President Wilson of the contents of the message. In World War II, the Battle of Midway was a stunning victory of intelligence, since cryptanalysis gave the United States full information on the size and location of the Japanese forces advancing on Midway, enabling the Navy to concentrate a numerically inferior force that otherwise might have been 3,000 miles (5,000 kilometres) away, and thus prepare an ambush that proved to be the turning point in the Pacific war. Admiral Isoroku Yamamoto, the commander-in-chief of the Combined Fleet of the Japanese Navy, was shot down when a highly secret message was solved, giving the itinerary of Yamamoto's plane on a tour of bases in the Solomon Islands. But the most spectacular role of cryptanalysis in World War II was revealed after the war in the hearings held by the joint congressional committee on the investigation of the Pearl Harbor attack. At that time it was made public that, shortly before the war, the United States, in a brilliant stroke of cryptanalysis, had been able to reconstruct the Japanese cipher machine that was used for the highest level diplomatic communications, enabling this traffic to be read throughout the war. It was also revealed that during the war, U.S. cryptanalysts had considerable success in solving other Japanese codes and ciphers. This success led the committee to state that all witnesses familiar with the intelligence produced throughout the war "have testified that it contributed enormously to the defeat of the enemy, greatly shortened the war, and saved many thousands of lives."

Intelligence in the Battle of Midway

BIBLIOGRAPHY. The most important and authoritative works on the subject are classified governmental publications, unavailable to the general public. A comprehensive and scholarly bibliography of cryptologic literature covering practically all publicly available books and important papers on the subject is JOSEPH S. GALLAND, *An Historical and Analytical Bibliography of the Literature of Cryptology* (1945). Such a compilation is more helpful to specialists than to the general reader; therefore, a selection of items from the Galland bibliography may be useful. Also, after 1945 a few items of importance were published that should be mentioned. For a detailed history of cryptology, DAVID KAHN, *The Codebreakers* (1967), is outstanding. The best technical works in the public domain are LUIGI SACCO, *Manuale di crittografia,* 3rd ed. (1947); and CHARLES EYRAUD, *Précis de cryptographie moderne* (1953). The best work in English is HELEN F. GAINES, *Cryptanalysis* (1956). Other works are ANDRE LANGE and E.A. SOUDART, *Traité de cryptographie* (1925); MARCEL GIVIERGE, *Cours de cryptographie* (1925); ANDREAS FIGL, *Systeme des Chiffrierens* (1926); ROGER BAUDOUIN, *Éléments de cryptographie* (1939); HERCULES MARTHANS GARRO, *Tratado de criptografía* (1965); ABRAHAM SINKOV, *Elementary Cryptanalysis: A Mathematical Approach* (1968). "Communication with Extraterrestrial Intelligence," *IEEE Spectrum,* 3:156–163 (1966), contains an extensive discussion on the subject by LAMBROS D. CALLIMAHOS from the point of view of cryptology.

(L.D.C.)

Crystallization and Crystal Growth

Crystallization is a general term for the process that results in the formation of crystals—solid materials in which the component atoms or molecules are arranged in a definite pattern. Crystal growth, specifically, is the enlargement of crystals at the expense of materials in contact with them.

Even in amorphous, or noncrystalline, solids strong forces usually hold together the constituent atoms, or molecules, which therefore occupy definite positions relative to one another. Except for slight vibrations about

their normal positions, these component atoms and molecules do not move about. Thus, many amorphous solids, such as glass, have fixed shape, rigidity, and mechanical strength. In crystalline solids, however, the atoms (and molecules) are not only held in place, they also are arranged in a definite order that is constantly repeated throughout the sample.

Crystals and crystallites

In most crystalline solids, small, individually ordered regions, which are called crystallites, are oriented differently from one another, with the result that the crystalline pattern is repeatedly disarranged; such materials are said to be polycrystalline. In certain rare crystals, such as the gemstones, and in crystals man takes great care to prepare, however, a single ordered region makes up the entire solid object; in such cases a single crystal may range in size from fractions of an inch to a few feet. One of the identifying features of single crystals is that they usually have flat planes or faces that are set at fixed angles to one another. Quartz crystals, and other crystals displaying such regular faces, are most commonly seen in mineralogical collections.

Crystals are formed in many natural processes. Snow flakes, for example, are single crystals and, when the cooling is slow, large, single crystals of ice occur in frozen lakes. When cooling is rapid, however, many small crystals, or crystallites, form, and the solid blocks of ice are polycrystalline. Almost all crystalline solids of nature and commerce are polycrystalline rather than single crystals.

Comparatively slow rates of formation are required to produce large single crystals free of crystallite grains. Other more subtle imperfections occur to some extent in all crystals; these may be missing atoms or vacancies in the crystal structure, or foreign atoms present at sites that should be occupied by the atoms of the material located between the atoms of the material.

Dislocations

In addition to the above atomic or "point" defects, crystals may have more extensive "line" defects, in which the disorder or dislocation extends along a line or plane of atoms. These are of two principal kinds: edge dislocations and spiral, or screw, dislocations. Figure 1, top, is a representation of an edge dislocation. The imperfection may be thought of as resulting from the addition of an extra plane of atoms, ABB'A', in the upper half of the crystal. Because of the added atoms in the upper half of the crystal, this portion is strained by compression; correspondingly, the lower half of the crystal is strained by tension as the fewer number of atoms seek to fill the same volume as that filled in the upper part of the crystal. If the dislocation consists of a gradually increasing shift of alignment between planes (as illustrated in Figure 1, bottom), the imperfection is called a screw dislocation. Dislocations are created when existing crystals are deformed or when new crystals are growing. In fact, dislocations are important in growth mechanisms. Under careful control several materials, including silicon (Si), germanium (Ge), gallium arsenide (GaAs), and quartz (silica, SiO_2), have been grown in the form of crystals free of dislocation, but theoretical considerations require that all crystals have vacancies except at absolute zero temperature (approximately —460° F, —273° C). Indeed, the more rapidly a crystal is grown and the less pure its growth environment, the greater the number and variety of imperfections it acquires. In relatively perfect crystals the misorientation between adjacent grains may be undetectable —that is, the only imperfections are point defects and dislocations—but as the conditions of formation become less favourable, misorientations increase to the point where a polycrystalline mass is the product.

Single crystals

Single crystals are important in science because measurements of the fundamental properties of matter in such a highly ordered state are interpreted more easily than when irregularities in structure are present. Impurities that often segregate at the grain boundaries in polycrystalline materials may mask such subtle properties as semiconductivity of electricity. Properties of crystals that depend on the direction through which the crystal measurement is made (such as refraction of light) either are not exhibited in polycrystalline material or are much harder to interpret. Partly for this reason, single crystals are of especial interest in electronics. Boundaries between grains in polycrystals scatter light to produce opacity, thereby generally making polycrystalline materials unsuitable for applications requiring optical transmission. Similarly, the laser and related devices have created new demands for single crystals for research and applications (see LASER AND MASER). Devices that depend on semiconductivity, such as transistors, utilize single crystals, as do devices such as piezoelectric oscillators and filters, which depend on unsymmetrical distribution of properties (anisotropy) with regard to the crystal structure. Indeed, it is fair to say that the electronics industry could not exist without single crystals. In addition, of course, the intrinsic beauty of natural crystals has fascinated man from prehistoric times to the present and the manufacture of synthetic gemstones is an expanding area of crystal technology.

From (Top) R. Laudise, *The Growth of Single Crystals* (© 1970 Prentice Hall, Inc.; (Bottom) W. Read, Jr., *Dislocation of Crystal* (Copyright 1953), used with permission of McGraw-Hill Book Compan

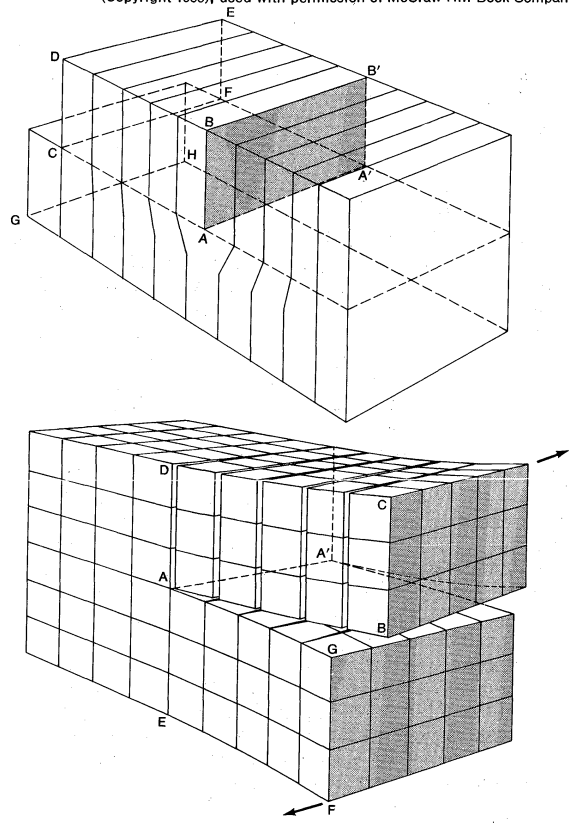


Figure 1: (Top) Edge dislocation. The planes of atoms in a single crystal are represented as vertical layers. AA' is the dislocation line and the dislocation can be produced by pushing (shearing) on CDEF while pulling on GCFH. (Bottom) Screw or spiral dislocations. AA' is the dislocation line. The dislocation can be thought of as having been created by pushing ABCD while pulling AEFG.

Disturbing of equilibrium

All crystals are grown by disturbing a balance, or equilibrium, that exists between the crystalline material and the gas, liquid, or amorphous solid that exists in contact with it. Under the conditions of the equilibrium, the relative amounts of the crystal and the material in contact with it do not change. Disturbing the equilibrium by altering the conditions, such as lowering the temperature in the case of an equilibrium between water and ice, may result in an increase in amount of crystal with respect to the noncrystalline material.

The material that is crystallizing may be the only substance present in the growth system or other substances may be added to lower the temperature of the system or otherwise alter its properties.

Crystal growth from pure, or monocomponent, systems presents several important advantages and is widely practiced. The purity of the crystals obtained is high, because

no additional components are present to be incorporated during the growth process. Furthermore, the speed of growth can be relatively rapid because, in the absence of additional components, slow movement of the additional components away from the growing crystal surface is not required.

The most common monocomponent growth process is growth by the freezing of a pure liquid, or melt, of the material in question. In contrast to most other growth processes, freezing of a pure melt is uncomplicated and readily controllable; it is probably the most widely practiced commercial growth process. Freezing is also the easiest process in which to consider the distribution of undesired impurities and impurities deliberately added to impart desired properties (dopants) to the crystal. The most important consideration regarding distribution of impurities is whether the crystallization process is conservative or nonconservative. In a conservative process no material is added to or taken from the melt except, of course, for the crystalline material that separates as a result of the freezing process. In a nonconservative process, however, additional material may be melted into the liquid during crystallization, or material may be lost from the melt as, for instance, by evaporation during the process. (For discussion of actual growth methods, see below *Preparation of crystals.*)

Conservative and nonconservative systems

THEORETICAL ASPECTS OF CRYSTAL GROWTH

Extensive theoretical work on crystal growth has been conducted, chiefly in the 20th century. One of the most important considerations seems to be the energy released by the attachment of atoms at different sites on a crystal face.

Energy considerations. When a new atom is added to a growing crystal, the most probable position for addition will be one that is energetically most favourable. Since energy generally is liberated during bond formation, the most favourable position for adding a new atom will be at the point where the greatest number of bonds to adjacent atoms can be formed, thus providing the most potential energy. Figure 2 may be considered as an example of

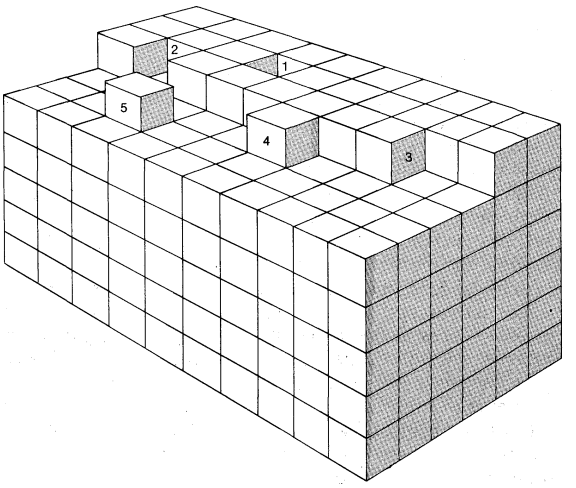


Figure 2: Positions for the addition of new atoms at the surface of a growing crystal (see text).

a developing crystal structure in which the atoms are identical and are represented as cubes. In this structure, the atoms are closely packed so that each has six nearest neighbours, which touch it on the faces of the cube. In this example, an atom is considered to bond only to its nearest neighbours. The most favourable position for the addition of a new atom, therefore, is at a "hole" in the crystal structure, such as position (1), at which the atom will be surrounded by neighbouring atoms on all sides but one. The next most favourable position is at a "notch" in one step of the crystal structure, such as position (2), in which location the new atom will find four neighbours. Growth at either of these positions, of course, cannot be repeated. The next most favourable position for the addition of a new atom is at a "kink" in a step, such as position (3), at which the new atom has three neighbours. In this case, the addition process can be repeated until the step in the structure is completed. The next most favourable position for addition is at a jog at the front of a step, position (4), and the least favourable position for addition, based on nearest neighbour considerations, is at a lone site on the surface, position (5).

Complications in the above picture arise in the case of ionic crystals, which are composed of ions carrying opposite electrical charges because, when an ion of a given charge approaches the face, it is attracted by ions of opposite charge while at the same time it is repelled by ions of like charge. Similar considerations apply to crystals composed of molecules bearing separate regions of positive and negative charge.

Ionic crystals

The problem of how crystallization begins on an already formed atomically smooth crystal surface is one that has received great attention. As seen in the discussion of Figure 2, the implantation of the first unit of a crystal, a process called nucleation, is energetically less favourable at a new step than at a kink in a step already partially formed. Once a step has grown across the surface, however, nucleation of a new step is required. Quantitative calculations indicate that either large supercooling (cooling far below the melting point) or supersaturation (addition of more of the crystallizable solid to a solution than it can hold) is required for the nucleation of a new step on an atomically smooth surface. Nevertheless, crystals are observed to grow at conditions of low supercooling or low supersaturation, contrary to what might be expected. What is obviously required in such cases is a step in the crystal structure that grows without coming to an end; that is, indefinitely. It was only recently realized that spiral dislocations, of the type pictured in Figure 1 (bottom), supplied locations at which growth could continue without completing the regular crystal structure. Careful observation of crystals grown under different conditions provided an understanding of how such growth took place.

The above energy considerations generally apply to circumstances in which a crystal is grown extremely slowly, so that new material is deposited on the crystal surface strictly in terms of the energy requirements of the system. In most cases, however, the growth of crystals is too rapid to achieve equilibrium, or balance, between the rate of addition to the crystal and rate of removal from it, so that the observed shape of crystals usually depends on the growth rate of the various individual faces. In the 1960s it was shown that the shape, or morphology, of the crystal could be correlated with a property known as the entropy of fusion, entropy being a measure of the loss of energy from an organized system, resulting in an increase of disorder. In particular, it was shown that the entropy of fusion could be used to distinguish four classes of materials: (1) in which growth is equivalent in all directions (isotropic) and cells and dendrites (see below) occur when impurities are added; (2) in which growth occurs as facets, or as facetted dendrites, when sufficient impurity is added; (3) in which facets are always observed; (4) in which growth of spherulites is usual. Cells, or bumps on the surface, result from an instability caused by supercooling. These produce a fairly regular pattern of segregation of impurities within the crystal. Dendrites are crystals with a branched, treelike morphology resulting from growth that is controlled by the rate of diffusion. Facets are flat faces parallel to the crystal planes. Spherulites are oriented spherical masses of crystals. Figure 3 illustrates some of these structures.

Classification by entropy of fusion

Constitutional supercooling. Supercooling that results from concentration changes within the crystallizing system is called constitutional supercooling. Constitutional supercooling is important in the crystallization of all polycomponent systems and has been much studied, especially in melt systems. When growth takes place in the presence of a very high concentration of impurities, or in a polycomponent system, there is a buildup of the components that are rejected from the growing crystal interface. Crystallization, therefore, depends on the rate of

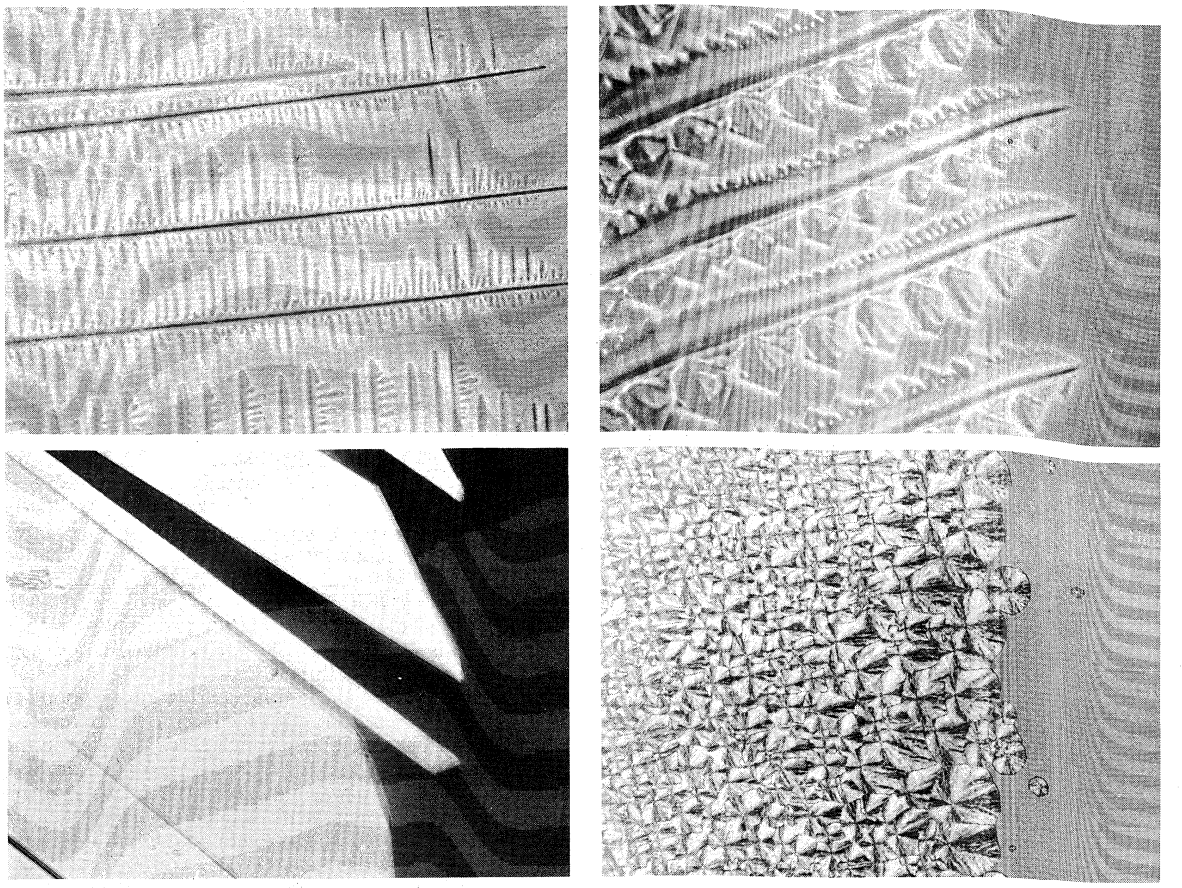


Figure 3: Observed morphologies of crystals as related to the values of the entropy of fusion. As the entropy of fusion increases, it is possible to observe (top left) dendrites (carbon tetrabromide), (top right) facetted dendrites (*tert*-butyl alcohol), (bottom left) facets (benzil), and (bottom right) spherulites (tristearin).
By courtesy of Bell Telephone Laboratories

diffusion of the impurities away from the growing crystal surface. If crystal growth is so slow as to be almost at equilibrium, diffusion has ample time to remove impurities, and the effect of diffusion, therefore, is negligible. In many cases, however, the diffusion problem is severe. Figure 4 shows the melting points of the compositions ex-

From R. Laudise, *The Growth of Single Crystals* (© 1970); Prentice Hall, Inc.

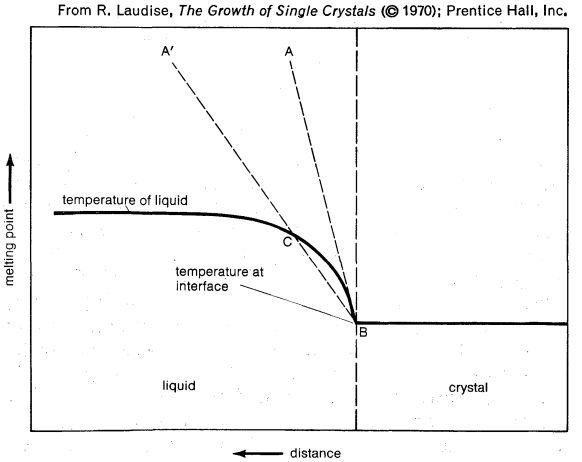


Figure 4: Relationship between melting point and distance in front of a growing crystal (see text).

isting near a growing crystal as a function of distance (in a system in which diffusion is important). As can be seen, the melting point of the solution decreases as the interface is approached. The broken lines AB and A'B are two different possible temperature gradients in the solution. For the larger gradient, AB, at all positions in front of the interface the temperature is above the melting point (*i.e.*, there is no supercooling); for the smaller gradient A'B, the region CB is below the melting point; it is supercooled, and crystallization tends to take place in front of the solid interface. The region CB is referred to as constitutionally supercooled, and large temperature gradients, such as can be obtained easily in crystal pulling (see below *Preparation of crystals*), lessen the tendency toward constitutional supercooling. In a crystal growing from a solution in which the liquid is maintained at uniform temperature, supercooling is least near the crystal and increases farther from it. Supercooling should lead to instability of the crystal surface.

A crystal face that has a very slow intrinsic growth rate is not as susceptible to instability as are more rapidly growing faces of the same crystal. The crystal faces that become unstable under these conditions are the fastest growing faces and, as a result, they break up into small segments of more slowly growing faces. Under such conditions, the peaks of some of the facets will be in regions where the supercooling, or supersaturation, is greater than the valleys. The peaks then grow ever more rapidly, with a tendency toward extreme dendritic growth.

The behaviour of eutectics. A eutectic is a mixture of two phases that freeze together in a fixed ratio and at a constant temperature. A great variety of structures can be found among eutectics; the common feature, however, is that two phases can always be seen. These two phases usually form a structure in which they alternate by growing around and over each other, starting from a single pair of nuclei. The structures can be broadly classified as regular or irregular. The former includes both layered, or lamellar, structures and rod types, in which one phase is rod-shaped and is imbedded in a continuous matrix of the other. The irregular structures include a variety of complicated intertwinings of the two phases, as well as discontinuous eutectics, in which one phase consists of isolated crystals imbedded in a continuous matrix of the other. If the intrinsic growth rate of the two phases is iso-

Lamellar growth

tropic, eutectics can crystallize in the lamellar form with regular spacing between the lamellae. The boundaries between the lamellae are parallel to the growth direction; that is, lamellae of both phases crystallize at the same time. The lamellar widths are determined by the interaction of two opposing considerations: (1) diffusion of solids must occur over great distances to form thick lamellae, a factor that favours narrow lamellae, and (2) the narrower the lamellae, the more surface energy that is associated with boundaries between the two phases, a factor that favours wide lamellae. Rods occur whenever the volume fraction of one of the phases is small. Many complex and irregular eutectic structures result from unevenly directed growth rates of the two phases.

Nucleation and impurities. The formation of a crystal in the absence of a seed involves nucleation, or initiation of a crystalline centre. When nucleation occurs on a surface, such as a wall of the container, or on a foreign body, such as a dust particle, it is called heterogeneous nucleation. If the structure and interatomic spacing of the surface on which heterogeneous nucleation takes place approximate that of the crystal, growth on the surface can resemble growth on a normal seed crystal, a phenomenon called epitaxial growth. When nucleation occurs in the absence of a surface—*i.e.*, in the bulk of a liquid—it is said to be homogeneous.

In some cases, nucleation can be looked upon as the aggregation of particles, and precursors of nuclei large enough to grow are formed by association of particles in the system. If the number of particles in the aggregate and the equilibrium constant for the crystallization reaction are large enough, the nucleus will tend to grow instead of redissolving. Aggregates too small to grow in the average supersaturation in the medium are sometimes called embryos. Most embryos never survive to grow. A few do so, however, because of statistical fluctuations of supersaturation or because energy is added locally to assist their growth (as, for example, when a beaker containing a supersaturated solution is scratched with a stirring rod). The higher the supersaturation, the larger the average embryo size and the greater the chance that statistical fluctuations will allow nucleation. Given enough time, there is a statistical chance that nucleation will occur in any supersaturated medium. In the presence of dust, of course, nucleation is heterogeneous and comparatively easy. Some form of particulate contamination appears to be inevitable in most actual nucleation. The deliberate, controlled addition of impurities is very important in controlling the properties of commercial crystals. In semiconductor crystals the type and degree of conductivity are controlled by the nature and concentration of impurities present.

Crystal embryos

The basis of most semiconductor devices, such as diodes and transistors, is a junction at which materials of different conductivity type meet. Such junctions are produced: (1) by growing material of one conductivity type and diffusing appropriate impurities into it or (2) by altering the concentration of impurities during growth. Single crystals of high perfection and carefully controlled impurities are the materials over which man has achieved the greatest control of structure and properties at the atomic level.

PREPARATION OF CRYSTALS

Monocomponent methods. *Liquid–solid.* Growing crystals in systems consisting of only one chemical component present in both the solid and liquid phases can be considered to be conservative or nonconservative depending upon whether the amount of material in the system is or is not altered during the process. Figure 5 illustrates the important conservative and Figure 6 the important nonconservative crystal processes. In a typical example of one of the conservative (the Bridgman-Stockbarger) methods (Figure 5, top), the container is a cylindrical crucible with a conical bottom. The material to be grown is melted in the crucible and lowered through a furnace with a temperature gradient such that the first freezing takes place in the conical tip of the crucible. In the ideal case, one crystallite comes to dominate the liquid–solid interface and this interface moves through the melt as the crucible is lowered.

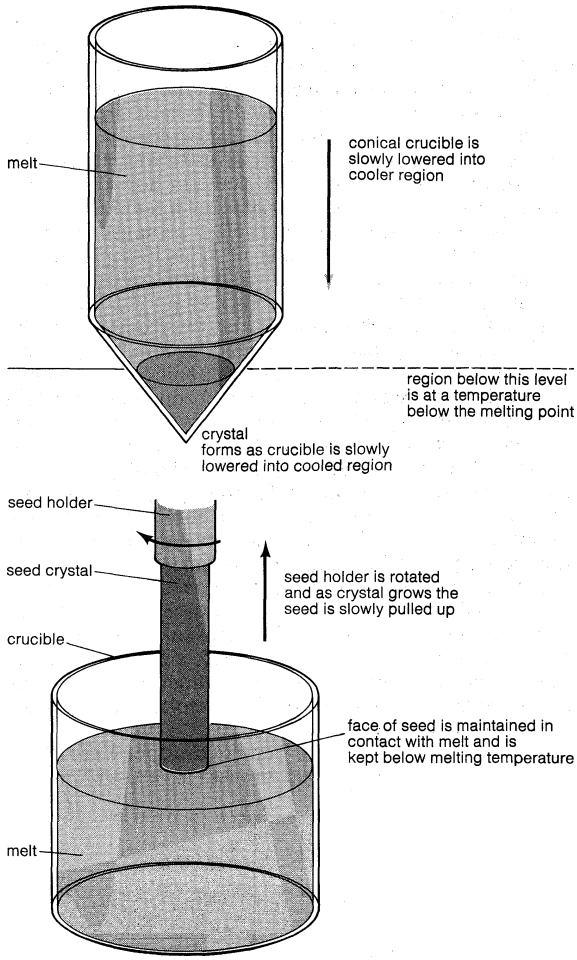


Figure 5: *Conservative growth processes.*
(Top) Bridgman-Stockbarger method, (bottom) Czochralski method.
From R. Laudise, *The Growth of Single Crystals* (© 1970); Prentice Hall, Inc.

Another conservative method is the crystal-pulling (Czochralski) method (Figure 5, bottom), in which a seed is introduced into a melt contained in a crucible arranged so that a temperature level corresponding to the melting

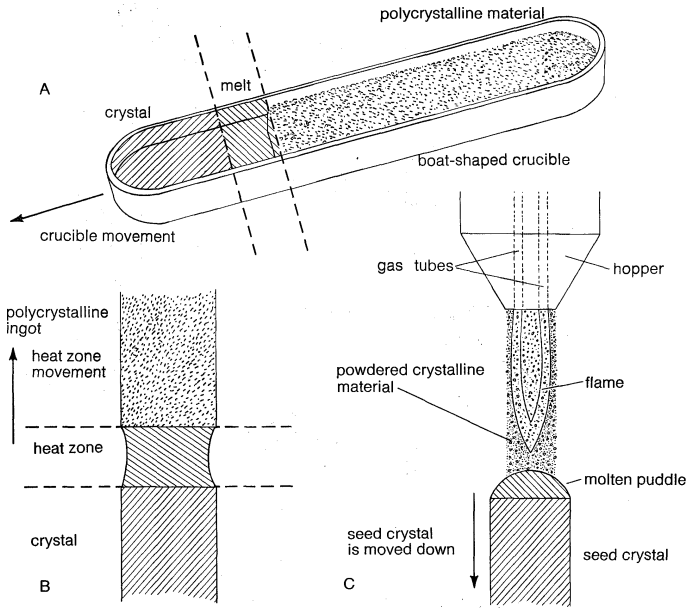


Figure 6: *Nonconservative growth processes.*
(A) Horizontal zone melting using a boat-shaped crucible,
(B) vertical zone melting, (C) Verneuil method or flame fusion.

point lies just above the liquid surface. The seed is then slowly withdrawn from the melt and growth takes place on it. In a third conservative (Kyropoulos) method a cooled seed is immersed in a melt and the temperature of the melt is lowered to cause growth.

Zone melting

Zone melting is an example of an especially powerful nonconservative process, which is used for purification and for the growth of single crystals. In horizontal zone melting (Figure 6A), the material is held in a boat-shaped crucible and the temperature profile is arranged so that a narrow molten zone is produced. This zone is moved down the crucible rather slowly in order to effect a purification or to grow a crystal. If further purification is required, the zone is re-established at the front end of the crucible and the process is repeated. Seeding may be effected by placing a seed at the front end of the crucible and beginning the zone so as to avoid completely melting the seed. Alternatively, the ingot to be melted can be arranged vertically (Figure 6B). This technique, called float zoning when no container is used for holding the melt, allows purification, homogenization, or growth without crucible contamination. The molten zone is held in place by surface tension. The diameter of the melting ingot and the growing crystal need not be the same.

Another nonconservative method is the flame-fusion process (6C). In this growth technique a molten puddle, which is held in place by surface tension, is produced on the surface of the seed by a flame, or plasma, or by radiant heating from a focussed high-intensity light. Solid particles of the material or molten droplets are added to the puddle.

Gas–solid. Gas–solid (or vapour transport) growth in a monocomponent system, a process called sublimation, is of relatively limited applicability. Gas–solid growth in a polycomponent system (growth by reaction) is, however, a very powerful technique. Both reversible and irreversible reactions may be used.

Solid–solid. Growth by means of solid–solid transformation is difficult to control because the density of sites (nuclei) for growth in a solid is so high that growth tends to occur at many places in a transforming solid, with the result that good quality single crystals usually are not obtained. The most generally useful method is strain annealing, a process in which a polycrystalline material is deformed plastically by stretching it and then heating (annealing) the specimen to promote crystal growth as strain is relieved. Transforming an amorphous material to a single crystal by direct conversion from a glassy to a crystalline state (devitrification) is possible, but the high density of nuclei in amorphous materials has so far prevented the use of this method for the growth of crystals.

If a particular material can exist in several different structural modifications (polymorphs), it is often possible to convert a single crystal of one form to the other form by carefully moving the specimen from a region of temperature stability for the one polymorph to that of temperature stability for the other. The geometry and technique is analogous to that used, for instance, in lowering a liquid–solid sample through a temperature gradient (Figure 5, top). Such polymorphic transitions result in twins in the newly formed polymorph, or even in polycrystallinity, when the structures of the two polymorphs are too dissimilar, twins being composite crystals in which the individual parts are related to one another in a definite crystallographic manner. Sometimes appropriately

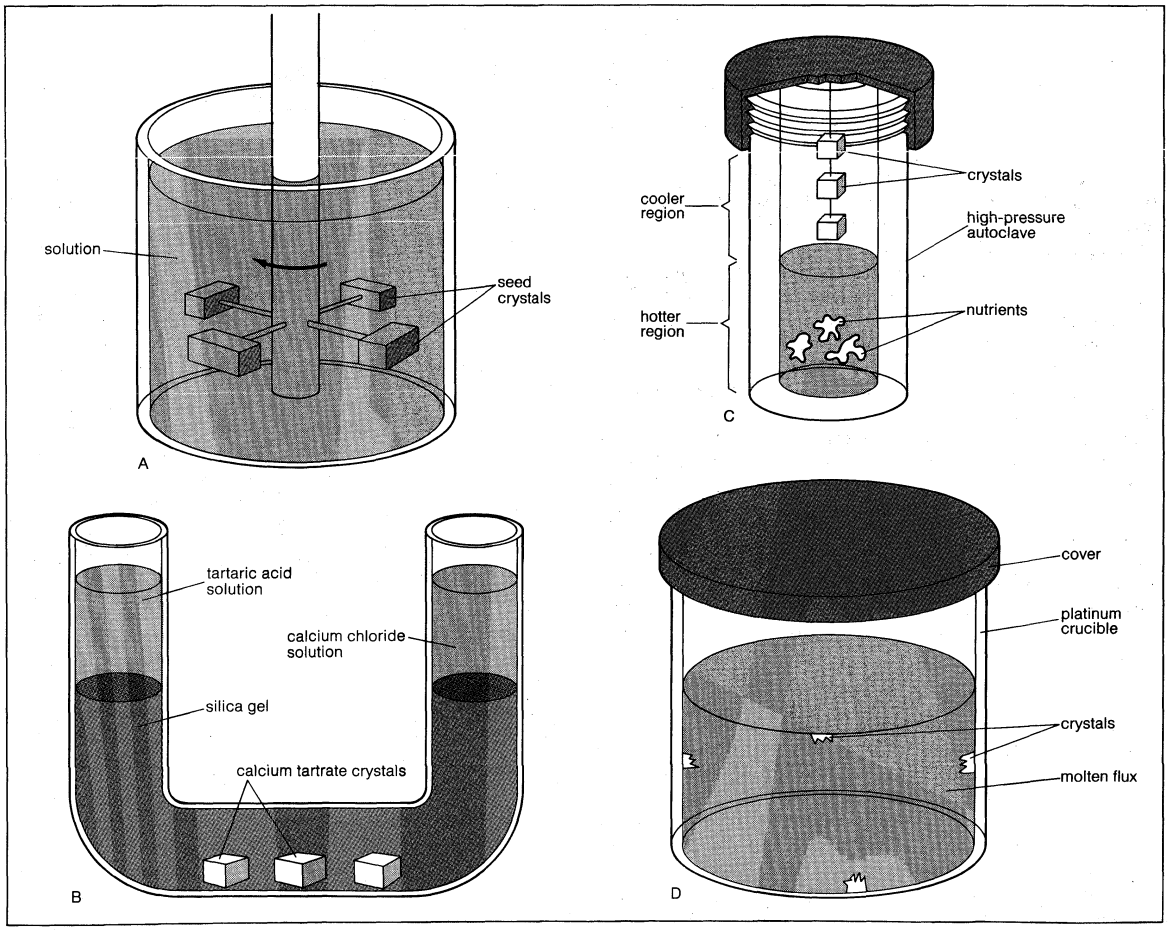


Figure 7: *Solution growth methods.*
(A) Glass-jar crystallizer; growth is produced by gradually lowering the temperature of the aqueous solution. (B) Gel growth; growth is by the reaction of the two solutions that diffuse toward one another through the gel medium. (C) Hydrothermal growth; the high-temperature–high-pressure aqueous solution moves by convection to the cooler region where it is supersaturated and growth takes place. (D) Flux growth; growth is produced by gradually lowering the temperature of the molten salt solvent.

applied stresses can be used to "de-twin" samples or convert them to single crystals.

Polycomponent methods. *Liquid–solid.* Polycomponent methods are particularly important in growth from liquids, in which the addition of a second component can lower the melting point of refractory material to a point where advantages in ease to growth are obtained. Such growth is ordinarily called growth from solution, and the driving force that makes it possible is supersaturation.

A useful classification of the growth methods can be made on the basis of the procedure used for bringing about supersaturation. It includes both isothermal and non-isothermal methods. The isothermal methods are: (1) solvent evaporation or change in solvent concentration, (2) temperature differential, and (3) chemical or electrochemical reaction. A non-isothermal method is slow cooling. Isothermal methods have an advantage over non-isothermal methods in that the concentration of impurities (or dopants) in the grown crystal is uniform. Non-isothermal methods, however, usually require less elaborate equipment and generally do not utilize such long diffusion paths, so that growth rates can be higher. Temperature-differential methods are capable of producing the largest crystals because, in this case, crystal size is not limited by the solubility of the material being grown.

Isothermal methods

Another convenient method of classifying solution-growth methods is based on the nature of the solvent used. This may be (1) water at ambient conditions, (2) water at high temperature and pressure (hydrothermal conditions), (3) molten inorganic salts at high temperature (flux) conditions, or (4) other nonaqueous solvents.

Figure 7 shows the various solution growth methods. The most generally used aqueous-solution growth apparatus is the glass-jar crystallizer (Figure 7A). In this apparatus a saturated solution is exposed to suitably mounted seeds and the solution is cooled a few degrees every day to produce growth. The seeds are rotated at the rate of a few revolutions per minute, first in a clockwise direction and then counterclockwise, to aid in mixing and to hasten diffusion.

If the solubility of a material is so low (less than 5–10 percent) that it makes growth by other methods impractical, growth by chemical reaction can be used; for instance, although the salt calcium tartrate is virtually insoluble in water, tartaric acid and calcium chloride are appreciably soluble, and the reaction between them can be used to form crystals of calcium tartrate. Direct mixing of solutions of tartaric acid and calcium chloride, however, inevitably results in such a large local supersaturation that many centres for growth form, and the product will consist of many small crystallites. In a similar reaction, however, solutions of calcium chloride and tartaric acid are allowed to diffuse through a substance called silica gel from the opposite ends of a U-tube (Figure 7B); the region of high supersaturation is localized near the centre of the tube, and a few comparatively large crystals are formed. This method is called gel growth and is quite useful for the growth of insoluble materials by means of a chemical reaction.

Single crystal growth by an electrochemical reaction, although little used, in principle also could be an important means of preparing certain crystals. Careful control of voltage and current, for example, could be used to grow single crystals of metals on an appropriate single-crystal seed electrode.

Hydrothermal method

Under hydrothermal conditions the increased solvent power of water at high temperatures is used to dissolve refractory materials. The water is heated in a high-pressure autoclave where it expands and fills the vessel (Figure 7C). Depending on the temperature and on the percent of the free volume of the autoclave initially filled with water, the pressure can range from a few hundred to many thousand atmospheres. The solvent power of water can be increased further by the use of "complexers"—substances that react with the material to be crystallized in such a way as to form soluble complexes.

Flux method

In flux growth a molten salt is used as the solvent (Figure 7D). Generally, flux growth is carried out in the absence of deliberately added seeds. Initial nucleation occurs spontaneously when an appropriate saturation is reached in the cooling cycle, and these nuclei provide the centres for further growth.

Gas–solid and solid–solid. Gas-phase polycomponent growth has been referred to above (see *Monocomponent growth*). Solid–solid polycomponent growth (by solid-state precipitation) is difficult to control because of the high density of nuclei in solids, and it is slow because solid-state diffusion is slow. As a result, the process has not been used.

Vapour–liquid–solid. Vapour–liquid–solid growth makes use of a gas-phase reaction for the transport of material to a liquid solvent. The material dissolves in the liquid solvent, supersaturates it, and deposition takes place by a transformation from liquid to the crystal.

The choice of growth method. An important problem in crystal growth is the choice of the most suitable growth method. Generally, monocomponent methods are to be preferred because of their comparative rapidity and simplicity. In them, growth rate is limited essentially only by the rapidity with which the heat produced by crystallization can be dissipated. The purity of crystals grown in monocomponent systems is high, because the growth system does not contain additional components to contaminate the grown crystal. If the temperature required for melting or sublimation is too high, however, additional components must be added to produce growth from solution or by vapour-phase reaction. The reasons for growing crystals at lower temperatures include: (1) avoidance of undesired high-temperature polymorphs, (2) elimination of decomposition or irregular melting behaviour, (3) elimination of high vapour pressure, which would be present at the melting point, and (4) convenience of operation. When an additional component is added to make possible lower operating temperature, however, the rate of growth usually must be reduced, in order to accommodate diffusion of the growth component to, and the additional component from, the growing interface. If the additional component has appreciable solubility in the growing crystal, its inclusion may become a problem.

Limiting temperatures

BIBLIOGRAPHY. R.A. LAUDISE, *The Growth of Single Crystals* (1970), a comprehensive discussion of the techniques of crystal growth and theory relevant to practical growth; J.J. GILMAN (ed.), *The Art and Science of Growing Crystals* (1963), on the major techniques of growth; B. CHALMERS, *Principles of Solidification* (1964), on the theory of melt growth; J.P. HIRTH and G.M. POUND, *Condensation and Evaporation* (1963), on the theory and technique of vapour growth; K. LARK-HOROWITZ and V.A. JOHNSON (eds.), *Solid State Physics* (1959), with particular emphasis on electronic materials; A. HOLDEN and P. SINGER, *Crystals and Crystal Growing* (1960), an elementary work on the growth of water soluble crystals and symmetry; A. VAN HOOK, *Crystallization* (1961), on industrial scale crystallization; R.H. DOREMUS, B.W. ROBERTS, and D. TURNBULL (eds.), *Growth and Perfection of Crystals* (1958), classic theoretical papers; H.S. PEISER (ed.), *Crystal Growth* (1967), a collection of papers on all aspects of the subject; W.G. PFANN, *Zone Melting*, 2nd ed. (1966), including information on impurity distribution.

(R.A.L.)

Crystallography

The word crystal stems from the Greek word *krystallos*, meaning "clear ice." It was first applied to the clear crystals of quartz found in the Swiss Alps, because these were thought to be formed from water under conditions of intense cold. Such crystals actually reflect a highly symmetric and periodic internal arrangement of atoms; this is characteristic of all crystalline materials—metals, alloys, minerals, ceramics, and some organic materials. Crystallography is the science of crystals and of the crystalline state. It is concerned with the nature of the regular atomic or molecular arrangements within a crystal, the bonding of the atoms or molecules, and the physical and chemical properties that result from certain arrangements. Many rows and planes of atoms occur within a crystal, for example, and some rows (and planes) exhibit equal spacing of atoms in different directions, whereas others do not. Because the atomic arrangement is generally different in different directions, properties

such as elasticity or light transmission commonly depend on direction.

Crystallography is also concerned with atomic arrangements in amorphous materials, liquids, gases, and the structure of living matter and the stresses in metals. All forms of atomic arrangements are in the crystallographic domain; and patterns, arrays, and symmetry in general have fascinated thinkers down through the centuries. Interest initially was most keenly concerned with crystals, and Johannes Kepler wrote a monograph for one of his patrons on the symmetry of snowflakes in 1611. The versatile 17th-century British natural scientist Robert Hooke wrote:

> There was not any regular Figure, which I have hitherto met withal, of metals, minerals, precious stones, salts and earths, that I could not with the composition of bullets or globules, and one or two other bodies, imitate even almost by shaking them together.

This article treats the classification, symmetry, and structure of crystals, including the determination of crystal structure by X-ray diffraction methods. It also includes a section on imperfections and crystal defects. For further information on the important properties of crystalline materials, see those articles treating properties—*e.g.*, MAGNETISM or ELASTICITY and the theoretical articles CHEMICAL BONDING and SOLID STATE OF MATTER. See also METALS, THEORY OF and MINERALS for insight into some applications of crystallography and the article CRYSTALLIZATION AND CRYSTAL GROWTH for coverage of the formation of crystals from solution.

THE PATTERNS OF ATOMS IN CRYSTALS

Interfacial angles

One of the first activities undertaken by those interested in natural crystals was an attempt to classify their shapes. This was not easy because even crystals of the same compound may fail to resemble one another; the difference in shape results from different degrees of development of the several faces. In 1671 the Danish physician Nicolaus Steno took the first grand step in this field. He cut sections from a quartz crystal and discovered that, even when the shapes were vastly different, there was a constancy of interfacial angles (the angles formed by normals or perpendiculars from crystal faces). For this reason it is common practice to use "normals" drawn perpendicular to crystal faces to represent the relationship between faces. This can be accomplished by stereographic projection as shown in Figure 1. The intersection of normals with an imaginary sphere surrounding the crystal (Figure 1A) is represented on the two-dimensional map (Figure 1B) of the shaded plane. The high degree of symmetry present is evident.

Crystal planes and their notation. In the late 18th century the idea that a large crystal could be built of a basic unit or building block took hold. This block is now known to be the unit cell—the smallest group of atoms within a crystal that will yield the crystal structure and properties by simple repetition of pattern. This arrangement is indicated in two dimensions in Figure 2, in which two planes of atoms perpendicular to the page are shown by the dark lines. Such planes will appear perfectly flat to the eye (rather than composed of a series of steps) because the building blocks are very small, about 10^{-9} to 10^{-10} metre on an edge. The angle formed by the plane on the right is 45° with the horizontal plane. Note that it is not possible to have just any angle; a plane must contain atoms in the structure, and the intercepts on the edges of the crystal are ratios of whole numbers of blocks. The intercepts of the plane on the left and the horizontal plane have a ratio 1:2 and the angle involved is 26°34′.

A simple notation for naming planes has developed based on the fact that the intercepts of planes on the edges of the crystal's unit blocks are ratios of whole numbers. A basic unit cell or block is assumed, and three noncoplanar edges (a_1, a_2, a_3) are defined as unit lengths in Figure 3. These need not be equal in all directions. The intercepts of planes on these edges are written as

$$\frac{a_1}{h}, \frac{a_2}{k}, \frac{a_3}{l}.$$

Thus, if a plane intercepts the axes at their tips,

$$\frac{a_1}{h} = \frac{a_1}{1}, \frac{a_2}{k} = \frac{a_2}{1}, \frac{a_3}{l} = \frac{a_3}{1},$$

and (111) is the designation. If the intercepts are ½,½,½, then h,k,l are 2,2,2. Common factors are cleared and the designation is still (111)—because all such planes are parallel and have identical atomic arrangements in the structure (Figure 3E). If a plane intercepts one axis at unit distance and is parallel to the other axes, as in Fig-

From E.E. Wahlstrom, *Optical Crystallography* (1951); John Wiley and Sons, Inc.

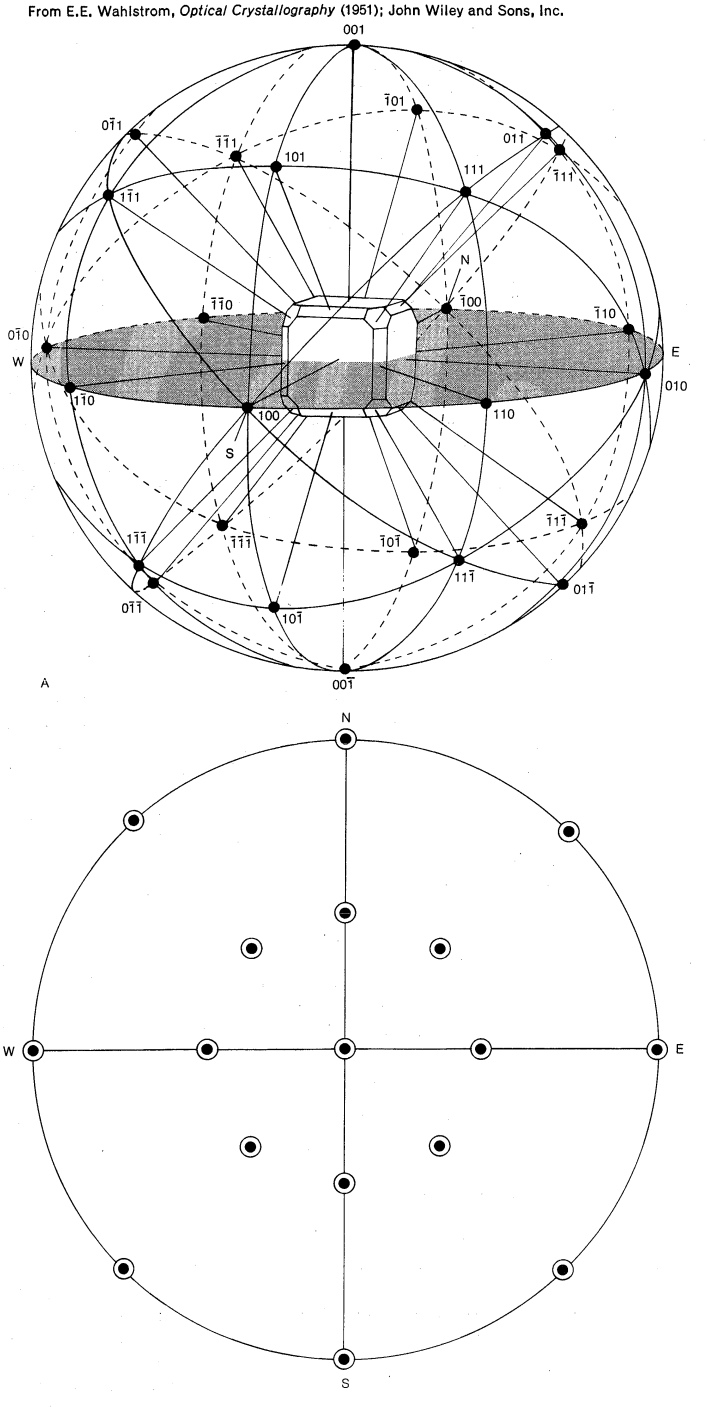


Figure 1: A two-dimensional representation (stereographic projection) of a crystal and its symmetry that can be obtained by surrounding the crystal with a sphere, erecting perpendiculars to the planes until they intersect the sphere as in A, and then projecting these intersections through the shaded circle of the sphere. The appearance is then as in B. This process eliminates the shapes of the planes and their relative sizes, which depend on growth conditions and can confuse the similarity of crystals.

ure 3D, then two integers will be zero—for example, (001) in this case.

From F.C. Phillips, *An Introduction to Crystallography*, 2nd ed. (1956); Longman Group Ltd.

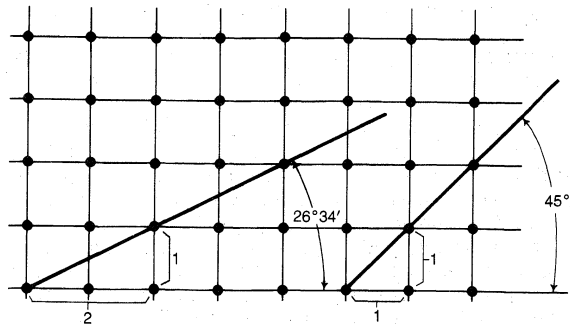


Figure 2: Two planes of atoms in a crystal represented by their edges (dark lines). Note that these make intercepts that are whole numbers along the edges of the cells or building blocks.

Symmetry considerations. Rotation of 90° around the centre of the map of normals to planes in Figure 1B leads to the same figure as before the rotation, as does rotation of 180°, 270°, or 360°. This rotation is generally described by the number n, by which 360° (a complete turn) is divided. In this case $360°/n = 90°$, so that n is 4, and then this apparent rotational symmetry axis is thus called "4-fold." To return to the starting point after n rotations, it is then necessary that n be an integer. In examining many natural crystals, n has been found to have values of 1,2,3,4, or 6; sometimes several of these values are found to be present in one crystal. But the values are never 5 or 7, although certain flowers and starfish are excellent examples of these symmetries. In order to understand the reason such values are excluded in crystals, it is necessary to ask more precisely what results from periodic arrangements of atoms. One-dimensional arrays are easiest to consider, and the different ways a one-dimensional molecule can be periodically arranged in one dimension provides a suitable example.

Rotational axes of symmetry

A periodic set of crosses, as shown below, is first set down as a frame of reference; this imaginary collection is called a lattice.

\+ + + + +

The actual spacing of the crosses is not important because it is the kinds of arrangements that need to be considered, not their sizes. The entire array can be constructed after one point and the spacing between the points are designated. This basic framework—a point and the spacing to the next point—is called a unit cell, and in one dimension there is obviously just one such cell. If a molecule (represented by a dot–dash) is placed within this framework, two arrangements may be considered, as follows:

(1) +·— +·— +·— +·—

(2) +·— —·+·— —·+·— —·+·—

In the second lattice the "molecules" are arranged at each lattice point, so that one molecule leads to another after rotation of 180° around an axis normal to the paper (at the crosses or halfway between them). This is called 2-fold rotation because repetition occurs twice in one complete (360°) rotation. There are two molecules per lattice point. A mirror can also be imagined at these locations, with the molecules related by reflection through the mirror. Two reflections yield the original molecule. This indicates that the interaction of symmetry leads to some redundancy—one symmetry element produces another; rotational symmetry at lattice points produces rotational symmetry between lattice points, as well as the mirrors, but only one symmetry element or type is required to describe the figure. Only the unit cell and its contents are needed to produce the entire structure by translation, although the unit cell in lattice (2) contains two molecules.

Another possible pattern is the following:

(3) +·— —··— +·— —··— +·— —··— +
c a b d

But this is really not different from pattern (1); the unit in each cell is just longer—three molecules long. It might appear that there is rotational symmetry about an axis between a and b, but c would not be carried into d by the rotation. This is not a symmetry element of the whole structure. Hence the requirement of periodicity leads to only two patterns in one dimension. In these periodic structures, examination of some point within the structure, a lattice point, or any point on a molecule, reveals the same surroundings at every such point in every unit cell.

Symmetry in two and three dimensions

The entire array is called a space group, and the symmetry about a point is called a point group. There are one unit cell, two point groups, and two space groups in one dimension. Compact notations have been developed for these in one, two, or three dimensions and will be considered shortly. Symmetry in two and three dimensions will be treated first, however. Symmetry at a point in two dimensions involves the question: How many ways can a molecule be arranged periodically about a lattice point? Ways without any symmetry are also included because a 1-fold rotation axis produces this effect—*i.e.*, the entire crystal will exhibit no symmetry at all. Arrangements that do include some symmetry are shown in Figure 4, in which a stick hand has been chosen as a general figure for two dimensions because it has no symmetry of its own. Hands A and C are related by 180°, or 2-fold rotation, and A and B are related by reflection through the

Reprinted with permission of the Macmillan Company from J.B. Cohen *Diffraction Methods in Materials Science*; Copyright © by Macmillan Company (1966)

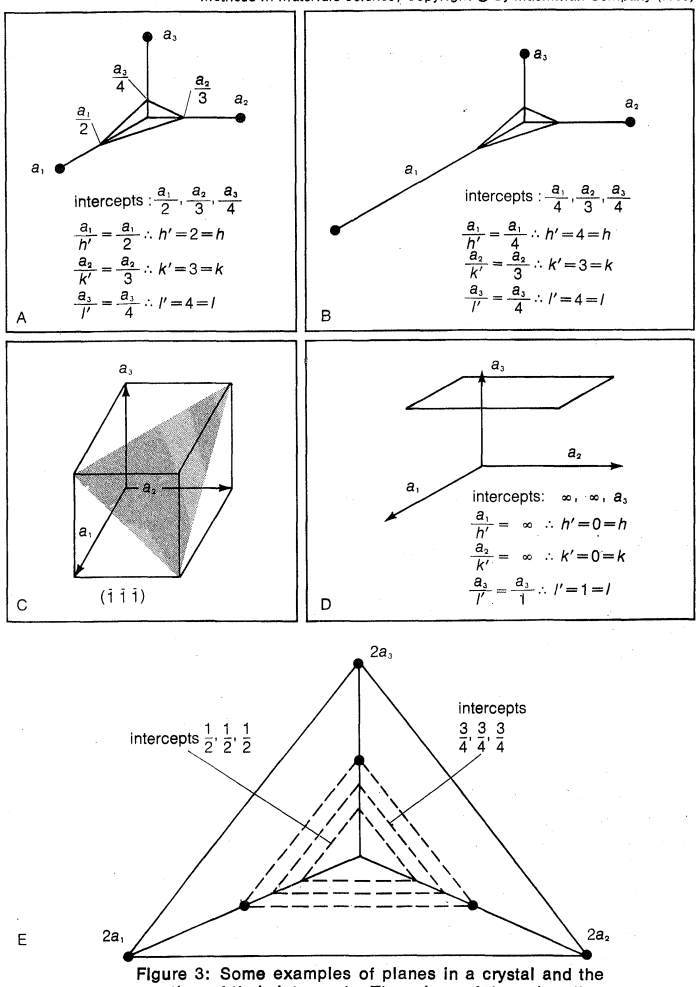


Figure 3: Some examples of planes in a crystal and the notation of their intercepts. The edges of the unit cell are defined by a_1, a_2, and a_3 (see text).

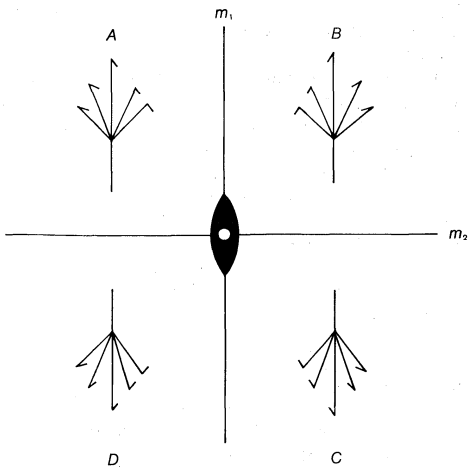


Figure 4: A set of stick hands arranged symmetrically around a point. Both mirror and rotational symmetries are present. The first relates *A* and *B* through m_1. The second relates *A* and *C*; the axis is normal to the page and is represented by the oval (see text).

Reprinted with permission of the Macmillan Company from J.B. Cohen, *Diffraction Methods in Materials Science;* Copyright © by Macmillan Co. (1966)

imaginary mirror plane m_1 perpendicular to the page. But *A* and *C* could also be related by reflections through m_1 and then through m_2. A rotation of 180° and reflection still retain some equivalence, as in one dimension, but there is no way to get *B* from *A* except by reflection.

Another symmetry element has arisen at the intersection of the two mirror planes; figures *A* and *C* are related by inversion through this point of intersection. That is, *C* can be obtained from *A* by drawing lines from every point on *A*, through the intersection, over an equal distance on the other side of the intersection. In two dimensions then, *C* can be produced by starting with *A* and rotating 180°, or by inversion. The presence of one of these symmetry elements leads to the other. Inserting m_1 and 180° rotation leads to all four hands and, in turn, to two other symmetry elements (m_2 and an inversion centre) that were not needed in the construction. The mirror plane m_2 leads to nothing new in terms of the pattern produced by m_1 and 180° rotational symmetry, for example.

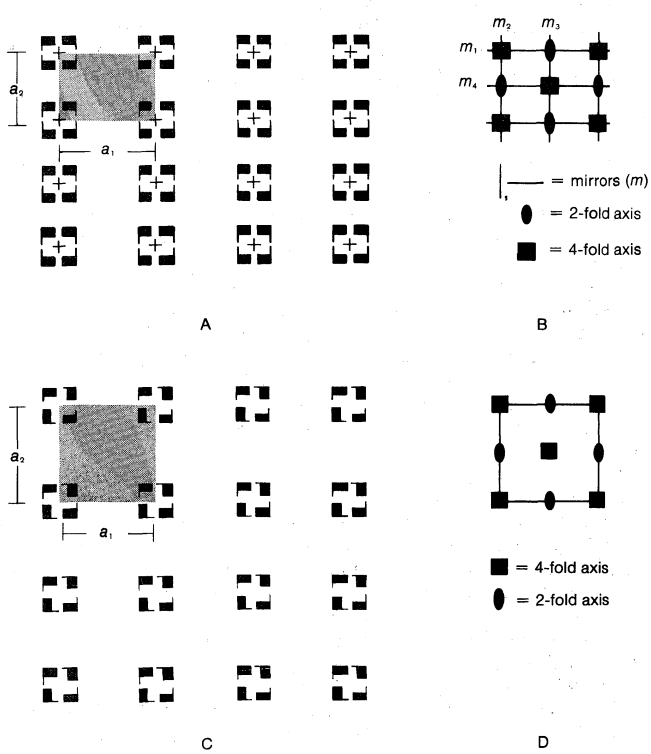


Figure 5: Two-dimensional space groups in which a unit cell is shaded (A and C). At each lattice point there are four molecules (flags). These are related by two perpendicular mirror planes, m_1, m_2 in A but additional symmetry arises, such as m_3, m_4 and 2-fold and 4-fold rotational axes normal to the paper (B).

A periodic array, or space group, is shown in two dimensions in Figure 5A. A flag is employed to represent a two-dimensional molecule; one unit cell is shaded and contains four such molecules. To create the entire periodic array, the lengths of the two sides of the unit cell, a_1 and a_2, and the angle (α) between them are required. One cell can then be constructed and translated in the directions of a_1 and a_2 to fill all space. If one molecule is placed near each lattice point and if each lattice point has two perpendicular mirrors, then the four molecules at each lattice point result, together with the additional symmetry elements shown in Figure 5B. Another arrangement, involving a 90° rotational axis without mirrors, is shown in Figure 5C.

Impossibility of 5-fold or 7-fold symmetry

It is clear that any axis of rotational symmetry in a crystal must be $360°/n$, in which n is an integer, because to have periodicity at a point there must be a return to the original figure after n rotations. The reason that n is restricted to 1, 2, 3, 4, 6 in a crystal, excluding 5 or 7, becomes clear with the aid of Figure 6. Assume that there is a rotational axis, with rotation $\alpha°$ at each lattice point. Not only must all molecules or atoms around any lattice point be related by rotation of $\alpha°$ around this axis, but also this must be true for all points in the entire structure; otherwise, the symmetry would relate only to molecules at one point and not to the entire crystal. The lattice point itself must be carried by rotation to another lattice point; this preserves the internal periodicity and the external symmetry.

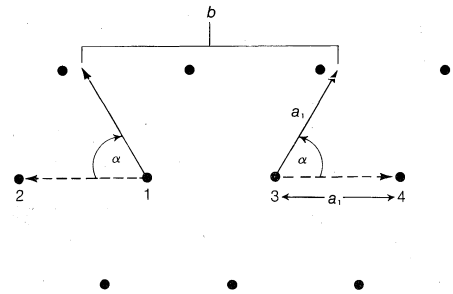


Figure 6: Rotational symmetry in crystals: rotation of α around lattice point 1 must bring lattice point 2 into coincidence with some other point; the same is true for 4, around 3. If not, as shown here, the periodicity, and hence the external symmetry, is destroyed (see text).

In Figure 6, m is an integer and the relation between the distances a_1, b, and the angle of rotation, α, can be expressed by the equation: $b = ma_1 = a_1 + 2a_1 \cos \alpha$, or

$$\cos \alpha = \frac{m-1}{2} = \frac{M}{2}, \qquad (1)$$

in which M must be an integer if m is an integer. The cosine of an angle cannot exceed plus or minus unity; and with M an integer the possible values for the cosine are 0, plus or minus ½, plus or minus 1. For cos α to be zero, α must be 90° ($n = 4$). Furthermore, for this 4-fold rotation axis, any side a_2 must equal a_1, because after a 90° rotation a_2 must coincide with a_1. The cell must be square —it cannot be a rectangle. The other allowable values for the cosine correspond to 1-fold, 2-fold, 3-fold, and 6-fold rotations and to two other shapes, or unit cells, a rhombus and a parallelogram. It is possible to have a molecule with 5-fold symmetry—like a 5-pointed star—at each lattice point, but the symmetry of the entire structure will not be 5-fold; similarly, 7-fold symmetry is prohibited.

Combined symmetry operations

Combined operations—other than just rotation, reflection, inversion, and translation—are also periodic in two dimensions. A glide, for example, is a combined opera-

tion in which the molecule is related to another by translation followed by reflection; the translation is half the spacing between lattice points, and thus two glides produce the same appearance at each lattice point. Proceeding from one lattice point to another, a molecule is seen in left-handed orientation, one halfway between lattice points is in a right-handed orientation, and at the next lattice point the original left-handed orientation appears. Other possible combined operations in two dimensions, such as rotation followed by reflection or rotation followed by translation, give rise to figures which are equivalent to rotation alone or to combinations of rotation plus reflection at each lattice point. It is possible to show that there are only 17 two-dimensional patterns or space groups and only ten point groups that are consistent with periodicity.

In three dimensions, two new kinds of arrangements of symmetry become possible. It is possible to have three rotation axes intersecting at a point in certain combinations, and axes can be combinations of rotation and translation called screws, as shown in Figure 7. As with glide,

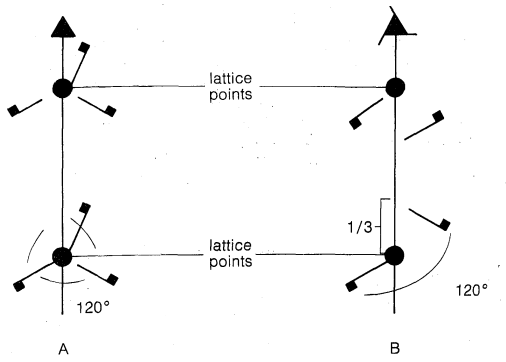


Figure 7: *Symmetry associated with screw axes.* (A) Two lattice points in a structure with three molecules rotated by 120° around each; the whole group at the bottom is translated to every other lattice point (like the one at the top) by the periodicity of the crystal. (B) Periodicity can also be obtained with screwlike motions (rotation plus translation along the rotation axis) where there is only one molecule at each lattice point; each molecule is related to the other molecules by a 120° rotation and a translation of ⅓ the distance between lattice points. Triangular symbols at the top represent rotational and screw axes in A and B, respectively.

this operation in a crystal must bring a molecule to a definite location and orientation near a lattice point in every cell. This serves to limit the possible types or pitches of the screws.

Displacements of planes of atoms or molecules that result from screw axes in natural crystals are approximately equal to the magnitude of atomic spacing—*i.e.*, 10^{-10} metre or one angstrom (Å). The effect, therefore, is indistinguishable from pure rotation to the naked eye, as shown in Figure 8.

Reprinted with permission of the Macmillan Company from J.B. Cohen, *Diffraction Methods in Materials Science;* Copyright © by Macmillan Company (1966)

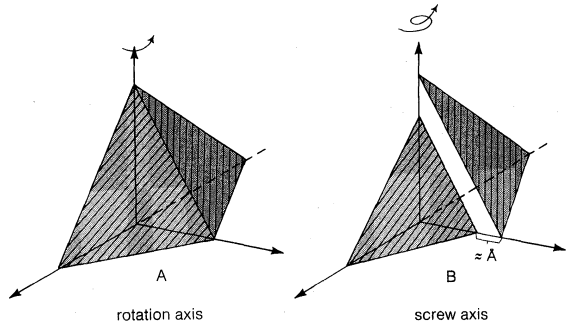
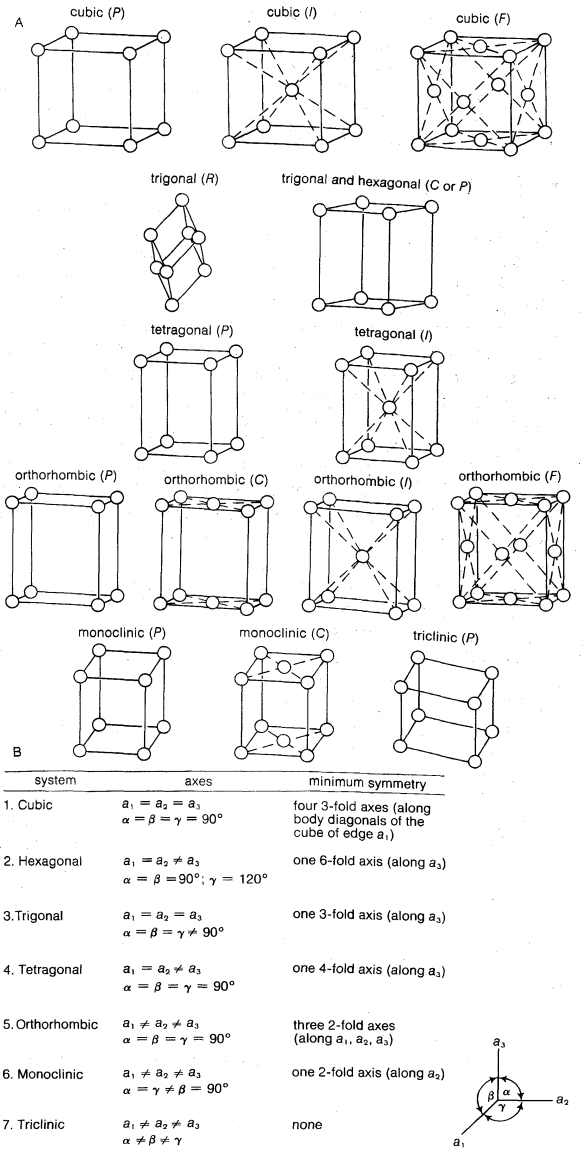


Figure. 8: *Distinction between rotational and screw axes.* (A) Rotation axis relates the shaded faces. (B) A screw axis relates the shaded faces, but there is also a translation. This is so small (about 10^{-10} metre, or one angstrom), however, that the differences in these two cases cannot be seen on a real crystal (see text).



system	axes	minimum symmetry
1. Cubic	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma = 90°$	four 3-fold axes (along body diagonals of the cube of edge a_1)
2. Hexagonal	$a_1 = a_2 \neq a_3$ $\alpha = \beta = 90°; \gamma = 120°$	one 6-fold axis (along a_3)
3. Trigonal	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma \neq 90°$	one 3-fold axis (along a_3)
4. Tetragonal	$a_1 = a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$	one 4-fold axis (along a_3)
5. Orthorhombic	$a_1 \neq a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$	three 2-fold axes (along a_1, a_2, a_3)
6. Monoclinic	$a_1 \neq a_2 \neq a_3$ $\alpha = \gamma \neq \beta = 90°$	one 2-fold axis (along a_2)
7. Triclinic	$a_1 \neq a_2 \neq a_3$ $\alpha \neq \beta \neq \gamma$	none

Figure 9: *Classification of crystal systems.* (A) The 14 unique three-dimensional unit cells. Balls represent lattice points around which, or on which, atoms or molecules can be arranged. (B) Symmetry characteristics of each crystal system.

Reprinted with permission of the Macmillan Company from J.B. Cohen, *Diffraction Methods in Materials Science;* Copyright © by Macmillan Company (1966)

The seven crystal systems

Seven crystal systems and 14 unique unit cells arise, as shown in Figure 9. There are 32 point groups and 230 possible structures based on atomic position. Other arrangements can arise if such properties as the magnetic moment are considered, but the number of unique periodic arrangements is still finite.

Optical determinations. Optical techniques can be quite useful in the study of crystals of transparent materials, because only in cubic crystals does light travel at the same velocity in all directions. In a tetragonal structure, for example, light is broken into two rays for any direction of propagation other than that of the 4-fold axis. These rays travel at different velocities and vibrate in different planes. If light is transmitted through a sheet of Polaroid, which only allows passage of light vibrating in one plane (direction), then rotation of the crystal will lead to elimination of one ray when its plane of vibration is perpendicular to the vibrations passing through the Polaroid sheet. If another piece of Polaroid is placed over the crystal and turned 90°, the light passing through the crystal cannot pass this second sheet. Every turn of the crystal by 90° produces this extinction, but between these positions the crystal is visible. If a thin slice of the crystal is examined between crossed polarizing lenses along the

4-fold axis, this splitting does not occur, and the crystal is always dark. This procedure can help distinguish the presence of a unique axis of symmetry. If a screw axis is present the position of the two Polaroids for extinction varies with the thickness of the crystal. Glide cannot be detected in this way, however.

SCATTERING OF RADIATION FROM CRYSTALS

In 1912 Max von Laue, an instructor at the University of Munich who was well versed in optics, particularly in the diffraction of light from ruled gratings, became intrigued with the question of what would happen when radiation with a wavelength equal to the spacing of the atoms in a solid was scattered from the atoms. Light scattering could not reveal atomic structure because atomic spacings are 10^{-10} metre, whereas light has a wavelength 10^{-7} metre. Laue thought, however, that scattering of radiation from a row of atoms in a crystal (if the wavelength of the radiation was the same size as the atomic spacing) would produce interference patterns. In a three-dimensional crystal, there are rows in three directions and, hence, such patterns in three directions. Only at their points of intersection will all the waves be in phase, and a set of dark spots should occur on a film placed in front of a small crystal when an X-ray beam impinges on it. It was not known whether X-rays had the proper wavelength, but Laue persisted in his experiment, and he succeeded on his second try. Laue and his associates demonstrated the existence of a new tool for examining the internal arrangements of atoms in crystals and proved that X-rays behave as electromagnetic waves (like light) but with a shorter wavelength.

Bragg's law and crystal structure. Laue's group had difficulties in understanding which planes of atoms and wavelengths produced the spots on the film. In England, Sir Lawrence Bragg was in his mid-twenties and finishing his studies in physics. His father was involved in research on the ionization of gases by X-rays, and it was quite natural for them to discuss Laue's exciting new result. The young Sir Lawrence was able to formulate a simpler and more physical understanding of Laue's discovery. The scattered waves from successive planes of atoms reinforce one another at certain angles if the path difference between rays from the two planes is a multiple (n) of the wavelength. As shown in Figure 10, this construction

Relation of wavelength and angle of incidence

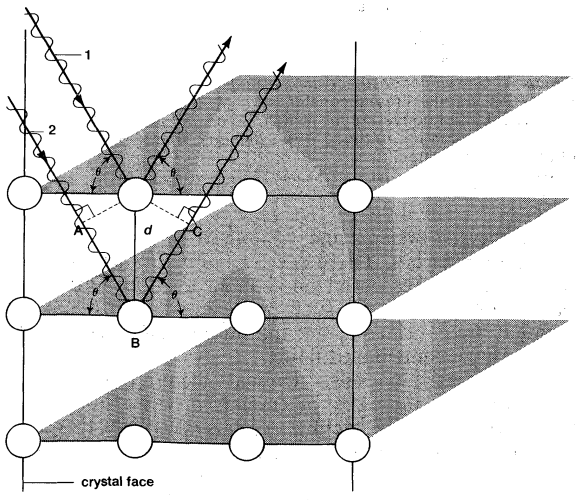


Figure 10: Incident rays (1 and 2) at angle θ on planes of atoms in a crystal. Rays will reinforce if their difference in path (AB + BC) is an integer times the wavelength (see text).

leads to a simple equation for diffractions or constructive interference. For diffraction the path difference, AB + BC, is equal to an integer times the wavelength, or $n\lambda$. But by simple trigonometry, AB = BC = $d \sin \theta$, and thus Bragg's law is derived: $2d \sin \theta = n\lambda$. As n increases from 1 to 2 to 3, etc., so does the scattering angle θ. Each plane, for each n, selects in effect the appropriate wavelength from the X-ray beam (which must consist of many wavelengths). This selection depends on the angle of the plane of atoms to the X-ray beam. With this formulation, Bragg was able to explain successfully all the spots from Laue's patterns from different planes of atoms. He then tried bouncing an X-ray beam from a crystal in reflection, as from a mirror, catching the scattering on a film. This was more like his formulation, and sharp increases in intensity were recorded at certain angles.

Bragg's law permits the atomic structure in crystals to be unravelled. If another set of planes is inserted at half the d spacing of the planes in Figure 10 and if these planes contain the same kind of atoms, then the prevailing conditions are as shown in Figure 11. If the angle em-

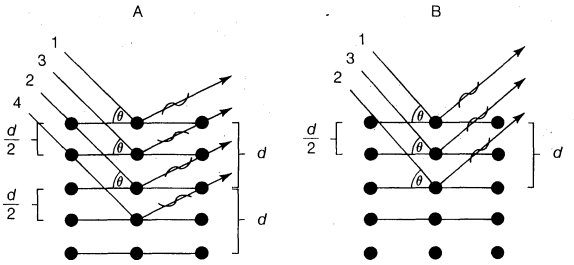


Figure 11: *Reinforcement and cancellation of incident rays.* (A) X-ray beams 1 and 2 are one wavelength difference in path; beam 3 to a plane at half the spacing is half a wavelength different in path with respect to 1, is thus out of phase, and cancels it. Similarly, beam 4 cancels beam 2, etc. (B) The angle of X-ray beam to crystal has been changed so that all rays are in phase and reinforce one another.

ployed is that for first-order diffraction without these additional planes—*i.e.*, $n = 1$ in Bragg's law—then the path difference between rays 1 and 2 is one wavelength (Figure 11A). But the path difference between rays 1 and 3 (to the inserted plane) is half a wavelength. The scattered waves 1 and 3 are out of phase, and their sum is zero; thus, the reflection is extinct. When $n = 2$ (Figure 11B) the scattering from the inserted plane differs in path length from plane 1 by one wavelength, and all planes are in phase. The second reflection is present and is even stronger than if there were no inserted plane because there are more planes (atoms) contributing. Reflections are labelled by multiplying the Miller (hkl) indices by n. Thus, a second order ($n = 2$) reflection from (100) planes is labelled (200). Bragg's law is rewritten:

$$\lambda = \frac{2d}{n} \sin \theta, \qquad (2)$$

and these (200) planes have half the spacing of the (100) planes.

Determination of a cubic structure

These simple ideas provided the solution of many inorganic structures almost immediately. Consider a cubic crystal with one of the three possible arrangements of atoms shown in Figure 9A and assume that there is only a single wavelength for the incident radiation. The spacing of planes (d) in terms of the length of the edge of the unit cell (a) and the indices of the reflection planes (hkl) can be expressed:

$$d = \frac{a}{\sqrt{h^2 + k^2 + l^2}}. \qquad (3)$$

Thus, Bragg's law can be restated by substituting this value of d as:

$$\lambda = \frac{2a}{\sqrt{h^2 + k^2 + l^2}} \sin \theta. \qquad (4)$$

Now λ is a known, fixed quantity, and the larger the values of h, k, l, the larger the angle θ. As shown in Figure 12, a face-centred structure contains (100) planes halfway between the faces of the cube. Thus, there will not be a (100) reflection as there would be for a cube with atoms only at corners; instead there will be a (200) reflection of larger intensity. The (100) reflection would also be missing for a cube with atoms at the corners and in the centre of the cube, however. How can

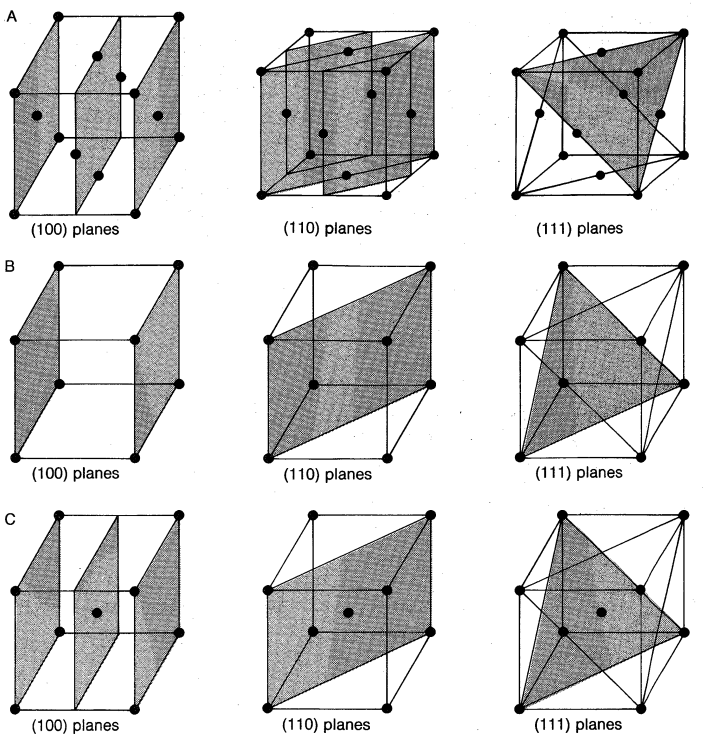


Figure 12: Various planes in (A) face-centred cube, (B) simple cube, and (C) body-centred cube. Note that there are planes of atoms halfway between cube faces in A and C but not in B. Similarly there are (110) planes in A halfway between those in B and C, and in C there is a (111) plane containing the atom at the cube's centre halfway between those in A and B.

the proper choice of structure be made? The (110) planes for the face-centred cubic crystal, simple cube, and body-centred cube also are shown in Figure 12. Because planes halfway between these occur only in the first case, a (110) reflection will not occur from the face-centred structure but, rather, from a simple cube or a body-centred cube. If the first reflection from the cube faces of a natural crystal is obtained, then a can be calculated from equation (4) above and the angle of the observed reflection, θ. Then θ can be calculated for the (110) planes from the same equation. The crystal can then be rotated 45° to get a reflection from these (110) planes, because the (110) planes are 45° from the (100) planes. If the crystal actually is face-centred cubic, then the first reflection from the cube face was really a (200), not a (100). If this were not known, then the calculated value of a from Bragg's law would be half as large as that of the real unit cell. If this incorrect value of a is used to calculate the θ position for the (110) reflection, the value obtained would actually be where a (220) reflection occurred, but this would not be known. If, however, the crystal were body-centred, the lowest angle reflection would be for a (110) reflection, which would be at a lower angle than calculated. Recalculation assuming that the first peak is a (200) reflection would give a correct value for a, from which the correct angle of the (110) reflection could be calculated. If this does not occur, then the crystal must be face-centred or simple cubic. Considering the (111) planes in the same way would distinguish these cases. If the atoms on these "in-between" planes are of a different species, a reflection will not be missing, only weakened. Scattering from different kinds of atoms is different, and the scattered waves do not exactly have the same heights, thus these waves cannot completely cancel the way the equal waves did in Figure 11.

The general method involved in such crystallographic detective work is repeated comparison of experiment and calculation for the angles and intensities of the various reflections. In this manner, it is possible to sort out some of the simpler structures. As the structures examined become more complex—involving in some cases thousands of atoms per unit cell—these simpler procedures provide only limited information. The progress of crystallography from the 1920s consisted of developing in detail systematic methods of sorting out the diffraction pattern. The process divides itself naturally into two parts. First, it is necessary to find all the diffraction spots (and to identify which ones are missing); and, second, it is necessary to record the intensity of these spots and devise procedures for interpreting these.

Missing reflections

Certain general information can be obtained from the indices of planes reflecting and not reflecting. If atoms are arranged with a face-centred cubic unit cell, for example, the only reflections that will appear are those with h, k, l all even or all odd. The (111), (200), and (220) reflections occur, but not the (100) and (110). Furthermore, the positions of reflections can be calculated with a single parameter, the edge of the unit cell. Other unit cells express themselves differently. Certain symmetry elements give rise to specific absences of reflections. A glide plane halves the spacing of planes perpendicular to it, and this requires h to be even for a reflection; for h odd, the molecules on the plane at half the spacing cancel the reflection. Such rules are tabulated for different symmetries and the different structures for easy reference. Those symmetries that are hard to detect from visual examination of a crystal—glides and screws—are revealed easily at this stage of a study. One way to determine which reflections are present and which are absent is to systematically tilt a crystal on a diffractometer and hunt for these. From the external appearance of the crystal it is often possible to decide if it is a cubic structure or one of the other crystal systems. It is then possible to choose dimensions for the cell edges that would give the measured reflection angles and to decide on the indices of reflections present and those that are extinct. If the crystal is orthorhombic (see Figure 9), for example, then the a_1, a_2, a_3 axes (but not their lengths) can generally be assigned from the shape of the crystal. Then it is possible to mount the crystal to get the (100) or (010) or (001) reflections. It is known that the relation between the spacing of planes (d), the indices of the reflections (h,k,l), and the lengths of the edges of the unit cell (a_1, a_2, a_3) for orthorhombic crystals can be expressed:

$$\frac{1}{d^2} = \frac{h^2}{a_1^2} + \frac{k^2}{a_2^2} + \frac{l^2}{a_3^2} . \qquad (5)$$

Substituting for d in Bragg's law then yields:

$$\lambda = 2\sqrt{\left(\frac{h^2}{a_1^2} + \frac{k^2}{a_2^2} + \frac{l^2}{a_3^2}\right)} \sin\theta. \qquad (6)$$

The (100), (200), (300), or (400) reflection will permit determination of a_1; a_2 is obtained from (010), (020), (030), or (040); and a_3 from (001), (002), (003), or (004). These calculated values can then be tested in computing θ for other reflections, to decide whether a reflection was (100), or (200), etc., and to decide on absences. A more common procedure, though, is to surround the crystal with film, to catch many reflections at once while continually moving the crystal, and then to examine the pattern for dimensions and shape of the unit cell and missing reflections.

The powder method. It is not always possible to obtain single crystals of suitable size and perfection for structural studies with X-rays or neutrons. This is the major problem in extending structural studies to enzymes and other biological molecules. It is far easier to obtain powders. Also, most engineering materials, such as steel and ceramics, are polycrystalline; that is, they are composed of many small crystals or grains that are held together across their faces by bonding. The way in which a diffraction pattern can be obtained from a random array of such grains or particles is illustrated in Figure 13. Uniform cones of diffracted intensity are produced instead of the spots characteristic of single crystals. An adequate specimen can be prepared by coating a hair with an adhesive and dipping it into the powder, which is, in effect, a randomly oriented array of crystals. A camera lens about 60 mm (2.4 inches) in diameter yields patterns of

Commercial importance

lines in a few minutes; they are recorded on a thin film wrapped in a cylinder around the sample to intersect the diffracted rays. Such patterns are quite useful in solving simple structures. Also, the spacing and intensity of the lines are characteristic of the material doing the scattering —just like a fingerprint. A file of more than 21,000 patterns is available, and the routine use of these for identification of unknown specimens constitutes one of the major uses of X-ray diffraction. One interesting field for this method stems from the fact that paint pigments contain

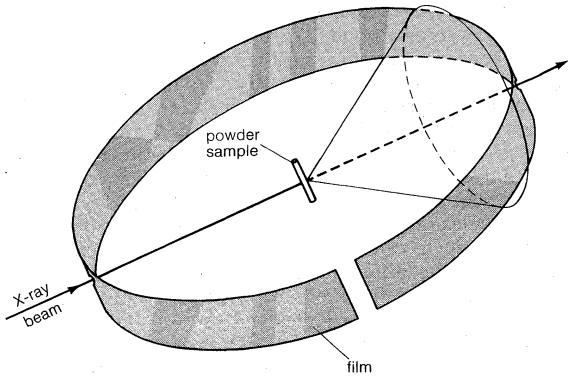


Figure 13: X-rays directed on a powder sample diffracting from the same planes in different grains. Because these planes are oriented differently in each grain, the result is a continuous cone of scattering at each diffraction angle. The continuous cones of diffracted radiation produce patterns of characteristic lines on a strip of film in a cylinder wrapped around the specimen.
From L.V. Azaroff, *Elements of X-Ray Crystallography*, copyright 1968; used with permission of McGraw-Hill Book Co.

crystalline materials, as do glazes. The composition of these has changed over the years with the art of paint manufacture, and it is sometimes possible to detect a forged art object from the diffraction patterns of the sample or to trace an automobile from a streak of its paint left after an accident.

By exploring the entire cones of scattering, it is possible to learn how random a specimen actually is. This can be very important in controlling properties, because these depend on direction in a single crystal. Processing of a polycrystalline material is often adjusted in order to provide a certain deviation from complete randomness as an aid in obtaining desired properties in certain directions.

IMPERFECTIONS IN CRYSTALLINE MATERIALS

Most crystalline materials are not perfectly regular. Accidents in growth lead to imperfections. Common among these are vacant lattice sites (vacancies) and missing parts of planes of atoms, as shown in Figure 14. Mechanical and electrical properties can be highly dependent on the number and arrangement of such imperfections. None of these defects can be seen with the naked eye or with an optical microscope, although in many cases they can be observed with an electron microscope. The presence of defects was first postulated by investigators to explain certain properties of materials long before they were seen.

Arrangement of defects. If the defects have a regular arrangement on an atomic scale, these can give rise to extra peaks in the diffraction pattern from which their arrangement can be discovered. As an example, there are whole groups of oxide compounds whose specific compositions can be understood as due to a regular arrangement of blocks of simpler structures with defects between the blocks.

If the defect arrangement is less regular but the defects are still numerous, there will be no sharp X-ray lines from the arrangement; but, instead, weak extra scattering of a variey of forms may be found between Bragg peaks, or the peaks may broaden or shift in position. A great deal is being learned from this extra scattering about the density and arrangement of such defects in materials and how they change with treatment, essentially from mathematical analyses of the effect of various defects on diffraction and comparisons of the calculations with experimental patterns. In fact, it is even possible to learn a great deal about amorphous materials from diffraction. Local atomic arrangements or groupings in which there is only a tendency toward certain arrangements at short interatomic distances are adequate to cause fluctuations in scattered intensity. The structure of glass consists of linked SiO_4 tetrahedra—that is, tetrahedra with silicon at the centre of each tetrahedron and oxygen at the corners. These corners are shared by two tetrahedra to hold the

Diffraction studies of amorphous materials

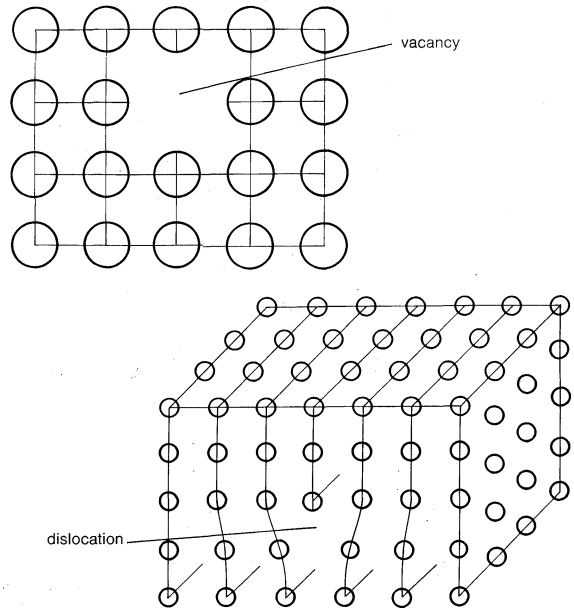


Figure 14: *Defects in a crystal caused by accidents of growth.*
(top) A missing atom, or vacancy. (bottom) A plane that does not pass entirely through the crystal; this dislocation leads to distortions of neighbouring planes.

structure together. This arrangement is similar to, but less periodic than, the structure of crystalline silica, as indicated from the analysis of scattering from amorphous glass. Considerable progress in understanding amorphous structures is expected in the near future because of increasing interest in amorphous materials. Glasses and also certain semiconductors and many plastics are amorphous.

Observation by electron microscopy and X-ray methods. Many materials have a sufficiently low density of defects to make attractive an exciting new expression of diffraction. With the electrons in modern electron microscopes operating at 50,000 to 1,000,000 volts, it is possible to penetrate foils of materials about 1000 angstroms and make pictures at magnifications of 1,000,000 times. Such foils can be prepared by electropolishing. The wavelengths of the electrons at these voltages are about 0.01 angstrom (the value depends on the voltage of the microscope). The diffraction angles, from Bragg's law, are very small, approximating one degree. Small tilts near an imperfection, such as the dislocation in Figure 14, can bring a group of planes into the proper orientation for diffraction. The diffracted beam is blocked, but the unscattered beam passes through an aperture. A dark line appears in the position of the dislocation on a fluorescent screen which is illuminated by the electrons to form the image. It is also possible to heat, cool, or deform the specimen in the microscope and observe the change in number and arrangement of defects, as well as their interaction.

With X-rays, quite similar pictures can be taken. One advantage of this approach is that it can be employed in reflection on thick specimens rather than on the foils needed for an electron microscope. One major disadvantage, however, is that the pictures are not taken at any enlargement. Photographic enlargement of the best of films is possible up to only about 200 times. Because this is far removed from the magnification of an electron

microscope, the density of defects must be much lower to resolve them. The principle of the method for the longer wavelength X-rays is also somewhat different. On scattering from a plane, there is a 90° phase change of the scattered wave compared to the incident wave. If a region is nearly perfect, the diffracted beam will rescatter back in the direction of the incident beam before leaving the crystal, but the 90° change in phase a second time leads to a total change of 180°; this twice-scattered beam is thus out of phase with the incident beam and reduces it, lessening its penetration into the specimen. This effect reduces the diffracted intensity in two ways. First, less material "sees" the incident beam; and, second, some of the diffracted beam is rescattered. In the vicinity of an imperfection, less of this occurs, and the scattered intensity is larger than for the perfect regions, making a darker region on a film placed above the specimen.

Such techniques are of great importance in the semiconductor industry because the good electrical properties of these devices require low defect densities. The entire industry, in fact, is built on the ability to produce in quantity highly perfect crystals of such elements as silicon—a triumph for the modern science of materials.

BIBLIOGRAPHY

Texts on crystal structure, morphology, optical techniques, and properties: E.A. WOOD, *Crystals and Light* (1964); ALAN HOLDEN and PHYLIS SINGER, *Crystals and Crystal Growing* (1960); E.E. WAHLSTROM, *Optical Crystallography*, 3rd ed. (1960); J.F. NYE, *Physical Properties of Crystals* (1957); M.J. BUERGER, *Introduction to Crystal Geometry* (1971); and F.C. PHILLIPS, *An Introduction to Crystallography*, 4th ed. (1972). The first two books are stimulating nonmathematical introductions; the last two are excellent accounts of crystallography with a minimum amount of mathematics.

Basic texts on diffraction: B.D. CULLITY, *Elements of X-Ray Diffraction* (1956), is an excellent introduction to diffraction, requiring a minimum of mathematics. It is particularly good for information on techniques not associated with structure determination. Four basic texts that require more advanced knowledge of mathematics are J.B. COHEN, *Diffraction Methods in Materials Science* (1966); L.V. AZAROFF, *Elements of X-Ray Crystallography* (1968); M.M. WOOLFSON, *An Introduction to X-Ray Crystallography* (1970); and H. LIPSON and C.A. TAYLOR, *Fourier Transforms and X-Ray Diffraction* (1958).

Sources of information on specific structure and other data on crystalline materials: W.L. BRAGG and G.F. CLARINGBULL, *Crystal Structures of Minerals* (1965); and C.S. BARRETT and T.B. MASSALSKI, *Structure of Metals*, 3rd ed. (1966), are excellent up-to-date summaries of the techniques and results of crystallographic studies on metals and alloys. The *International Tables for X-Ray Crystallography*, 3 vol. (1952–62), constitutes a compendium of brief descriptions of all techniques and theory in crystallography and diffraction with many references. Excellent tables and pictures of space groups are included.

Texts on effects of imperfections on diffraction: ANDRE GUINIER, *Théorie et technique de la radiocristallographie*, 2nd ed. (1956; rev. Eng. trans., *X-Ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*, 1963); B.E. WARREN, *X-Ray Diffraction* (1969).

History of crystallography: P.P. EWALD (ed.), *Fifty Years of X-Ray Diffraction* (1962), celebrating the 50th anniversary of Laue's discovery of diffraction, contains excellent histories of various aspects of the subject and personal reminiscences by many of the key people in the field. JOHN C. KENDREW, *The Thread of Life* (1966), presents an excellent nontechnical account of structural studies on biological molecules.

(J.B.Co.)

Ctenophora

Ctenophores, frequently called sea walnuts, sea gooseberries, cat's eyes, or comb jellies, are common inhabitants of the seas. Though they are, for the most part, of small size, at least one Mediterranean species, the Venus' girdle, may attain a length of one metre. One parasitic species is only three millimetres in diameter. Some ctenophores live in somewhat brackish water, but all are confined to marine habitats. They live in almost all ocean regions, but particularly in surface waters near shores. At least two species (*Pleurobrachia pileus* and *Beroë cucumis*) are cosmopolitan, but most have a more restricted distribution. When abundant in a region they consume most of the young of fish, larval crabs, clams, and oysters, as well as copepods and other planktonic animals that would otherwise serve as food for such commercial fish as sardines and herring. In turn, however, ctenophores are themselves consumed by certain fish.

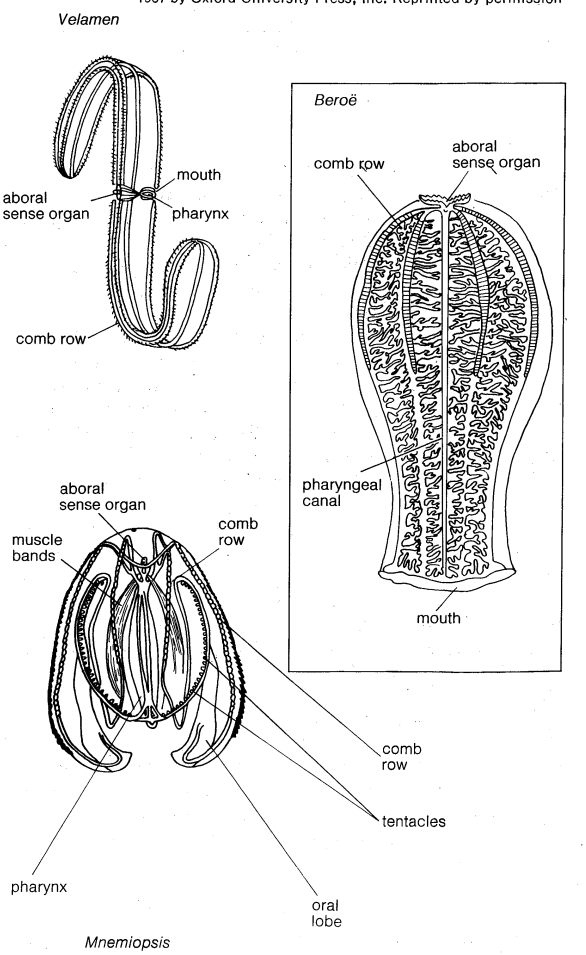
Natural history. A single ctenophore produces both eggs and sperms. The sex cells, earlier believed to be emitted through the mouth, are actually released by partial rupture of the meridional canals and the body covering over these canals (see below *Form and Function*). Only the highly specialized creeping species (order Platyctenea) have special gonoducts. Fertilization and embryonic development occur in the water. In *Pleurobrachia* and in other Cydippida, development is direct; there is no larval stage. More highly evolved ctenophores have a so-called cydippid larva that is similar in structure to the Cydippida adult. The cydippid larva undergoes metamorphosis to form the adult. Only in the parasitic *Gastrodes* has a free-swimming planula larva comparable to that of the coelenterates been reported.

Frequently comb jellies occur in vast swarms, especially in bays, lagoons, and other coastal waters. Except for a few creeping and parasitic species, comb jellies float freely suspended in the water, where they are important members of the plankton (see PLANKTON). Except for one parasitic species, all of them are carnivorous, devouring myriads of other small planktonic animals.

Some species are so extremely delicate that they are difficult to collect and preserve intact. Such forms avoid destruction in storms by descending from the surface into deeper waters.

Form and function. Most ctenophores are colourless, although *Beroë cucumis* is pink and the Venus' girdle (*Cestum veneris*) is delicate violet. The colourless species

From *Invertebrate Zoology* by Paul A. Meglitsch. Copyright © 1967 by Oxford University Press, Inc. Reprinted by permission



Diversity among ctenophores.

are as transparent as glass, so that, except for their beautifully iridescent rows of comb plates, they are nearly invisible when suspended in the water.

Most of the 80-odd known species are spherical or oval, with a conspicuous sense organ (the statocyst) at one end (aboral) of the body and a mouth at the other end (oral). Typically, eight rows of comb plates (or ctenes) extend orally from the vicinity of the statocyst. These serve as organs of locomotion. The comb plates are rows of very large, fused cilia. When they beat, the effective stroke is toward the statocyst, so that the animal normally swims oral end first. The more primitive forms (order Cydippida) have a pair of long, branched tentacles that function in the capture of food. These are completely retractile into recessed tentacle sheaths. The tentacles are richly supplied with peculiar lasso cells (or colloblasts), found only among ctenophores. Lasso cells produce a sticky secretion, to which prey organisms adhere on contact.

The mouth leads into a wide pharynx, from the aboral end of which arises a complex, branched series of canals that form the digestive tract. Since this structure serves both digestive and circulatory functions, it is designated a gastrovascular cavity. There is no true anus; the central canal opens toward the aboral end by two small pores, through which a small amount of egestion can take place.

Gonads develop as thickenings of the lining of the digestive canals. The nervous system is a primitive nerve network, somewhat more concentrated beneath the eight rows of comb plates. It is similar to the coelenterate nervous system. There is no trace of an excretory system.

Body covering

As in coelenterates, the outside of the body is covered by a thin layer of ectodermal cells, which also line the pharynx. A second thin layer of cells, constituting the endoderm, lines the gastrovascular cavity. Between the ectoderm and the endoderm is a thick gelatinous layer, the mesoglea. Because it contains not only many mesenchymal cells (or unspecialized connective tissue) but also specialized cells (*e.g.*, muscle cells), the mesoglea forms a true mesoderm. In this respect the comb jellies are more highly evolved than even the most complex coelenterates.

Most of the comb jellies are bioluminescent; they exhibit nocturnal displays of bluish or greenish light that are among the most brilliant and beautiful known in the animal kingdom. The bioluminescence begins only after the animals have been in the dark for at least 20 minutes; the ability to luminesce is terminated by exposure to light (see BIOLUMINESCENCE).

Taxonomy and evolution. The primitive order Cydippida, of which *Pleurobrachia* is a representative, has a pair of well-developed tentacles. In the order Lobata (which includes *Bolinopsis* and *Mnemiopsis*) the tentacles are greatly reduced and the oral lobes greatly enlarged. The body is somewhat flattened laterally. In the order Cestida the body is greatly extended, forming a thin band. Tentacles are very small in the adults. The cestids include the Venus' girdle mentioned earlier. The order Beroidea (*Beroë*) has no tentacles in the adult, the rows of comb plates are short, and the digestive canals are highly branched. The greatest modifications from primitive conditions occur in members of the order Platyctenea, in which the body is orally-aborally flattened, and the adults (*e.g.*, *Coeloplana*) creep around on solid surfaces.

Relationship to coelenterates

The relationship of ctenophores to other animals has caused much speculation. Formerly they were considered to be a class or subphylum of the phylum Coelenterata (or Cnidaria). Although some textbooks still maintain this classification, most modern authorities separate Ctenophora as an independent phylum. The separation is based upon (1) the lack of the characteristic stinging cells (cnidoblasts) of coelenterates, (2) the existence of a definite mesoderm in the ctenophores, and (3) fundamental differences in embryological development between the two phyla.

It is, however, generally conceded that ctenophores and coelenterates were derived from the same evolutionary stem. This is indicated by (1) the primitive radial or biradial organization of the body in both groups, (2) the extensive development of mesoglea between the ectoderm and the endoderm, and (3) the existence of a planula larva (characteristic of the coelenterates) in the ctenophore *Gastrodes*.

The body shape of creeping species of Platyctenea has suggested to some authorities that the ctenophores are transitional forms lying between coelenterates and turbellarian flatworms (*Platyhelminthes*). Although there is a certain superficial resemblance between the body of the Platyctenea and that of free-living flatworms, as well as some similarities in certain aspects of their embryonic development, many fundamental differences between them make it improbable that the ctenophores lie directly in a coelenterate-platyhelminth line of evolution. It is more likely that both turbellarian flatworms and the creeping ctenophores have evolved by parallel evolution from completely different ancestors.

BIBLIOGRAPHY. F.M. BAYER and H.B. OWRE, *The Free-Living Lower Invertebrates* (1968), a broad survey of the morphology, classification, physiology, and ecology of lower invertebrates; G.C. BOURNE, "Ctenophora," in E.R. LANKESTER (ed.), *Treatise on Zoology*, pt. 2, ch. 7 (1900), in a classic series of volumes dealing with technical details of structure and classification; A.C. HARDY, *The Open Sea* (1956), a modern comprehensive treatment with special reference to the planktonic organisms; S.J. HICKSON, "Coelenterata and Ctenophora," in *Cambridge Natural History*, vol. 1 (1906), in a classic series of volumes; L.H. HYMAN, *The Invertebrates: Protozoa Through Ctenophora*, vol. 1 (1940), an authoritative series dealing with the biology of invertebrates; G.E. and N. MACGINITIE, *Natural History of Marine Animals*, 2nd ed. (1968), a modern consideration of the behaviour and habitats of various marine animals.

(C.C.D.)

Cuba

Cuban archipelago

Territorially, the Republic of Cuba comprises the Cuban archipelago, a formation of no fewer than 3,715 islands, islets, and cays with a combined area of 42,827 square miles (110,922 square kilometres). The archipelago is situated near the geographical centre of the New World landmass, just south of the Tropic of Cancer at the entrance to the Gulf of Mexico, and forms an important segment of the Greater Antilles island chain, which continues east and then south in a great arc enclosing the Caribbean Sea. The island of Cuba itself—the "pearl of the Antilles"—is the largest island in the chain, covering 40,543 square miles (105,007 square kilometres). In general, the island runs from northwest to southeast and is long and narrow—777 miles (1,250 kilometres) long but only 119 miles (191 kilometres) across at its widest and 19 miles (31 kilometres) at its narrowest point. From Maisí lighthouse on the eastern tip, Haiti, the nearest neighbouring country, is visible 48 miles away across the Windward Passage. Jamaica, 87 miles to the south, is also visible on a clear day. The United States is about 90 miles to the north across the Straits of Florida.

The Isla de Pinos (Isle of Pines, now often called the Isle of Youth because of an influx of young settlers) is the second largest in the archipelago (849 square miles), rising to the southwest of Cuba itself. Other island and shoal groups are Los Colorados, to the northwest; the Archipiélago de Sabana-Camagüey, off the north central coast; the Jardines de la Reina (Queen's Gardens), off the southeast coast; and the Archipiélago de los Canarreos (technically including the Isla de Pinos), along the southwest coast.

Because of its geographical location and natural resources, Cuba was coveted by more than one foreign power over the centuries. A colony of Spain after its discovery by Columbus in 1492, it formally became a republic at the beginning of the 20th century, although its independent status was qualified by a high degree of political and economic dependence on the United States. On New Year's Day, 1959, the republic entered a new era with the victory of revolutionary forces led by Fidel Castro.

This article deals with contemporary Cuba. For history see CUBA, HISTORY OF, and LATIN AMERICA AND THE CARIBBEAN, COLONIAL; see also HAVANA.

THE LAND

The natural environment. *Topography.* Mountains cover about a quarter of the total area of the island of Cuba, often interrupted by the extensive plains that cover some two-thirds of the surface. The coastal basins of Santiago de Cuba and Guantánamo lie in the extreme east; a great central valley also begins in the east and then combines with a peneplain, continuing westward across the entire island. It is these plains that have been hospitable to the raising of sugarcane and livestock.

The *alturas*—regions of moderate elevation—are in some cases residues of formerly higher surfaces. More rugged relief includes the Guaniguanico range in the west, comprising the Sierra de los Órganos and the Sierra del Rosario, which attains 2,270 feet (692 metres) at El Pan de Guajaibón; the Sierra de Trinidad in the central region, with the 3,793-foot (1,156-metre) Pico San Juan; and the Sierra Maestra, more than 100 miles long, that rises abruptly from the southeast coast and contains the island's highest peaks, with Pico Turquino, at 6,476 feet (1,974 metres), pre-eminent.

Mountain regions and the coastline

Cuba possesses an irregular 3,570-mile (5,746-kilometre) coastline, made picturesque by the many bays, sandy beaches, mangrove plantations, swamps, coral reefs, and rugged cliffs. There are also some spectacular caverns in the interior, notably the Ensenada Don Tomás (Santo Tomás) of the Quemado ridge region, which has a linear extension of 16 miles. The 40,543 square miles of the island of Cuba are surrounded by a submerged platform that is an additional 30,000 square miles (70,000 square kilometres) in area; hence sea depths surrounding the island are no greater than about 650 feet (200 metres), dropping abruptly away at the outer edges of the platform.

Drainage and soils. Cuba's excellent supply of groundwater is utilized intensively throughout the island but especially in La Habana province. The rivers are generally short, with very meagre flow; of 563 watercourses classified as rivers, 236 discharge to the north, the remainder to the south.

The island's heaviest rainfall and therefore its largest rivers are in the southeast, where the Cauto (at 230 miles the country's longest) and its tributaries, notably the Salado, drain the Sierra Maestra and the uplands to the north. Other major rivers in this region include the Guantánamo, Sagua de Tánamo, Toa, and Mayarí. Proceeding westward, the most important rivers flowing south are the Sevilla, Najasa, San Pedro (of Camagüey), Jatibonico del Sur, Zaza, Agabama, Arimao, Hondo, and Cuyaguateje; north, the Saramaguacán, Caonao, Sagua la Grande, and La Palma.

Cuban lakes are small and more properly classified as freshwater or saltwater lagoons. The latter include the 26-square-mile Laguna de Leche (Milky Lagoon), which is technically a sound, as it is connected to the sea by three natural channels. Sea movements generate disturbances in the calcium carbonate bottom deposits to produce the milky appearance.

The complicated Cuban topography and geology have produced soils of no fewer than 13 different major groups, the majority of which are fertile, in good physical condition, and amenable to year-round cultivation.

Climate. Cuba lies in the tropic zone, located on the southwestern periphery of the North Atlantic High atmospheric pressure zone and hence influenced by the Northeast Trade Winds in winter and east–northeast winds in summer. The warm currents that later form the Gulf Stream are a year-round ameliorating influence along the coasts.

Temperatures

Annual mean temperature is 77.9° F (25.5° C), with little variation between January (at 72.5° F [22.5° C] the coolest month) and August (the warmest month, at 82.0° F [27.8° C]). The November–April dry season abruptly changes to the rainy May–October season. Annual rainfall averages 54 inches (1,380 millimetres). Between June and October, the country often is exposed to hurricanes, whose strong winds (up to 163 miles per hour) and heavy rains (up to 12 inches in a 24-hour period) have occasioned great economic damage and human suffering. In the western and central parts of the island, cold fronts occasionally induce precipitation even during the dry season.

Plant and animal life. Cuba's tropical plant life is very rich, with some 8,000 species represented, 6,000 of them "higher"; half of these are endemic. Much of the original vegetation has been replaced by sugarcane, coffee, and rice plantations, made possible by enormous and indiscriminate destruction of forests. The revolutionary government introduced extensive reforestation for both economic and conservation reasons, and official policy is to cover 20 percent of the island with forests. Significant planting is being carried out in Guane, in Pinar del Río province, and in Pinares de Mayarí, in the north of Holguín province.

Cuban timber is of excellent quality. Pine is found in abundance, and the durable mahogany is of great economic importance, while ebony (*Diospyros*) and granadillo (cocus, or West Indian ebony; *Brya ebenus*) are both beautiful and valuable. The royal palm, of which there are more than 20,000,000 specimens, reaching heights of 50 to 75 feet (15 to 23 metres), is the national tree and a characteristic element of the rural landscape. The ceiba (kapok) tree plays a role in many legends, while the extremely rare cork palm (*Microcycas calocoma*) of the western regions is a "living fossil" thought to have survived since the Cretaceous Period, more than 100,000,000 years ago.

Fruit trees include such citrus varieties as lemon, orange, and grapefruit; some species of the genus *Annona*, including the *guanábana* (soursop) and *anón* (sweetsop); and avocado and papaya. Banana plants are also common. The lower Cuban coasts and the shoals of the archipelago are lent marked character by the mangrove, and the tobacco plant is valued for its product and as an ornamental.

Cuban animal life is extraordinary in its abundance and variety, particularly the invertebrate species. The archipelago is the home of more than 7,000 insect species and 4,000 species of land, river, and sea mollusks. Sponges (the basis of an industry) are important off the southwestern coast, and edible crustaceans abound. The similarly profuse Arachnida include the tarantula and scorpion. Fish (more than 500 edible species) are economically the most important vertebrate group, and there are 37 shark species. There are far fewer freshwater fishes.

Invertebrates

Cuba is visited by many migratory birds, and only a third of the 300 or so species found on the island are typically Cuban; these include the flamingo, royal thrush, nightingale, mockingbird, and hummingbird.

Reptiles are distributed equally among sea, river, and dry land species. Marine species include the tortoise and hawkbill turtle; mud turtles inhabit the rivers; the marshes contain two species of crocodile (formerly almost extinct but now the object of a repopulation program); and land reptiles include the iguana and the nonpoisonous *majá de Santa María* (*Epicrates angulifer*), Cuba's largest snake. Amphibians are similarly varied, with 60 frog and toad species, the former including the plantain frog (*Hyla septentrionalis*) and bullfrog. The solenodon (*Atopogale cubana*), an almost extinct, ratlike insectivore, is found only in the remotest eastern regions. Other mammals include the hutia (an edible rodent) and the manatee, or sea cow, which inhabits river mouths. Bats (30 species) destroy mosquitoes as well as insects harmful to agriculture and in their roosting caves produce accumulations of guano that is valuable as fertilizer.

The human imprint. The various areas of Cuba do not differ greatly in customs and traditions, making a clear regional division on this basis difficult. The eastern end of the island (provinces of Guantánamo, Santiago de Cuba, Las Tunas, Granma, and Holguín)—with its own speech patterns, music, and customs, resulting from regional isolation—is an exception. A more significant division might be that between city and countryside,

especially between the dominant Havana metropolitan area and the rest of the country, a pattern of unequal development repeated on a much smaller scale in the various provincial capitals.

The homogeneity of Cuban society has been increased since the revolution of the 1950s by the emphasis placed by the government on the development of communications and on rural growth. Given these qualifications, a division of Cuba into five natural regions—Occidental (Western), Central, Camagüey-Maniabón, Oriental (Eastern), and Pinera (the Isla de Pinos, or Isle of Pines)—is still possible on the basis of natural endowment and human exploitation. Cuban specialists have made a further subdivision into 45 subregions and 38 physiographic zones.

The Occidental region. The largest of Cuban regions, the Occidental region is 332 miles (535 kilometres) long and includes the provinces of Pinar del Río, La Habana (Havana), Matanzas, and parts of Villa Clara and Cienfuegos. Its topography is notable for the *mogotes*, unusually shaped elevated hummocks, although it is primarily composed of an enormous plain that occupies its entire southern section. The 110-mile-long Cordillera de Guaniguanico and the serpentine highlands of northern and central La Habana and Matanzas provinces lend further character to the region. The Occidental region contains nearly 42 percent of the country's population and also Havana, the major Cuban economic, cultural, and administrative centre. The regional economy is based on agriculture (some of the best tobacco in the world, citrus fruits, sugarcane, cattle, rice, and coffee), copper and iron mining, and fishing.

Cordillera de Guaniguanico

The Central region. Central region extends from the Manacas plain (west of the city of Santa Clara) to the Ciego de Avila plain in Ciego de Avila province and consists primarily of the Trinidad and Sancti Spíritus highlands and lesser uplands, interspersed with plains and boggy regions in the province of Sancti Spíritus. It also includes the Archipélago de Sabana-Camagüey.

The major Cuban tobacco-growing area, the Central region also produces sugarcane, cattle, citrus fruits, and coffee as well as marble, copper, iron, and limestone. Industrial activity in the region is varied, and petroleum exploration continues.

The Camagüey-Maniabón region. The Camagüey-Maniabón region, Cuba's main stock-raising area, centres on an ancient peneplain that exends over central Camagüey province and northwestern Las Tunas province between the Trinidad, Sancti Spíritus, and Maniabón highlands. Other economic activities include a major emphasis on sugarcane, cultivation of rice and grains, mining of chromium, and processing of cement, fertilizer, and gypsum.

The Oriental region. At the eastern tip of the main island, the provinces of Las Tunas, Granma, Holguín, Santiago de Cuba, and Guantánamo have the nation's highest mountains, fastest streams, richest mines, and most spectacular bays. The Sierra Maestra rises along the southern coastline, the Sierra de Nipe on the northern, and between them lies the economically important Valle Central, its easternmost portion rising to the Sagua-Baracoa highlands, covered by the most extensive forests in Cuba. The region contains nickel reserves and also produces chromium, iron, copper, and manganese as well as sugarcane and other agricultural items; its industrial centres process minerals and food and generate electric power.

The Pinera region. The largest and most beautiful of

MAP INDEX

Political subdivisions

Camagüey....21·30n 78·10w
Ciego de Avila...21·55n 78·40w
Cienfuegos....22·15n 80·30w
Granma....20·25n 77·00w
Guantánamo....20·20n 75·00w
Holguín....20·50n 76·00w
La Habana....22·45n 82·10w
La Habana, Ciudad de....23·08n 82·22w
Las Tunas....21·00n 77·05w
Matanzas....22·40n 81·10w
Pinar del Río....22·35n 83·40w
Sancti Spíritus...22·00n 79·25w
Santiago de Cuba.... 20·20n 76·00w
Villa Clara....22·30n 80·00w

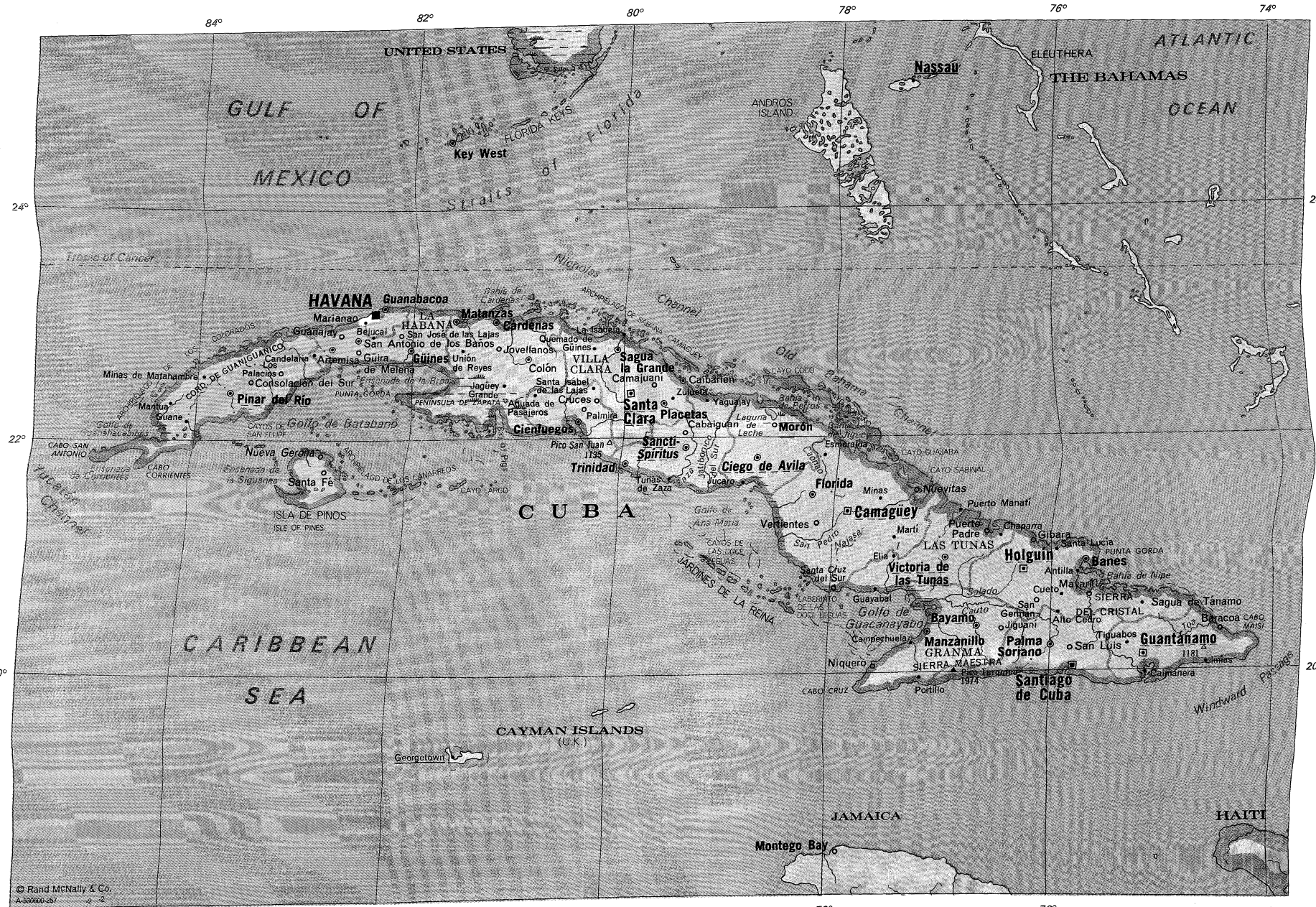
The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Aguada de Pasajeros....22·23n 80·51w
Alto Cedro....20·31n 75·58w
Antilla....20·50n 75·45w
Artemisa....22·49n 82·46w
Banes....20·58n 75·43w
Baracoa....20·21n 74·30w
Bayamo....20·23n 76·39w
Bejucal....22·56n 82·23w
Cabaiguán....22·05n 79·30w
Caibarién....22·31n 79·28w
Caimanera....19·59n 75·09w
Camagüey....21·23n 77·55w
Camajuaní....22·28n 79·44w
Campechuela....20·14n 77·17w
Candelaria....22·44n 82·58w
Cárdenas....23·05n 81·10w
Chaparra....21·10n 76·29w
Ciego de Avila...21·51n 78·46w
Cienfuegos....22·09n 80·27w
Colón....22·43n 80·54w
Consolación del Sur....22·30n 83·31w
Cruces....22·21n 80·16w
Cueto....20·39n 75·56w
Elia....20·59n 77·26w
Esmeralda....21·51n 78·07w
Florida....21·32n 78·14w
Gibara....21·07n 76·08w
Guanabacoa....23·07n 82·18w
Guanajay....22·55n 82·42w
Guane....22·12n 84·05w
Guantánamo....20·08n 75·12w
Guayabal....20·42n 77·36w
Güines....22·50n 82·02w
Güira de Melena.22·48n 82·30w
Havana....23·08n 82·22w
Holguín....20·53n 76·15w
Imías....20·04n 74·38w
Jagüey Grande...22·32n 81·08w
Jiguaní....20·22n 76·26w
Jovellanos....22·48n 81·12w
Júcaro....21·37n 78·51w
La Isabela....22·57n 80·01w
Los Palacios....22·35n 83·15w
Mantua....22·17n 84·17w
Manzanillo....20·21n 77·07w
Marianao....23·05n 82·26w
Martí....21·09n 77·27w
Matanzas....23·03n 81·35w
Mayarí....20·40n 75·41w
Minas....21·29n 77·37w
Minas de Matahambre....22·35n 83·57w
Morón....22·06n 78·38w
Niquero....20·03n 77·35w
Nueva Gerona...21·53n 82·48w
Nuevitas....21·33n 77·16w
Palma Soriano...20·13n 76·00w
Palmira....22·14n 80·23w
Pinar del Río....22·25n 83·42w
Placetas....22·19n 79·40w
Portillo....19·55n 77·11w
Puerto Manatí...21·22n 76·50w
Puerto Padre....21·12n 76·36w
Quemado de Güines....22·48n 80·15w
Sagua de Tánamo....20·35n 75·14w
Sagua la Grande.22·49n 80·05w
San Antonio de los Baños....22·53n 82·30w
Sancti Spíritus...21·56n 79·27w
San Germán....20·36n 76·08w
San José de las Lajas....22·58n 82·09w
San Luis....20·12n 75·51w
Santa Clara....22·24n 79·58w
Santa Cruz del Sur....20·43n 78·00w
Santa Fé....21·45n 82·45w
Santa Isabel de las Lajas....22·25n 80·18w
Santa Lucía....22·40n 83·58w
Santiago de Cuba....20·01n 75·49w
Tiguabos....20·14n 75·21w
Trinidad....21·48n 79·59w
Tunas de Zaza...21·38n 79·33w
Unión de Reyes..22·48n 81·32w
Vertientes....21·16n 78·09w
Victoria de las Tunas....20·58n 76·57w
Yaguajay....22·19n 79·14w
Zulueta....22·22n 79·34w

Physical features and points of interest

Ana María, Golfo de, *gulf*....21·20n 78·45w
Batabanó, Golfo de, *gulf*....22·15n 82·30w
Broa, Ensenada de la, *bay*....22·35n 82·00w
Canarreos, Archipiélago de los, *islands*....21·50n 82·30w
Caonao, *river*....22·05n 78·05w
Cárdenas, Bahía de, *bay*....23·05n 81·09w
Caribbean Sea...20·00n 83·00w
Cauto, *river*....20·33n 77·15w
Cochinos, Bahía de, see Bay of Pigs
Coco, Cayo, *island*....22·30n 78·28w
Colorados, Archipiélago de los, *islands*...22·36n 84·20w
Corrientes, Cabo, *cape*....21·45n 84·31w
Corrientes, Ensenada de, *bay*....21·50n 84·35w
Cristal, Sierra del, *mountains*....20·33n 75·30w
Cruz, Cabo, *cape*.19·51n 77·44w
Doce Leguas, Cayos de las, *islands*....20·55n 79·05w
Florida, Straits of.24·00n 81·00w
Gorda, Punta, *point*....21·05n 75·39w
Gorda, Punta, *point*....22·24n 82·10w
Guacanayabo, Golfo de, *gulf*...20·30n 77·35w
Guajaba, Cayo, *island*....21·50n 77·30w
Guanahacabibes, Golfo de, *gulf*...22·08n 84·35w
Guaniguanico, Cordillera de, *mountains*....22·30n 83·40w
Jardines de la Reina, *islands*...20·45n 78·50w
Jatibonico del Sur, *river*....21·33n 79·09w
Jigüey, Bahía de, *bay*....22·07n 78·06w
Laberinto de las Doce Leguas, *island*....20·35n 78·30w
Largo, Cayo, *island*....21·38n 81·28w
Leche, Laguna de, *lagoon*....22·13n 78·38w
Maestra, Sierra, *mountains*....20·00n 76·45w
Maisí, Cabo, *cape*....20·17n 74·09w
Mexico, Gulf of...24·30n 84·30w
Najasa, *river*....20·42n 77·58w
Nicholas Channel....23·25n 80·05w
Nipe, Bahía de, *bay*....20·47n 75·42w
Old Bahama Channel....22·33n 78·05w
Perros, Bahía de, *bay*....22·25n 78·36w
Pigs, Bay of....22·07n 81·10w
Pinos, Isla de (Pines, Isle of), *island*....21·40n 82·50w
Romano, Cayo, *island*....22·15n 78·00w
Sabana-Camagüey, Archipiélago de, *islands*....23·00n 80·00w
Sabinal, Cayo, *island*....21·40n 77·15w
Salado, *river*....20·38n 76·57w
San Antonio, Cabo, *cape*....21·52n 84·57w
San Felipe, Cayos de, *islands*....21·58n 83·30w
San Juan, Pico, *peak*....21·59n 80·08w
San Pedro, *river*..21·09n 78·30w
Siguanea, Ensenada de la, *bay*....21·40n 83·05w
Toa, *river*....20·23n 74·32w
Turquino, Pico, *peak*....19·59n 76·51w
Windward Passage....19·30n 74·30w
Zapata, Península de....22·20n 81·45w
Zaza, *river*....21·37n 79·33w



© Rand McNally & Co.
A-530600-257

CUBA

Size of symbol indicates relative size of town • ○ ◉ ▣ ■ ■

Elevations in metres

the islands surrounding the main island of Cuba, the Isla de Pinos, dotted with groves of pine and palm, has sand and clay plains in the north and hills in both the northwest and southeast; nearly 40 percent of the island, however, is taken up by the gravel bed in the south and by the bogs of the coasts and uninhabited interior. At one time sparsely populated, it had by the mid-1970s a population approaching 40,000, mostly young people whose migration was sponsored by the government. They are employed in fruit growing, stock raising, and fishing, as well as in small industries and in the mining of marble and kaolin.

THE PEOPLE

Ethnic origins and national composition. For more than four centuries diverse ethnic groups have been settling in Cuba. Not only Africans and Spaniards (the predominant elements) but also Chinese, European Jews, and Yucatecans have all superimposed their cultural and social characteristics on those of the earliest settlers. Contemporary Cuban society exhibits a remarkable diversity as a result.

The indigenous heritage. Cuba's original inhabitants came to the island from South America. They were the Guanahatabey and the Ciboney, the former living in the extreme west of the island of Cuba, the latter in various places in the island and particularly on the cays to the south. Both were hunter-gatherers. The Taino, who arrived later and who spread over not only Cuba but also the rest of the Greater Antilles and the Bahamas, lived in villages and had rudimentary agriculture; they also made simple pottery. The Taino constituted 70 to 80 percent of the island's population at the time of the Spanish conquest.

Aboriginal peoples

In 1511 the total indigenous population was estimated at between 80,000 and 100,000, very unequally distributed, with density decreasing westward. Half a century after the Spanish invasion, however, only about 4,000 scattered individuals remained. Harsh treatment by the invaders, hard labour in the gold mines, hunger resulting from low agricultural productivity, and contagious diseases introduced by Europeans had all taken their toll. A few families with Taino physical characteristics living in the Sierra del Purial of easternmost Cuba are perhaps the only surviving descendants.

The African heritage. The Spaniards soon imported African slaves (more than 800,000 in all, most of them for work on the sugar plantations) as a substitute for the rapidly disappearing Indians. The Africans came mainly from Senegal and the Guinea Coast, with diverse origins including Yoruba and Bantu tribal backgrounds. Between 1919 and 1926, some 250,000 black Antillean labourers, 90 percent from Haiti and Jamaica, arrived under labour contract; nearly all remained. The cultural influence of blacks has been considerable, especially in music and dance.

According to the census of 1899, 14.9 percent of Cubans were black and 17.2 percent mestizo (or mixed). By 1943—largely as a result of heavy white immigration—the proportions had dropped to 9.7 and 15.6 percent, respectively. By 1953, however, following a dramatic decrease in black death rates after World War II, accompanied by a continued high birth rate, the proportions had changed again to 12.4 and 14.5 percent. In contemporary Cuba, using more refined anthropological criteria, it may be estimated that about 45 percent of the population is either mestizo or black. With the removal of historic discrimination barriers, mixed marriages are tending to increase, and black racial identity is being lost in the general movement toward an integrated society.

The European heritage. White Cubans are almost entirely of Spanish origin. The white population rose from 66.9 percent of the total in 1899 to 74.3 percent in 1943; almost 1,000,000 Spaniards and Canary Islanders immigrated between 1900 and 1929. By 1953, however, the percentage had dropped to 72.8.

Throughout Cuban history, the dominant classes were recruited primarily from the Europeans and their descendants, white and mestizo (the latter being of particular importance in the 20th century), who monopolized not only the direction of the economy but also access to education and culture.

The Asian heritage. In order to supplement the interrupted slave trade, the Hispano-Cuban landholders imported 125,000 indentured Chinese labourers, nearly all of them Cantonese, between 1853 and 1874, to work under contract for eight years. Bad living conditions reduced their numbers to 14,000 by the census of 1899. In the 1920s an additional 30,000 Cantonese and small groups of Japanese also arrived; both immigrations were exclusively male, and there was rapid intermarriage with white, black, and mestizo populations. In contemporary Cuba, the remaining Chinese element is aged and near extinction.

Indentured Chinese

Language and religion. Spanish is the Cuban national language; there are no local dialects, although the diversity of ethnic origins has influenced speech patterns. Some words are of native Indian origin, and a few (such as *hamaca* ["hammock"]) have passed into other languages. Africans have also enriched the vocabulary and contributed the soft, somewhat nasal accent and the rhythmic intonation that distinguish contemporary Cuban speech.

Religions in Cuba include Roman Catholicism, Santería (a cult devoted to certain African divinities formally identified with Catholic saints), and a number of Protestant and other groups. In the early 1960s, church and state confronted one another with open hostility, the church seeing the revolutionary government as anti-religious (one aspect of this being the nationalization of all schools) and the government seeing the church (the largest mass organization in the country) as a repository of counter-revolution. Many priests and nuns, and some Protestant clergy, left the country; others were deported. By 1965, however, the church, with a clergy drastically reduced in number, and the government entered upon a period of better cooperation.

Cuba, Area and Population

	area*		population
	sq mi	sq km	1970 census
Provinces (*provincias*)			
Camagüey	5,457	14,134	540,000
Ciego de Avila	2,504	6,485	273,000
Cienfuegos	1,602	4,149	296,000
Ciudad de La Habana	286	740	1,900,000†
Granma	3,263	8,452	650,000
Guantánamo	2,458	6,366	417,000
Holguín	3,516	9,105	772,000
Isla de Pinos‡	849	2,199	30,000
La Habana	2,190	5,671	525,000
Las Tunas	2,461	6,373	386,000
Matanzas	4,505	11,669	495,000
Pinar del Río	4,193	10,860	548,000
Sancti Spíritus	2,601	6,737	366,000
Santiago de Cuba	2,449	6,343	793,000
Villa Clara	3,115	8,069	701,000
Total Cuba§‖	{41,448 {42,827	{107,351 {110,922	8,693,000†

*Provincial areas exclude certain portions of the national territory (totalling 1,379 sq mi [3,571 sq km], approximately the area of outlying islands and cays). †Population figure for Havana is an estimate made in 1975; for this reason the total population figure disagrees with the 1970 population figure given in the text. ‡Special municipality administered by the central government. §First total includes areas under the administration of the province named; second total includes all areas under national sovereignty. ‖Detail may not add to total given because of rounding.

Demographic trends. The 8,569,100 Cubans recorded in the census of 1970 represented a 47 percent increase over the figure in 1953 and a 2.16 percent annual geometric intercensal growth (compared with 2.09 percent in 1943–53 and 1.58 percent in 1931–43). Taking into account the fact that the pre-1953 national migration balance was always positive and that since 1953 the balance has tended to be negative (*i.e.*, there are more emigrants than immigrants, especially during 1959–73 when an estimated 600,000 persons left the country), it may be deduced that the population increase in 1953–70 was attributable to changes in birth and death rates. After

making further allowances for the change in Cuba's socio-economic structure after 1959 (which both broadened the social services available to the general population and simultaneously reduced demographic underregistration), it can be concluded that there has been an increase in the actual birth rate, coupled with a decrease in the death rate. By the mid-1970s, per 1,000 of population, these rates were 20.7 and 5.4 (compared with 36.9 and about 6.5 in 1962), respectively, according to the government's statisticians. Former areas of Oriente (the most populous) and Pinar del Río provinces have the highest birth rates, while La Habana province—the most densely populated—has the lowest.

In the 1953–70 period the age distribution of the population also changed, with a broadening of the base and the top of the age pyramid. Thus, the percentages in the 0–4 years, 5–9 years, and over 64 age categories were 12.6, 12.2, and 4.2 in 1953 but 13.5, 13.7, and 5.8 in 1970.

Internal migration

The most dramatic change, however, has been in internal migration trends. The former drain from the countryside to the slums of the Havana metropolitan area (paralleling the trend in many developing countries) has been halted by government planning, which calls for a more diversified net movement outward from the regions with high birth rates (area of the former provinces of Oriente and Pinar del Río) to all the other provinces, which now have actual net inward movements.

THE ECONOMY

The pre-revolutionary economy. The revolutionary authorities inherited an economy that was essentially agrarian and was widely held to be characterized by extensive production (*i.e.*, the utilization of large areas of land with minimal outlay and labour), by emphasis on a single crop, and by the presence of *latifundios,* large landed estates. These basic traits, the revolutionary government decided, had to be eliminated before any serious effort could be made to transform the essentially underdeveloped country. Two agrarian reform laws (1959 and 1963) made the state, by the 1970s, the owner of 70–80 percent of all farmland. Small-scale farmers were permitted to retain their farms, and many tenant farmers were given title to the land they had worked. With state control established over the former North American- and Cuban-owned *latifundios,* Cuban authorities initiated, through investment of resources and effort, a substantial modification of agriculture.

Dependence on foreign commerce

The structure of Cuban industry (especially the sugar sector) likewise was not in harmony with the new economic development plans. At the time of the revolution, industry was extremely dependent on foreign commerce, while foreign investments (notably of North American capital) were covering more than 50 percent of its costs. The structural deformations in the economy induced by this dependence were also evident in the prospects for future industrial development, which were then limited to those branches using domestic sugar, nickel, and tobacco resources.

During the 1960s the Cuban authorities devoted every effort to removing the dependence on Western capital. In 1972 a series of major agreements was concluded with the Soviet Union, postponing until 1986 repayment of an accumulated debt estimated at more than $3,000,000,000 U.S.; payments would be spread over 25 years without additional interest. The agreements covered a further loan of some $350,000,000 for industrial development and guaranteed prices for Cuban sugar and nickel.

The extent and distribution of resources. Cuban soil is very fertile, allowing for up to two crops a year, but agriculture was traditionally plagued by the unreliability of the annual rainfall. During 1959–75 dams with a total water-holding capacity of 139,600,000,000 cubic feet (3,950,000,000 cubic metres) were constructed, and it was hoped that the water-shortage problem finally was solved. Subterranean waters are an important additional resource for both agriculture and industry. Tractors and cane-gathering machines were imported, and there was an average of one tractor per 113 cultivated acres by 1973.

The search for new petroleum deposits continues; production in 1974 was only 167,968 metric tons, and the greater part of domestic needs is met by imports. Peat, concentrated in the Península de Zapata, is still the most extensive fuel reserve. Nickel, cobalt, and iron are the most important minerals, the laterite (iron ore) beds in Holguín province having world significance. There are also major chromite, magnetite, manganese, and copper reserves and lesser amounts of lead, zinc, gold, silver, and tungsten. Abundant limestone, rock salt, gypsum, and dolomite reserves and large kaolin and marble beds are also found on the Isle of Pines.

Fishing resources (there are 500 different species of edible fish) are localized in four zones around the major shoal and island groups bordering the island of Cuba.

Resource exploitation and sources of national income. *Agriculture and fishing.* The most important agricultural crop is sugarcane. Between 1970 and 1974, production ranged between 44,000,000 and 80,000,000 tons annually. Vast areas have been levelled and brought under irrigation, thus greatly expanding the acreage in sugarcane, and yields per acre have been greatly increased with the application of ever larger amounts of fertilizers. Introduction of the Australian mowing system is expected to diminish the use of hand labour.

Improvements in stock breeding

The number of cattle was increased by cessation of slaughter of reproducing cows, by irrigation projects that added to available pasture, by artificial insemination, and by expanded veterinary services. Crossbreeding was also very important, the aim being to produce acclimatized meat and milk producers. Thus Brahman (or Zebu) cattle, the dominant breed (resistant to tropical climate but low in milk yield), were crossed with Holsteins (productive but prone to illness in the Cuban environment) and also with Brown Swiss.

Apart from sugarcane, the chief agricultural products are rice (the main source of calories in the traditional diet) and citrus fruits. Tobacco, traditionally the country's second crop, is grown mainly in the Pinar del Río area in the west and also in the centre of the island. Other products include bananas, coffee, pineapples, sweet potatoes, potatoes, and beans. It is necessary, however, to import large amounts of food (especially grain) and oilseeds and cotton.

Between 1958 and 1974 the Cuban fishing catch increased sevenfold, and it continues to increase, largely as a result of heavy investment in fishing vessels.

Industry. Processing of food and sugar are the two most important sectors of Cuba's domestic industry, and sugar is by far the most important export earner. In 1969–70, Cuba's record year (when personnel and resources from other areas of the economy were diverted to the sugarcane fields for the harvest), 8,537,600 metric tons of centrifugal sugar were produced. A combination of prolonged drought and dropping world prices led Cuban planners to try to develop more diversified sources of industrial income in the late 1970s. Other important industrial products include cement, electric power, machinery (linked to agricultural needs), food and beverages, tobacco, cigars and cigarettes, paper, fertilizers, textiles, and footwear.

The volume of expenditure on construction has increased in recent years, the portion spent on servicing (rather than on new projects) having fallen from two-thirds to a third of the total over the 1960s.

Trade patterns. In 1958 some 70 percent of Cuban imports came from the United States. By 1961 this was down to 4 percent, and it soon dropped to zero under United States government blockade measures. Communist bloc countries now account for between 65 and 70 percent of Cuban import and export totals. Import composition, too, has changed: production equipment accounts for about a fifth of the total, while food and live animals account for nearly a quarter. Principal exports include sugar (crude and refined and molasses), nickel, tobacco (especially the cigars for which Cuba is famous), fish, and coffee.

Management of the economy. The Cuban economy is centrally directed in accordance with a national plan

worked up by the government under guidelines laid down by the directorate of the Partido Comunista de Cuba (PCC; Communist Party of Cuba). Practically all economic activities are state run, a small private sector existing only in agriculture. The institutional economic structure consists of the Central Planning Board (Junta Central de Planificación, or Juceplan), headed by the economics minister; the ministries and national organizations that control the economic sectors and basic activities; the various state enterprises (*empresas*); and the provincial delegations that direct the work of the factories and related services.

The national planning mechanism

Each work centre follows a distinctly Cuban organizational pattern. It is managed by a collective organization: the administrative board is responsible for production and service activities; the cells of the PCC support the outlining of jobs and politically orient members of the work centre; and the trade union, representing the workers, acts as a vehicle for the expression of their opinions.

Transportation. *Maritime transport.* In 1959 there was a small merchant fleet of 14 ships with 63,800 deadweight tons and a cargo capacity of 53,100 tons. Sixteen years later the number of vessels had been multiplied by more than 19, and there was nearly 10 times as much deadweight tonnage and eight times as much cargo capacity; refrigerated cargo capacity had increased more than 14 times. These increases were accompanied by a major increase in foreign commerce. Authorities intend that ultimately half the total volume of freight entering and leaving Cuban ports will be carried in Cuban ships. Coastal trade movements are based on the vessels of the Empresa de Navegación Mambisa (Mambisa Navigation Enterprise) and the Empresa Cabotaje de Oriente (Oriente Coastal Trading Enterprise). A passenger service on the Cauto River was inaugurated in 1968.

Rail transport. The line constructed between Havana and Bejucal in 1837 was the first in the Americas after those of the United States. The railway system deteriorated in the first years after the revolution of 1959 but was being restored by the mid-1970s. Track mileage is approaching 10,000 (more than half of it connecting sugar fields and mills), and some 10,000,000 passengers and 11,000,000 tons of freight are transported annually.

Other land transport. Total road mileage increased by 50 percent during the 1960s. The most important highway is the Carretera Central (Central Highway), which runs along almost the entire length of the island. Other major routes are the Vía Blanca (linking Havana with the Playa Varadero) and the Vía Mulata (connecting Baracoa, at the east end of the island, with the rest of the country). There are eight state transport enterprises (seven of them for passenger traffic); one of these is national, the rest are provincial in scope.

Automobile transport has been substantially downgraded in an attempt to adjust transport facilities to the realities of Cuban life. In the 1950s automobiles, trucks, and buses accounted for 71, 10, and 19 percent of vehicle imports, respectively. In the 1970s the percentages were 9, 81, and 10 for the same categories.

Air transport. The Empresa Consolidada Cubana de Aviación (Cuban Aviation Enterprise) operates the former four private airlines. There are daily flights between Havana and the major Cuban cities, some subsidiary flights from Santiago, and weekly flights to a number of European and South American cities and to Rabat, Morocco. Among international flights, the Moscow–Havana nonstop flight of the Soviet Aeroflot line is the world's longest.

ADMINISTRATION AND SOCIAL CONDITIONS

The constitutional background. Until the adoption of the constitution of February 24, 1976, Cuba had for some 36 years been governed either by the constitution of 1940 or by the post-revolutionary Ley Fundamental (Fundamental Law) of February 7, 1959, modelled upon the constitution but centralizing governmental power. The 1940 constitution had been suspended twice—from 1952 to 1955 by the dictatorship of Fulgencio Batista, and after 1959 when it was supplanted by the Ley Fundamental and by legislation that included the Agrarian Reform Law (May 17, 1959), the Urban Reform Law (October 14, 1960), the Nationalization of Education Law (June 6, 1961), and the Second Agrarian Reform Law (August 3, 1963).

Drafting of a new constitution to succeed that of 1940 began in 1965 and continued for the next 10 years; a preliminary draft was approved by the Politburo of the Central Committee of the Communist Party of Cuba in 1975, and the final version was approved by referendum on February 15, 1976, entering into force on February 24.

In October 1976, for the first time in 17 years, representatives for 169 municipal assemblies "of the people's power" were elected to give the people a more effective role in the running of their urban centres. President Castro said the election was "a significant step forward in the process of consolidation and institutionalization of the revolution." These 169 assemblies met subsequently in November to elect a 481-member National Assembly, as they had somewhat earlier elected delegates to the 14 provincial assemblies; each of the bodies in turn elected executive committees to carry on the day-to-day work of their respective administrative organs.

The National Assembly, at its inaugural session in December 1976, appointed a State Council of 31 members, headed by Castro, as president, and a Council of Ministers, headed by a chairman, also Castro, who was thus both head of state and of government. The State Council is the executive body of the state, carrying on the daily administration of the country between the twice-yearly sessions of the National Assembly.

Following the revolution, political parties were dissolved, and a single party was created out of the participating revolutionary organizations: the Movimiento 26 de Julio (26th of July Movement), the Partido Socialista Popular (Popular Socialist Party), and the Directorio Revolucionario 13 de Marzo (13th of March Revolutionary Directorate). In 1965 this single national party was officially designated the Communist Party of Cuba (Partido Comunista de Cuba). In the elections of 1976, candidates represented both the Communist Party and mass organizations.

The Communist Party of Cuba

The mass organizations were created after the revolution to replace former social organizations and are under the supervision of the government. They include the Confederation of Cuban Workers (Confederación de Trabajadores Cubanos, or CTC), reconstituted in 1970, with stated objectives to support the government, help improve managerial performance and labour discipline, and raise the political consciousness of workers. The National Association of Small Farmers (Asociación Nacional de Agricultores Pequeños, or ANAP) is composed of independent farmers, outside the system of collectivized state farms, who own about 20 percent of the total cultivated land. In 1960 the Committees for the Defense of the Revolution (Comités de Defensa de la Revolución, or CDR), which now enroll about 70 percent of the adult population, were created to maintain vigilance against "enemies of the revolution"; they are organized in every city, factory, and place of work and in many rural counties. The objective of the Federation of Cuban Women (Federación de Mujeres Cubanas, or FMC) is "to raise the ideological, political, cultural, and scientific level of women in order to incorporate them into the tasks assigned by the revolution"; it replaced a number of social clubs.

Committees for the Defense of the Revolution

Justice and the armed forces. In 1973 Cuba's judicial system was reorganized. The Tribunal Supremo Popular, divided into four chambers, became the main body of the new structure. Its jurisdiction includes criminal offenses, civil and administrative offenses, crimes against state security, and military offenses. The Tribunales Provinciales Populares deal with cases that warrant sentences of up to six years' imprisonment.

Cuban defense is based on the revolutionary armed forces, now equipped with sophisticated weaponry. The land forces are grouped in eastern, western, and central armies, and there is an effective air force and a navy.

Auxiliary forces include a part-time civil-defense body. Military service is compulsory for citizens between 17 and 45 years of age.

The Ministry of the Interior is charged with the maintenance of public order and state security, rehabilitation of prisoners and prison management, and fire fighting.

Social services. *Education.* The eradication of illiteracy was given high priority by the revolutionary government. In 1961, the "Year of Education," 707,212 men and women were officially declared to have achieved literacy. Dissolution of all private schools was one of the first acts of the revolutionary government, and a fundamentally altered, state-directed education system was introduced. It includes general education, 12 grades preceded by a preschool stage; higher, or university, education; teacher-training education; adult education, directed toward the eradication of residual illiteracy and toward continued study by working people; technical education, parallel to middle-level general education; language instruction; and specialized education. Education is free at all levels, with supplementary scholarships to cover living expenses and medical assistance.

The campaign against illiteracy

Fidel Castro once remarked that "education *is* the revolution," and the revolutionary government has acted on that view. Education expenditures are the highest in Latin America. In all schools, fewer than 1,000,000 students were enrolled in 1959, in contrast to nearly 3,100,000 in 1976.

Health. Medical care is free, and mortality rates have been reduced. It was reported that by 1974 infant mortality rates were about 28 per 1,000 live births—a rate that compares very favourably with the rest of the Americas, including some developed areas. By 1974 the official ratio of doctors to population was 1:1,000, a favourable ratio for a developing country. The Ministry of Public Health (Minsap) requires physicians to work for the rural medical service in needy areas for two years after graduation.

Housing. Urban real-estate rental was prohibited under the Urban Reform Law of 1960, and it was made possible for families to own their homes by paying the current rental sum for not less than five or more than 20 years. By 1970 some 270,000 families had acquired titles to houses and apartments in this fashion, and 100,000 rural families had achieved free use of formerly rented lands. The traditional rural *bohío* ("hut") is being slowly replaced by more modern housing units. New towns and fishing villages have also been constructed.

Social assistance and welfare. Homes for the aged are under the direction of Minsap, but the *círculos infantiles*, institutions for the day care of children under seven years of age, are run by the Federation of Cuban Women. The *círculos infantiles* and *jardines de la infancia* are both intended to free women to work. The women's federation also supervises more than 8,000 social workers. Physical education and sports, under a national body, are an integral part of Cuban education. The Ministry of Interior Commerce is in charge of the fixing of prices for, and distribution and sale of, foodstuffs and notions (*abarrotes*), working as closely as possible with community groups. In 1968 more than 55,000 till-then privately operated sales outlets were placed under the ministry.

Social indexes. Cuban society has been transformed since 1959. The traditional public graft and corruption and also prostitution, gambling, and vice became targets of a strenuous campaign, and the official claim that the campaign has been successful seems to be warranted. The long-term investment program has been accompanied by the elimination of mass unemployment as it existed in the 1950s and access to consumption by formerly economically marginal sectors of the population. There is, however, a marked imbalance between money in circulation and goods and services available. Wages and prices are rigidly controlled and quota systems strictly enforced. For example, workers are moved to areas where labour is scarce by simply assigning them ration cards valid only in the sectors where their services are required. Farm incomes are controlled by regulation of the storage of agricultural products and by production limits. Overall, private ownership of the means of production is now limited to the 200,000 or so small farmers, and the working class is officially ascribed a leading role in Cuban society.

Quota systems

CULTURAL LIFE AND INSTITUTIONS

The cultural life of the Cuban people has also undergone a major transformation as part of what is regarded by the authorities as an ongoing revolutionary process. The government believes that mass culture is essential to the fulfillment of its economic and social aims and since 1959 has played a leading role in cultural life. This policy has been embodied in a set of directives, such as that creating a system of cultural organizations to take artistic displays to the remotest regions. Traditional Cuban culture, it was felt, was generally limited to Havana (and, to a much smaller extent, the provincial capitals), and was almost entirely privately endowed and thus subject to the vagaries of private fortunes, tastes, and interests. The National Cultural Council (Consejo Nacional de Cultura, or CNC) directs a program of education in music, plastic arts, ballet, dramatic arts, and modern dance; by 1970 more than 10,500 students were enrolled in provincial and regional schools of the National School of Fine Arts (Escuela Nacional de Arte Cubanacán). Amateur groups led by CNC instructors are also popular.

The theatre. Cuban theatre has been state-supported since 1959, coming under CNC direction, except in the case of the university cultural extension departments. There are six national dramatic groups, whose directing councils create their own repertory. Their activities are centred on Havana, but they travel frequently in the provinces. Provincial theatre groups characteristically present their entertainments on tour. Cuban critical opinion acknowledges that the national theatre suffers from lack of maturity but emphasizes that the drive to seek a unique national expression while utilizing the full range of tradition has produced a number of important works. The Casa de las Américas (House of the Americas) has held a number of Latin-American theatrical festivals. The Casa del Teatro (House of the Theatre; attached to the International Institute of Theatre) is a major centre of information and research on Cuban and international themes, while the lyric theatre stimulates interest in opera and operettas by conducting exchange programs with other countries.

Touring theatre groups

Music and ballet. Cuban music has Spanish and African roots. Various organizations have disseminated modern influences, and soloists participate in exchange programs and tour rural areas under CNC sponsorship. Cuba's foremost contemporary artistic figure is undoubtedly the prima ballerina and founder (1948) of the Ballet Nacional de Cuba, and head of its school, Alicia Alonso, a dancer of international acclaim. The Ballet of Camagüey was established in 1971.

The National Symphony Orchestra has a chamber orchestra and instrumental ensembles attached and accompanies opera, operetta, and ballet. It makes annual tours of the island, as do its provincial equivalents. Festivals of Cuban music and song are held at intervals, encompassing works of every genre from every period. Performing musical groups range from the traditional *charangas* to popular orchestras.

Folklore. In 1959 the Institute of Ethnology and Folklore was created within the Academy of Sciences of Cuba, with the aim of collecting and classifying the Cuban cultural heritage. It formed the National Folklore Group (Conjunto Nacional Folklórico), which performs Afro-Cuban dances nationally and internationally. The activities of the folklore group are complemented by the Institute of Literature and Linguistics of the Academy of Sciences. The legacy of Fernando Ortiz, a pioneer investigator of black customs, and the work of the Central University of Las Villas on popular culture deserve special mention.

National Folklore Group

Art. Cuba has some two dozen galleries, three art museums, and community cultural centres ready to display the works of Cuban painters. The Palacio de Bellas Artes and the Casa de las Américas both organize major

exhibitions from time to time. The revolutionary government emphasizes that it is not enough to open the galleries for increased public attendance if there has not been some prior education of the viewers, and effort is directed toward increasing public interest and artistic awareness.

Literature and publishing. Publishing is centralized in the Book Institute, an arm of the government, which produced 33,900,000 volumes in 1974—34 times the 1958 figure. About 70 percent of the titles are reference works, usually with scientific or technical content, and many of them are distributed free in support of national educational goals. The government-controlled press includes *Granma* ("Grandmother"), organ of the Communist Party, named for the yacht that brought Fidel Castro to Cuba in 1956; and *Juventud Rebelde* ("Rebel Youth"), the organ of the Unión de Jóvenes Comunistas (Organization of Communist Youth), as well as a wide range of provincial newspapers, magazines, and specialized reviews. The National Union of Cuban Writers and Artists (Unión Nacional de Escritores y Artistas Cubanos, or UNEAC), founded by Nicolás Guillén (a "social" poet whom the government considers the national poet), sponsors writers' competitions.

Granma and *Juventud Rebelde*

Although Cuba's pre-revolutionary literary production had, in many cases, achieved a very high quality, activity was sparse and unintegrated. New, often younger, writers now make their living principally (and in some cases solely) by writing. The annual competition of the Casa de las Américas, one of the most important in the Spanish language, brings world figures to Cuba.

Cuban poets maintain, perhaps surpass, pre-revolutionary standards, but many critics feel that fiction writing began to decline in the early 1970s, when works tended to become more directly related to revolutionary topics. The winner of the Casa de las Américas prize in 1971, *La última mujer y el próximo combate* ("The Last Woman and the Next Battle"), by Manuel Cofiño López, has a typical theme: the efforts (successful) of a new state farm director to eliminate corruption in the enterprise and to draw the peasants into the revolution.

Film making, radio, and television. Cuban film making has been stimulated by the Cuban Institute of Cinematographic Art and Industry (Instituto Cubano del Arte e Industria Cinematográfica; ICAIC), which aims to make serious films and to see that those films are available to the general populace. The government program of distribution has 560 mobile film units. The ICAIC also has an extensive film library and supports a cinematographic studies centre, which trains future technicians for the industry. The very active Cuban Broadcasting Institute (Instituto Cubano de Radiodifusión; ICR) has five national stations and more than 30 provincial stations. Two national television channels are supplemented by a local channel in Santiago de Cuba. Both institutes are agencies of the government.

BIBLIOGRAPHY. RAMIRO GUERRA Y SANCHEZ, *Manual de historia de Cuba*, new ed. (1971), offers an analysis of the economic conditions, social and political institutions, and external influences from the discovery of Cuba until 1868; JULIO LE RIVEREND, *La Habana: biografía de una provincia* (1960), deals with the historical development of the province in which Cuba's capital is located; the same author's *Historia económica de Cuba*, new ed. (1975; English ed. 1967), is a very important source for the study of the economic development of Cuba to 1958; J.K. BLACK *et al.*, *Area Handbook for Cuba*, 2nd ed. (1976), presents basic facts about social, economic, political, and military institutions; *A Study on Cuba* by the CUBAN ECONOMIC RESEARCH PROJECT, UNIVERSITY OF MIAMI, contains serious papers contributed by knowledgeable scholars; *Cuba 1968* (1970), by the LATIN AMERICAN CENTER, UNIVERSITY OF CALIFORNIA, LOS ANGELES, is a compilation made possible by exchanges of statistical data with Cuban agencies; *La enciclopedia de Cuba*, 9 vol., 2nd ed. (1975), was edited in Puerto Rico; *Anuario estadístico de Cuba*, issued by the CENTRAL PLANNING BOARD, is an annual compilation of statistical data on the economy and society.

(I.E./R.E.Cr./Ed.)

Cuba, History of

At the time of the Spanish discovery of Cuba, the native population formed two groups totalling 50,000. The Ciboney and Guanahatabey occupied western Cuba. The more numerous Taino, who occupied the rest of the island, were highly developed agriculturalists and a peaceful people, related to the Arawakan peoples of South America who had migrated to the Greater Antilles. Their houses, called *bohíos*, formed villages ranging from single families to communities of 3,000 persons. They had pottery, polished stone implements, and religious spirits called *zemis*, which were represented by idols of wood, stones, and bones. The Taino diet included potatoes, manioc, fruits, and fish.

Pre-Columbian Cuba

The Spanish colonial regime, to 1898. Christopher Columbus discovered Cuba for Spain during his first voyage, on October 27, 1492. Diego Velázquez began permanent settlement in 1511, founding Baracoa on the northeastern coast with 300 Spaniards and their African slaves.

Establishment of the colony. Within five years the island had been divided into seven municipal divisions, including Havana, Puerto Príncipe, Santiago de Cuba, and Sancti Spíritus. Each municipality had its own cabildo, or town council, governing its legal, administrative, and commercial affairs. From 1515, elected representatives of each cabildo formed a body that defended local interests before the royal council, especially matters such as the slave trade and the encomienda (an estate of land and the Indians inhabiting it). A bishopric, subordinate to Santo Domingo, was founded at Baracoa in 1518 but later moved to Santiago de Cuba.

Colonial life before 1763. The island's limited gold deposits discouraged early settlement. The colony became a staging ground for the mainland exploration of Yucatán, Florida, and the Gulf Coast. Such expeditions as that of Cortés, which attracted 400 Spaniards and 3,000 Indians, depleted the colonial population. The small number of permanent resident Spanish colonists used the Indians in encomienda. But by 1550 the encomienda was no longer feasible, because the island's Indian population had declined dramatically to about 5,000 because of social dislocation, maltreatment, epidemic diseases, and emigration. Throughout the 17th century, colonial life was made more difficult by the ravages of hurricanes, epidemics, the attacks of rival European countries trying to establish bases in the Caribbean, and freebooters. By 1700 peace had returned, and the population had grown to about 50,000. The *flota* system (regularly scheduled fleets between Spain and Spanish America) increased the commercial and strategic importance of Havana, while ranching, smuggling, and tobacco farming occupied the colonists. The administrative costs depended, however, on irregular subsidies from New Spain until 1808.

The plantation society, 1763–1898. The 18th century brought intensified agricultural development. The main changes came with the growing dependence on sugarcane cultivation and the importation of African slaves. In 1740 the Havana Company was formed to stimulate agricultural development by increasing the importation of slaves and regulating the export trade. The company was unsuccessful, selling fewer slaves in 21 years than the British sold during a ten-month occupation of Havana in 1762. Reforms of Charles III of Spain (ruled 1759–88) further stimulated the development of the sugar industry.

Dependence of the economy on sugarcane

Between 1763 and 1860 the island's population increased from less than 150,000 to more than 1,300,000. Slaves made the most dramatic growth, increasing from 39,000 in the 1770s to some 400,000 in the 1840s. In the 19th century Cuba imported more than 600,000 Africans, most of them after an Anglo-Spanish agreement to terminate the slave trade in 1820. The Cuban insistence on the slave trade raised considerable diplomatic controversy between Spain and Great Britain between 1817 and 1865.

In 1838–80 the Cuban sugar industry became the most mechanized in the world, utilizing steam-powered mills and narrow-gauge railroads. Expanding *ingenios* (sugar mills) dominated the landscape from Havana to Puerto Príncipe, expelling small farmers and destroying the is-

land's famous large hardwood forests. By 1850 the sugar industry accounted for 83 percent of all exports. In 1860 Cuba produced some 500,000 tons of sugar—nearly one-third of the world production. The sugar revolution propelled a new rich class of slaveowners to political prominence. Mexican Indians and Chinese augmented the labour force, and in 1865 the African slave trade ended; but slavery remained a thorny issue until its abolition in 1886.

The end of Spain's empire. The demands of sugar—labourers, capital, machines, technical skills, and markets—strained the interracial relations of the society, aggravated political and economic differences between metropolis and colony, and laid the foundation for the break with Spain in 1898. Spanish colonial administration had been corrupt, inefficient, and inflexible. The United States had shown a lively and growing interest in the island, and U.S. filibuster expeditions won support in the United States, especially among the Southern slave states; but several expeditions from the United States failed to free Cuba. After the 1860s the United States tried many times to purchase the island.

United States involvement

Spain's failure to grant political autonomy, while increasing taxes, led to the outbreak of the first war of independence—the Ten Years' War (1868–78)—which emphasized Spain's weakness even though the divisions within the island led to a military stalemate. The rich sugar producers of western Cuba and the vast majority of the slaves failed to rally to the Nationalist cause. The Nationalists themselves were divided over the questions of slavery, complete independence, and annexation to the United States. At that time the United States was already accepting 83 percent of all Cuban exports, compared with the 6 percent that went to Spain. Unable to find a military solution, Spain promised to reform the island's political and economic system at the Convention of Zanjón (1878). Many Cubans, including the Nationalist leader Antonio Maceo, however, refused to accept the Spanish conditions and left the island to prepare for a new attempt at independence.

By 1895 the political and economic crisis had grown more severe. America's investment had reached $50,000,000, and its annual trade with Cuba was about $100,000,000. Cuban political organizations in exile were coordinated and mobilized by the poet and propagandist José Martí. War broke out again on February 24, 1895.

Fighting quickly spread throughout the island. Spain deployed more than 200,000 troops. Both sides killed civilians and burned estates and towns. By 1898 commercial activity had fallen to a standstill. Sugar production declined from more than 1,500,000 tons in 1894 to 200,000 tons in 1897. The United States' economic interest intensified its concern and attracted intervention. Excited by the "yellow press" and a mysterious explosion aboard the USS "Maine" in Havana Harbour, the United States declared war against Spain on April 25, 1898. In August Spain signed a peace protocol in Washington, ending hostilities. The Cubans gained their independence in the Treaty of Paris (December 10, 1898), signifying the end of the Spanish American empire.

Cuban independence

United States occupation, 1899–1901. Cuba began its independence on January 1, 1899, under U.S. occupation. The first United States military governor was Gen. John Brooke, who had accepted the Spanish surrender. American attempts to exclude the Cubans from the government of the island met with friction. Brooke disbanded the Cuban Army and took a census before yielding his post to Gen. Leonard Wood, a former military governor of Santiago City. Wood, a competent and politically astute administrator, sought to mollify the rift between the two sides and supervised elections that gave Cuba its first elected president, Tomás Estrada Palma.

The military occupation restored normality. The Americans built a number of schools, roads, and bridges; they deepened Havana Harbour, paved streets, repaired and extended the telephone and telegraph system, started sewer works, and made a significant advance against yellow fever. But they were primarily interested in preparing the island for incorporation into the United States. American economic, cultural, and educational systems prevailed, and even the franchise was designed to eliminate Afro-Cubans from political participation. The Platt Amendment (1901), forced on the Cubans, gave the United States the right to oversee international commitments, dominate the economy, intervene in internal affairs, and establish a naval station (Guantanamo Bay) on the island.

The Republic of Cuba, from 1902. The first Republican administration, which began on May 20, 1902, under Tomás Estrada Palma, faced many difficulties involving relations with and the continued indirect influence of the United States. Estrada Palma, who favoured annexation by the United States, made an unabashed attempt to return to power at the end of his term in 1905. A Liberal revolt contesting the electoral and administrative procedures brought about the second United States occupation, on September 29, 1906.

Second U.S. occupation

The U.S. secretary of war, William Howard Taft, failed to resolve the dispute between the rebellious Liberals, who had lost the 1905 elections, and Estrada Palma, who subsequently resigned. Charles Magoon, who took over the administration for the United States in 1906, judiciously continued the Cuban constitution and kept the Cuban flag flying over public buildings. His civilian government was assisted by an advisory law commission, which revised the electoral procedure. On January 28, 1909, Magoon handed over the government to the Liberal president José Miguel Gómez. Meanwhile, political instability did not affect Cuba's steady economic progress, as the price of sugar continually rose until the 1920s.

Presidencies and dictatorships, 1909–58. Cuban politics before 1959 consisted of graft, corruption, maladministration, fiscal irresponsibility, and social insensitivity, especially toward Afro-Cubans. During the Gómez administration (1909–13) the Afro-Cubans, led by Evaristo Estenoz and Pedro Ivonet, organized to secure better jobs and more political patronage and to protest a law aimed at eliminating political associations based on colour and race. In 1912 government troops put down large demonstrations in Oriente, and 3,000 Afro-Cubans lost their lives. Gómez set the pattern of corruption, which was followed by Mario García Menocal (served 1913–21), Alfredo Zayas (1921–25), Gerardo Machado (1925–33), Fulgencio Batista (1940–44 and 1952–59), Ramón Grau San Martín (1944–48), and Carlos Prío Socarrás (1948–52). Machado and Batista were the most notorious, using assassinations, constitutional manipulation, and military power to gain office and remain there. Batista overthrew Machado with United States support in 1933 and was the most powerful politician in Cuba until 1959.

The income from sugar was augmented by a vigorous tourist business based on hotels, casinos, and brothels; Havana became especially attractive during the years of U.S. prohibition (1919–33) as a source of liquors and pleasure. Yet the prosperity of the 1920s, '40s, and '50s enriched only a few Cubans—mainly politicians and their families. For the majority of Cubans, poverty (especially in the countryside) and the lack of public services were appalling: with a national per capita income of $353 in 1958—among the highest in Latin America—unemployment and underemployment were still rife, even in the sugar communities, and the average rural worker earned merely $91 per year. Foreign interests controlled the economy, owning about 75 percent of the arable land, 90 percent of the essential services, and 40 percent of the sugar production. Nevertheless, there was no widespread economic discontent in 1959, when the revolutionary leader Fidel Castro ousted Batista.

The Castro regime from January 1, 1959. There was, however, widespread dissatisfaction with the Batista regime in 1959, even within the army. Castro had gained considerable admiration because of his early opposition to the Batista government. He had led a suicidal attack against the military fortress of Moncada in Santiago on July 26, 1953—hence the name of his movement, the 26th of July. Granted an amnesty along with the few

survivors, Castro went to Mexico and prepared for a renewed challenge to the dictatorship. Castro's group, including the Argentine revolutionist "Che" Guevara, landed in Cuba in December 1956 and was almost immediately wiped out. The survivors retreated to the Sierra Maestra and fought a two-year guerrilla campaign, gaining victory with fewer than 1,000 Fidelistas in January 1959. Their political plans were vague, their support rather insignificant, and their governmental skills totally untested; yet, they set about building a movement based on the poor peasants, the urban workers, the young, the nonwhite, and the idealistic. Under Castro's personal direction, the government gradually adopted a pragmatic Socialism designed to bring equality and social justice via agrarian, educational, and social reforms. Under Castro there were no elections, and what political organization took place was modelled after the Communist nations, based on the Cuban Communist Party, which dated back to 1925. Although Osvaldo Dorticós Torrado became president in July 1959, Castro firmly controlled the government as premier.

Overthrow of the Batista regime

The course of the revolution was neither smooth nor clear. The first three years were the most confused and difficult. The nationalization of hundreds of millions of dollars of U.S.-owned property brought the undisguised hostility of the American government, which reduced the sugar quota in 1960, dealing the Cubans a severe economic blow, declared a total trade embargo the following year, broke diplomatic relations (January 1961), supported an unsuccessful invasion by Cuban exiles (April 1961) at the Playa Giron, or the Bay of Pigs, in south central Cuba, and voted for the expulsion of Cuba from the Organization of American States (January 1962). Cuba's erratic drift into Communism and its economic dependence on the Soviet Union divided not only the country at large but also the 26th of July Movement. A great number of Cubans, especially among the wealthy and the skilled, emigrated mainly to the United States, reducing the efficiency of the administration. An attempt to diversify the economic base of the island in Castro's first three years under the National Agrarian Reform Institute (INRA) disrupted sugar output without producing the desired results. Constant mobilization for war and the ad hoc nature of the government created havoc among organizations and frustrated long-term planning. Attempts to export revolution, especially to the Dominican Republic, Venezuela, and Bolivia, alienated Cuba from most of the other Latin American countries. Only Mexico maintained full diplomatic and commercial relations with Cuba throughout the early revolutionary period.

Relations with the Soviet Union

Castro achieved full diplomatic relations with the Soviet Union in 1960, and that country replaced the United States as the major trading partner and source of funds and military supplies. Beginning in July 1960, the Soviets provided the major market for Cuban sugar. That same year they extended a loan to Castro of $100,000,000 at an annual interest rate of 2.5 percent. In order to defend the Castro government, the Soviets continued to extend loans, petroleum, and war material. The installation of Soviet missiles in Cuba in the summer of 1962 brought the world to the brink of war, as the United States forced the Soviets to remove the missiles under the threat of nuclear attack. The dominant Soviet role forced Cuba to support the Soviets in their dispute with China, although certain of Castro's followers were fascinated by Mao Tse-tung's revolutionary methods, and although the Chinese—along with many other Communist countries—were also important trade partners.

Soviet military support for the Cuban Revolution was crucial for defense against outside attack. But Soviet technology failed to provide an equal nonmilitary assistance. The Soviets were unable to supply desired consumer products, and an attempt to construct cane-cutting machines was a dismal failure. As early as 1962, the Cubans began to ration supplies, and rationing became a way of life. In 1969 the Cubans even rationed cigars and sugar, two products for which the island had been known for its surplus.

Rationing, however, was only partly the result of political and economic measures. Bad weather, especially floods and droughts, and the unavailability of manpower, especially to cut the canes, also reduced harvest expectations. Despite the efforts of INRA to produce 10,000,000 tons of sugar in 1970 by greatly extending the *zafra* (season of the sugar harvest), the manpower reserves proved inadequate for the task. Volunteer and forced labour were insufficient to offset technological deficiencies.

As external threats to the island increased and the internal economic situation deteriorated, the Castro government moved toward greater organization. In 1965 Castro merged his followers with the Communist Party of Cuba to form a single monolithic political structure. In the early years, Castro had gained popular support from the powerful feeling of involvement in the revolution and from the anticipated invasions by counterrevolutionaries, but, by the 1970s, with Castro maintaining his personal rule, the sacrifices required demanded more than a sense of involvement or self-interest. No major headway was made against chronic failures and shortages. Instead, in March 1968, the government expanded its nationalization and instituted stronger measures to encourage workers to leave the cities and do agricultural labour. Despite its vagaries, sugar remained Cuba's most valuable product on the world market, and so the obsession to produce greater quantities continued.

The Cuban Revolution was still incomplete in the 1970s. Yet, despite its admitted failures, it had injected national pride, social consciousness, and honesty for the first time in Cuban history, and, while it had not solved all the problems, the situation no longer looked bleak.

BIBLIOGRAPHY. R. GUERRA Y SANCHEZ *et al.* (eds.), *Historia de la Nación Cubana*, 10 vol. (1952), is a lengthy, solid, but rather ponderous study that provides the basic background from the colonial period to modern times. F.W. KNIGHT, *Slave Society in Cuba During the Nineteenth Century* (1970), examines the social, political, economic, and demographic impact of the rise of the sugar industry from the middle of the 18th century until the abolition of slavery. DAVID F. HEALY, *The United States in Cuba, 1898–1902* (1963), meticulously analyzes the actions and personalities involved in the first military occupation. HUGH THOMAS, *Cuba: The Pursuit of Freedom* (1971), is a monumental narrative of the period from 1763 until 1970, with an authoritative text enhanced by apt illustrations and useful appendixes. LEE LOCKWOOD, *Castro's Cuba, Cuba's Fidel*, new ed. (1969), is a very human treatment of the Cuban leader, based on extensive tape-recording sessions, the transcripts of which had the Cuban leader's approval prior to publication.

(F.W.Kn.)

Cuculiformes

The bird order Cuculiformes is a cosmopolitan group containing two very distinct families, the cuckoos, Cuculidae, and the turacos, or plantain eaters, Musophagidae. There has been considerable discussion as to the degree of relationship of the two, and authorities increasingly tend to give the turacos ordinal rank. The family Cuculidae is much the larger group, containing about 127 species, found in the tropical and temperate zones of all the continents (none in Antarctica) and on many oceanic islands; the Musophagidae contains 20 species, found only in Africa.

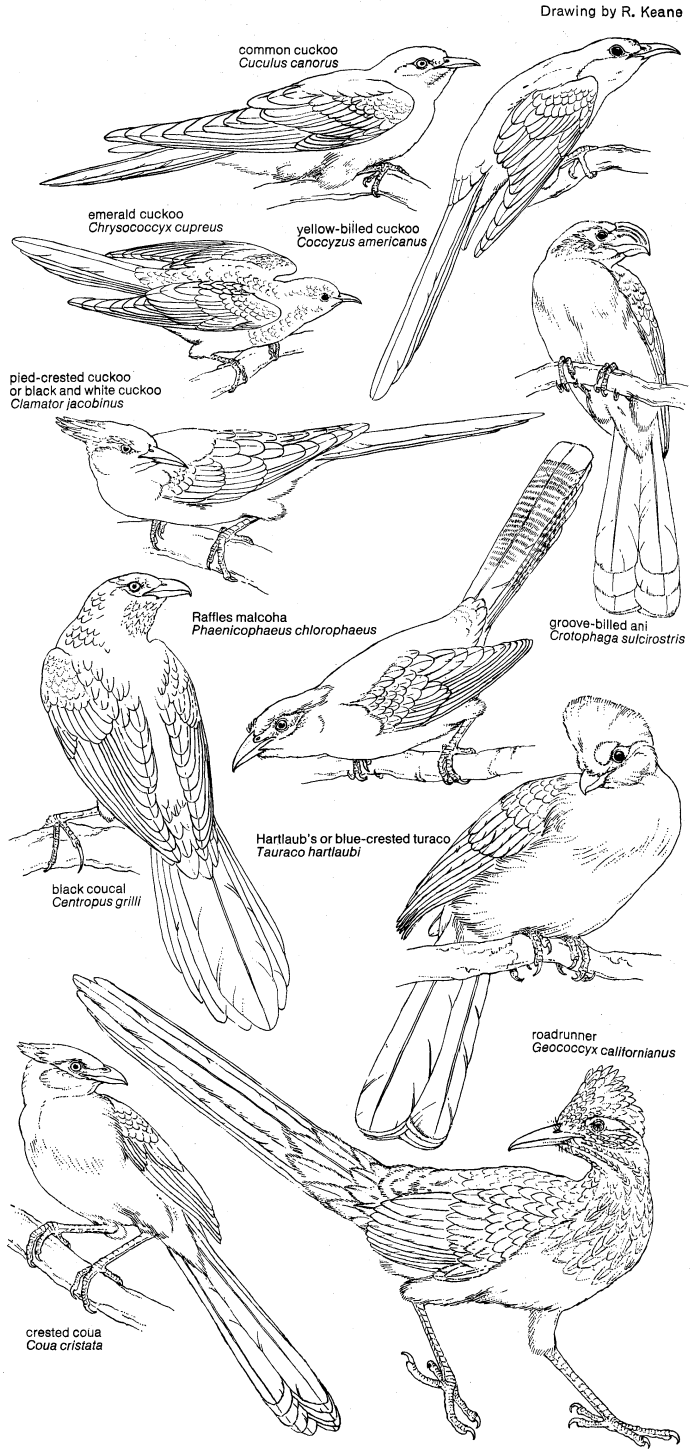
The cuckoos are an ancient group with uncertain phylogenetic affiliations and no living near relatives, even the turacos being quite distinct and with no intermediate, or connecting, species. The cuckoos are of unusual biological, especially ethological, interest because many species are brood parasites; *i.e.*, they lay their eggs in the nests of other species, which then rear the young cuckoos. Other cuckoos make their own nests, in which they incubate their eggs and rear their young as do most birds; and still others (*Crotophaga* and *Guira*) build communal nests. Some cuckoos are of interest because they are among the few birds that feed extensively on hairy caterpillars.

General features. The cuckoos cover a great range in size, from the small glossy or emerald cuckoos of the

genus *Chrysococcyx*, which are about 15 centimetres (six inches) long, to the large ground cuckoos (*Carpococcyx*) and the larger species of coucals (*Centropus*), which reach nearly 90 centimetres (three feet), including the tail, which is often strikingly long. Most cuckoos have fairly loose-webbed feathers, varying in colour from subdued browns, grays, olive, and black, to brilliant, iridescent greens and purples and bright yellow. The beak is of moderate length and often slightly downcurved. The turacos are all sizable birds, the smallest species having a total length of about 37 centimetres (15 inches) and the largest attaining a length of almost 75 centimetres (30 inches).

Size ranges

In keeping with the greater number of their included species and their worldwide distribution, the cuckoos show far more diversity in structure than do the relatively homogeneous turacos. The cuckoos include the arboreal "typical" cuckoos of both the Old World and the New World, the terrestrial roadrunners (*Geococcyx*) of southwestern United States and Mexico, and the more compact but also largely terrestrial coucals (*Centropus*) of Africa and Australasia.



Body plans of representative Cuculiformes.

Distribution. Although the family Cuculidae is virtually worldwide in temperate and tropical regions, most of the subfamilies are restricted to one hemisphere or the other. Three (Cuculinae, Centropinae, and Couinae) are limited in distribution to the Old World; the subfamily Crotophaginae is wholly New World; the Neomorphinae is mostly New World, with one genus in Southeast Asia; and the sixth subfamily, Phoenicophaeinae, is represented in the tropics of both hemispheres. Many cuckoo genera are peculiar to certain parts of the world; others are cosmopolitan. Three genera (*Rhamphomantis*, *Caliechthrus*, and *Microdynamis*), for example, are known only from New Guinea; *Dasylophus* and *Lepidogrammus* occur only in the Philippine Islands; and *Saurothera* (lizard cuckoos) is found only in the West Indies. Some other genera are wide ranging: the typical cuckoos of the genus *Cuculus* are found in Europe, Asia, Africa, and Australia; the small, brightly coloured *Chrysococcyx* species occur in Africa, Asia, Australia, and some Pacific islands; and the coucals (*Centropus*) live in Africa, Asia, Australia, and adjacent islands.

Most cuckoos are solitary, often furtive birds that are inconspicuous even when relatively common. They do not form large flocks or even, except for the communal nesting anis (*Crotophaga*), small parties. The turacos, some of which form loose bands, are often conspicuous because of their bright colours, large size, and loud voices.

Natural history. *Habitat utilization.* As a group, cuckoos are forest birds, often inhabiting dense thickets that may make them difficult to observe. Some species, such as many of the genus *Cuculus*, inhabit rather open woodland. The guira (*Guira guira*) of South America and many members of the Old World genera *Clamator* and *Chrysococcyx* are found in open savanna (grassland), but only where trees are present. The roadrunners live in open scrub and cactus desert, often in the absence of any large vegetation.

The green and iridescent turacos (*Tauraco*, *Gallirex*, *Musophaga*, and *Corythaeola*) are primarily residents of dense, evergreen, broad-leafed forest; the grayer forms (*Crinifer*), most of which are called go-away birds (because the calls of some are "g'-way, g'-way"), are found in more open woodland, including savanna.

Food habits. Cuckoos are largely insectivorous, preying mostly on crawling insects. Orthoptera (grasshoppers, locusts, and mantids) are often taken. The greater, or North American, roadrunner (*Geococcyx californianus*) is reported to take flying grasshoppers by leaping after them from the ground. The larger cuckoos, such as the lizard cuckoos, roadrunners, and coucals, take substantial numbers of lizards, snakes, and other small vertebrates, sometimes including birds. Alone among the Cuculidae, the Madagascar couas (*Coua*) are reported to eat some fruit.

Musophagids feed almost entirely on fruit; only a few species are known to take insects and small vertebrates.

Vocalizations. Compared with birds generally, cuckoos must be considered a highly vocal group, a fact consistent with their forest habitat. A variety of songs, contact calls, and alarm notes are known for most species, some melodious, many harsh and discordant. The "song," associated with territorial assertion and courtship, is usually characterized by the repetition of loud, short notes, often on a descending scale or with a downward break in the middle. The individual syllables have variously been described as whistling, piping, cooing, tooting, laughing, grating, and clicking, depending on the species. The familiar clear, two-note call of the common cuckoo (*Cuculus canorus*) of Europe, Asia, and Africa is uttered by the male alone, the female giving a low bubbling call; apparently in most other species of cuckoos as well, the song is given by the male alone.

"Songs" of cuckoos

In addition to vocal sounds, at least one cuckoo, the North American, or greater, roadrunner, makes nonvocal sounds, clacking the mandibles to produce a rattling noise.

The calls of turacos are described as barking or laughing notes given in a series. A few species utter softer cooing notes.

Courtship behaviour. Except for features of parasitism, the behaviour of cuckoos is one of the least studied aspects of their natural history. A few species, in several subfamilies, have been observed closely, and some generalizations are possible. In territorial advertisement and in courtship, the singing male (in both parasitic and nonparasitic cuckoos) may posture, spreading the tail feathers, which are white tipped in most species, and dropping the wings. Courtship feeding of the female by the male has been observed in several species.

Courtship behaviour has been described for a few musophagids. In forest species, the male walks or leaps among branches with his tail fanned and with his wings half spread, displaying the crimson wing patches. Courtship feeding has also been observed, the male regurgitating fruit pulp or presenting whole fruit to the female.

Nesting. Two basic types of nesting behaviour are found among members of the order: parasitic and nonparasitic. About three-fifths of the Cuculidae and all of the Musophagidae mate, nest, and rear their young in the manner seen in most birds. About 50 species of cuckoos, including all of the subfamily Cuculinae (sometimes called "typical" cuckoos) and three species of the Neomorphinae, exhibit brood parasitism (*i.e.*, they lay their eggs in the nests of other species, which then rear the young cuckoos). Brood parasitism in the common cuckoo was recognized by Aristotle in the 4th century BC, and that of the koel (*Eudynamys scolopacea*) appears in the ancient Vedic literature of India, dating from about 2300 to 1800 BC, but the scientific study of this form of parasitism remained largely neglected until the late 19th century, when intensive study by many observers revealed a number of remarkable behavioral and physiological adaptations in the cuckoo.

In western Europe the common cuckoo is migratory, arriving at its breeding area in spring. The female establishes a territory and defends it against intrusion by other female cuckoos. The male is less restricted to an individual territory and more tolerant of other males; often more than one male will be in attendance on one female. Moving about within her territory, the female learns of potential host nests largely by watching their construction. Males of some species of cuckoos may accompany the females in searching for potential host nests. By the time she is ready to lay her eggs the female cuckoo has under surveillance several possible host nests and knows when the first eggs are laid by their owners. Most of the species parasitized by the common cuckoo are smaller than the cuckoo and lay their eggs at the rate of one each day until the clutch is complete, only then beginning incubation. Visiting when the hosts are away from the nest the cuckoo is able to place her egg in the nest after the host has started the clutch but before incubation has begun. It is well established that the common cuckoo lays eggs at 48-hour intervals, and apparently some glossy cuckoos lay one every 24 hours. The cuckoo places only one egg in each host nest and so must find a new nest for each of her 15 to 20 eggs. When parasitizing a species that builds a domed or globular nest, the female cuckoo clings to the outside of the nest and lays her egg through the opening without entering the nest. The repeated finding of cuckoo eggs in nests too small to admit a cuckoo or to bear her weight has led some authorities to maintain that the cuckoo lays the egg on the ground and carries it to the nest in her beak, but extensive observations have failed to substantiate this supposition. Some of the glossy cuckoos, however, parasitize nests so small and frail that mandibular egg placement cannot be ruled out.

Egg laying by the common cuckoo

In western Europe common cuckoos lay their eggs in the nests of many species, mostly songbirds (Passeriformes); more than 140 hosts species have been reported, including some that could hardly rear a young cuckoo successfully. Careful study of this and several other cuckoos indicates, however, that despite the wide range of hosts recorded, a few host species bear most of the load of raising the young of each cuckoo species. In some instances, such as two of the little Australasian glossy cuckoos (*Chrysococcyx*), the parasite relies mainly on a single host species.

Egg mimicry

Many birds react to a foreign object in the nest by deserting the nest, building another nest on top of the first, or removing the offending object. It is therefore of great advantage to a brood parasite to lay eggs that resemble those of its host, or, conversely, to parasitize species whose eggs are the same colour as its own. The utilization of several host species by a single population of cuckoos presents a difficult adaptive problem, for the host species may have many different egg colours. Early naturalists noted that there was often a marked resemblance between the egg of a cuckoo and those of the host, and a German ornithologist, Eduard Baldamus, in 1892 showed that the frequency and degree of similarity were too great to be coincidental. Subsequent studies by a number of workers, especially by an English naturalist, Edgar P. Chance, have revealed much of the basis for the resemblance, which is now called egg mimicry.

Each female common cuckoo lays eggs of a particular shape and coloration throughout her life, but the eggs of different females vary widely. Most authorities now agree that a given female normally restricts her parasitism to the single host species by which she herself was reared, and that the common cuckoo population is composed of numerous clans, called gentes, each of which parasitizes only one species of bird, for which the females of that gente have evolved egg mimicry. Only the gente of the female is important in egg coloration; the foster parentage of the male (*i.e.*, his gente) has no effect on his choice of females, and his genetic makeup apparently does not influence the coloration of eggs laid by his female offspring. Gentes are not given taxonomic designations but are sometimes known by their host species ("redstart cuckoo," "dunnock cuckoo," etc.). Evidence for this theory is largely circumstantial because of the difficulty in establishing the identity of individual cuckoos, but no plausible alternative theory has been proposed. The presence of cuckoo eggs in a wide variety of host nests in some small areas has been taken as evidence refuting the single-host hypothesis, but most of the nests with nonmatching eggs are found in areas in which the ecology has been seriously altered by human activity, with the result that female cuckoos have suddenly found themselves without adequate numbers of the proper host species and have been forced to utilize hosts for which their own egg colours are inappropriate. In the forested portions of northern Europe, where the habitat is relatively undisturbed and consists of homogeneous microecological areas (*e.g.*, reedbeds, birch forests), the common cuckoo parasitizes mainly a small thrush, the redstart (*Phoenicurus phoenicurus*), and both lay unspotted bluish eggs; in the reedbeds around Hungarian lakes the usual host is the great reed warbler (*Acrocephalus arundinaceus*), and the cuckoo eggs are strongly blotched with gray and black, like those of the warbler.

Studies of the reaction of the host to the presence of a cuckoo egg have indicated that egg mimicry is of considerable importance to the success of the parasite. Only about 5 percent of well-matched eggs are rejected by their hosts, as compared with up to 72 percent of mismatched eggs. Few cuckoos have been studied intensively in terms of egg mimicry, but the phenomenon is known to occur in at least some species. The great spotted cuckoo has an egg pattern mimicking that of the magpie (*Pica pica*), its usual host in southern Europe. In Africa, where it is apparently a recent colonist, this cuckoo exhibits what has been termed "evolutionary escape from specialization"; its new hosts, certain starlings, nest in holes, in which the semidarkness renders the dissimilarity of eggs less visible and hence unimportant. Among the 12 species of emerald cuckoos, the degree of host specificity and egg resemblance is variable, some species being highly host adapted, others less so.

The genus as a whole shows no clear continuity of evolution in egg adaptation; individual species evidently have evolved their egg mimicry separately.

A notable feature of the egg of the common cuckoo is its small size; its weight is only about 2.5 percent of that of the adult cuckoo. The eggs of other parasitic cuckoos weigh more than 6 percent of the adult weight, and those of birds that rear their own young vary from 8 to more than 20 percent of adult weight. The adaptive value of small eggs is not well understood, but it is assumed to be related to the reliance of the common cuckoo on small hosts. Recent experiments have shown, however, that incubating birds show no adverse reaction to eggs significantly larger than their own.

Nest-mate eviction

Once the egg has been accepted and incubated by the hosts, the newly hatched cuckoo faces the problem of securing enough to eat from its small foster parents. The young cuckoo has evolved an astonishing form of behaviour, that of nest-mate eviction, that ensures that it will not have to compete with members of the foster brood for food. Within a few hours of hatching, the blind, naked, young cuckoo develops a strong urge to evict any objects, such as eggs or other nestlings, from the nest. It does this by working itself under the offending object and, aided by the presence of a depression between the shoulderblades, heaving the object over the rim of the nest. Within about 24 hours of hatching, the young cuckoo has the nest and the attentions of its foster parents to itself. The eviction habit is well developed in some parasitic cuckoos and seems to operate effectively in favour of the evictor but is surprisingly absent in many others, such as the crested cuckoos (*Clamator*).

Another habit that, like nest-mate eviction, tends to reduce the competition faced by the young cuckoo is the removal of one or more of the host's eggs by the female cuckoo, sometime before or after or at the time of the laying of her own egg.

Soon after hatching, the young cuckoo, while still in the nest, grows rapidly in response to the efforts of its foster parents, who regard it as their own. By the time it is ready to leave the nest, the young bird dwarfs its smaller hosts. Normally it does not receive attention from its real parents, but adult common cuckoos have occasionally been observed feeding young cuckoos; in the genus *Chrysococcyx* such behaviour appears to be more frequent than in other genera.

Social nesting in the guira and anis

Most of the nonparasitic cuckoos form stable pair bonds and defend territories, within which they build their nests and rear their own young. The guira and the anis are exceptional in that they live in flocks of five to 20 individuals, each flock defending a territory within which its members feed and nest. Several birds of the flock may cooperate in building a nest, in which two or more females may lay eggs and share incubation. Many members of the flock participate in feeding the young.

The nests of nonparasitic cuckoos are loose platforms of twigs placed in low vegetation or, rarely, on the ground. The coucals are unusual in that they build sizable domed nests of grass and twigs, on or near the ground, with side entrances.

The musophagids also build large twig nests, placed in trees often at considerable height. They lay unmarked whitish eggs, from which the young are believed to hatch completely covered with down; certainly they are downy from an early age.

Migration. Like most insectivorous birds in temperate climates, temperate-zone cuckoos migrate toward or across the equator for the winter. The migrations of some species are remarkable in that the young of the year travel completely independently of their parents and may cross up to 3,200 kilometres (2,000 miles) of open ocean unguided. Adults and young of some of the small Australasian glossy cuckoos separately migrate long distances to relatively small oceanic islands. It is obvious that the means of orientation and navigation must be unusually precise, but what they are remains as yet unknown.

Form and function. In general body plan, most cuckoos resemble the "perching birds" (order Passeriformes). They are slender bodied, with medium to long wings and medium to extremely long tails, the latter sometimes constituting more than half the total length of the bird. The tail is usually graduated (the outer feathers shorter), and the individual feathers often are white tipped.

The foot structure of cuckoos and musophagids is distinctive and at close range distinguishes these birds from any with which they might be confused. The foot of cuckoos is zygodactylous, or yoke-toed (*i.e.*, the outer toe is directed backward); that of musophagids is semizygodactylous, the outer toe being movable into either a forward or a backward position.

The cuckoo bill, although usually slender, is occasionally quite stout, as in the ground cuckoos and coucals, in which it makes a formidable weapon for subduing prey. In the anis the bill is heavy, with a strong ridge along the culmen (top), and, in the channel-billed cuckoo (*Scythrops novaehollandiae*) of Australasia, it is extremely large, almost like that of toucans. The bill is strong and rather short in the Musophagidae, usually with serrated edges. In the genus *Musophaga* it extends onto the forehead as a broad shield.

Most musophagids have prominent crests, a feature rare in cuckoos and well developed only in the genus *Clamator*. A number of cuckoos and turacos have bare skin, usually brightly coloured, around the eye or between the eye and the bill.

Turaco feather pigments

Perhaps the most distinctive feature of the turacos is the presence of two unique feather pigments. Turacin, a rich crimson pigment found in the primary flight feathers of *Tauraco* and *Musophaga*, is a copper salt of the pigment uroporphyrin III; and turacoverdin, the green pigment in turacos, is an oxidized derivative of turacin.

Paleontology and classification. The order Cuculiformes is not well represented in the fossil record, but both cuckoos and musophagids apparently existed by the Early Tertiary Period, some 60,000,000 years ago. The dearth of fossil evidence is not surprising, in view of the group's preference for forest habitats; woodland animals, particularly small ones and particularly birds, are less frequently fossilized than those in open areas or aquatic habitats. Two species of musophagids have been described from the upper Eocene or lower Oligocene of France (about 40,000,000 years ago) and one from the middle Miocene (about 15,000,000 years ago). Three species of fossil cuckoos are known from North America, the earliest from the middle Eocene (about 45,000,000 years ago), and a fossil coua has been found in deposits from the Quaternary Period (the last 2,500,000 years) of Madagascar.

Although the cuckoos and musophagids are frequently united in a single order, as in the present article, most authorities acknowledge the distinctness of the two groups by placing them in separate suborders, Musophagi and Cuculi. The parrots (Psittaciformes) are often considered on osteological evidence to be the closest relatives of the cuculiforms, but even this relationship, if it exists at all, is not a close one.

BIBLIOGRAPHY. E. BALDAMUS, *Das Leben der europäischen Kuckucke* (1892); and W. VON CAPEK, "Beiträge zur Fortpflanzungsgeschichte des Kuckucks," *Orn. Jb.*, 7:42–72, 102–117, 146–157, 165–183 (1896), are two classic works, in German, about the common cuckoo. E.P. CHANCE, *The Cuckoo's Secret* (1922), is an important and highly readable book about the common cuckoo. D. E. Davis, "Social Nesting Habits of the Smooth-Billed Ani," *Auk*, 57:179–218 (1940), presents the results of extensive field work on anis. H. FRIEDMANN, *The Parasitic Cuckoos of Africa* (1948); "Evolutionary Trends in the Avian Genus *Clamator*," *Smithson. Misc. Collns.*, 146:1–127 (1964), and "The Evolutionary History of the Avian Genus *Chrysococcyx*," *Bull. U.S. Natn. Mus. 265* (1968), are technical papers on Old World tropical cuckoos. A.C. BENT, "Life Histories of North American Cuckoos, Goatsuckers, Hummingbirds and Their Allies," *ibid. 176* (1940), provides readable accounts of the natural history of North American cuckoos and of some Asiatic *Cuculus* species. R.E. MOREAU, "A Contribution to the Biology of the Musophagiformes, the So-Called Plantain-Eaters," *Ibis*, 2: 639–671 (1938), is one of the few detailed studies of musophagids.

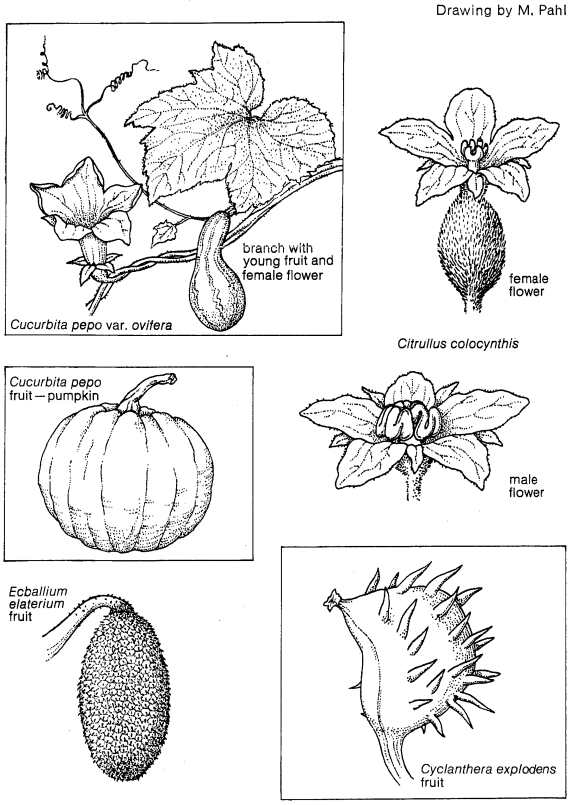
(He.F.)

Cucurbitales

Modern classification schemes of the flowering plants by most authorities suggest that the order Cucurbitales (the gourd order) is more or less isolated and composed of a single family, the Cucurbitaceae. The affiliations of the Cucurbitaceae are not well understood, and more research is necessary before its relationship within the system of flowering plant orders can be firmly established. Most authorities, however, have suggested a relationship with the Passifloraceae (Passiflorales order).

Plants of the Cucurbitales order for the most part are a homogeneous group, easily distinguished from other groups of flowering plants by amateur and professional alike. The plants are usually rapid growing, sprawling vines or climbers, annual or perennial, with palmate alternate leaves, long spiral tendrils, an inferior (or enclosed) ovary, and normally with unisexual flowers that have yellow or white petals. Seeds are without endosperm (a special nutritive tissue for the developing embryo) but have large cotyledons ("seed leaves"). The fruits are readily recognized; typical examples are the cucumber, muskmelon, watermelon, pumpkin, squash, and gourd. Because numerous cultivated varieties of most members of the gourd order exist, and because the group is popular and well known, many species have received several common names. It is likewise true that one common name has often been applied to several different species. Thus, the Latin or "scientific name" of a plant in this order provides the best guide to its exact identity.

Drawing by M. Pahl



Some representative plant structures in the order Cucurbitales.

GENERAL FEATURES

Size range and diversity of structure. There is a great range in the size, shape, and colour in the fruits of the Cucurbitales order. Some of the species (*e.g.*, pumpkin and winter squash of the species *Cucurbita maxima*, and the white-flowered gourd, *Lagenaria siceraria*), have fruits among the largest in the plant kingdom. *Cucurbita maxima*, variety Mammoth, has jumbo-sized fruit that may weigh 75 pounds (34 kilograms) or more, while the huge oval fruits of certain cultivars (varieties originating and persisting in agriculture) of *Lagenaria siceraria* may weigh as much as 50 to 125 pounds (23 to 57 kilograms). The cylindrical fruits of some cultivars of *Lagenaria siceraria* attain a length of 2 to 3 feet (6 to 9 metres), with a diameter of 2 to 4 inches (5 to 10 centimetres). The "snake melon," a cultivar of *Cucumis melo*, has about the same dimensions but tends to be slightly thicker in the centre.

Range of fruit colours

Fruit colours of cucurbits (members of the Cucurbitaceae family; hence, the Cucurbitales order) range from almost pure white to yellow, tan, brown, orange, dark orange, light green, and dark green. In the bicolour gourds, a single fruit has two distinct areas, usually green and white, or occasionally green and orange. Many fruits are striped or mottled, as in the watermelon and some gourds. Fruits of several species have prominent spines as in the chayote and certain species of *Cucumis* (cucumber, gherkin, etc.). In *Cucumis melo* var. *reticulatus* (cantaloupe) the fruit is covered with corky tissue, commonly known as the net.

The vines of most species of cucurbits are of modest size, with runners 10 to 15 feet (3 to 4 metres) long, but some species have thread-like, matted runners that are relatively short, as in *Brandegea* and *Echinocystis*. Others have giant runners 40–50 feet (12 to 15 metres) in length with a diameter of one to three inches (2.54 to 7 centimetres) at the base of the plant. Plants of such dimensions occur in some cultivars of *Lagenaria siceraria* (white-flowered gourd) and *Cucurbita mixta* (pumpkin, squash).

Distribution and abundance. There are 90 to 100 genera and about 700 species in the order. The species are about evenly distributed between the tropics and subtropics of both hemispheres. There is, however, an overlap into temperate regions. *Marah*, the chilicothe, occurs in North America as far north as southern Oregon, and in the Middle West of the United States, *Sicyos*, the bur cucumber, extends as far north as southern Canada. Some cultivars of *Cucurbita* (pumpkin, squash, gourd) thrive as far north as the Canadian border.

The Cucurbitales is essentially a tropical order, and for this reason it is assumed to be a relatively old one, but there are no fossil materials to substantiate this informed guess. While the cucurbits never become aggressive weeds in either the tropics or subtropics, certain genera such as *Cucurbita* (pumpkin, squash, gourd), *Cucumis* (cucumber, gherkin, muskmelons), and *Citrullus* (watermelon) retain a great deal of plasticity producing numerous variants in both vine and fruit as is demonstrated by their response to selection and hybridization under cultivation.

Economic importance. For a primarily tropical order, members of the Cucurbitales are of tremendous economic importance as food plants. Surprisingly, those of most importance for food grow well in temperate regions of the world. They are of minor importance for fibre and ornamentals, and one species is used for household purposes in primitive societies.

It is difficult to estimate the economic importance of cucurbits as food plants because they are used in such a manner that they do not enter into commerce; hence production statistics are not only unavailable, there are none. For instance, the huge quantities of cucumber, squash, pumpkin, watermelon, muskmelon and others, grown by almost every backyard gardener in North and South America go unreported, and there is no way to estimate the volume or value of such crops. Likewise, in several Latin American countries the perennial *Cucurbita ficifolia* (fig leaf gourd) is allowed by the peasants to grow around their dwellings, and the fruit, mostly immature, is hauled to the markets. This constitutes a sizeable trade that is never reported. Also, cucumbers, muskmelons, and watermelons are an important item in the diet, seasonally, for peoples in vast areas of Asia Minor, U.S.S.R., India, China and Japan. Obviously, these items are never reduced to trade statistics.

The use of cucurbits as food plants is not primarily for their caloric, mineral, or vitamin values, because they are poor or only modest sources of these nutrients. They are liked for their attractive appearance, pleasant taste, and refreshing qualities. As pickles they also add diver-

sification to the diet for people with an otherwise undistinguished menu.

There are some exceptions to the generally low nutrient content of the fruits of the Cucurbitales order. The winter or baking squash (some cultivars of *Cucurbita maxima*, *C. moschata*, *C. pepo*, etc.) are relatively high in energy and carbohydrates, and they are an excellent source of vitamin A. Besides, they contain modest amounts of minerals and other vitamins. The winter squash along with maize and beans were the foundation diet of the great pre-Columbian civilizations in the Americas. For peoples lacking modern refrigeration, cucurbits such as squash are an ideal food because they can be stored intact for long periods of time. They can also be cut into strips and dried in the sun. In this dehydrated form they can be stored almost indefinitely.

Edible and useful species

Species used for food include the following: *Cucumis sativus* (cucumber); *C. anguria* (gherkin); *C. melo* (muskmelon, including the cantaloupe, Persian, casaba, and honeydew melons); *Citrullus lanatus* (watermelon); *Cucurbita pepo* (summer squash, pumpkin, yellow-flowered gourd); *C. mixta* (pumpkin, squash); *C. moschata* (pumpkin, winter squash); *C. maxima* (pumpkin, winter squash); *C. ficifolia* (fig leaf gourd); *Lagenaria siceraria* (white-flowered gourd); *Benincasa hispida* (Chinese preserving melon); *Sechium edule* (chayote); *Sicana odorifera* (cassabanana); *Momordica balsamina* (balsam apple, Chinese bitter melon); *Trichosanthes cucumeroides* (snake gourd); *Cyclanthera pedata* (achocha); and *Hodgsonia heteroclita* (lard fruit). There are other species used for food but in a minor way and mostly by primitive tribes in Africa.

Species used for ornamental purposes include *Momordica charantia* (balsam pear), and *Cucurbita pepo* (yellow-flowered gourd).

Species used for fibre include *Luffa cylindrica* (vegetable sponge, dishcloth gourd), and *L. acutangula* (vegetable sponge, dishcloth gourd).

The main species used for containers, rattles, floats, and cutlery is *Lagenaria siceraria* (white-flowered gourd).

NATURAL HISTORY

Life cycle. The life cycle of species of the Cucurbitales order is in general similar to that of other dicotyledonous plants. There are, however, certain unique features that deserve mention. For example, polyploidy (*i.e.*, the presence in the plant cells of three or more sets of chromosomes) is relatively uncommon in the family, and chromosome number is identical for species within many genera as in *Cucurbita*, *Citrullus* and others. It is true, however, that only a small number of species have been sampled for chromosome number, and a more broadly based study is needed, particularly of tropical species, to confirm these trends.

Nearly all species in the family have unisexual flowers, and the pollen is heavy and sticky. For these reasons, pollination is entirely dependent upon insects. Except for the genus *Cucurbita*, the cultivated species are pollinated by the domestic honey bee. For cantaloupes and other muskmelons raised commercially one or two hives of bees per acre are used to obtain adequate pollination of the crop. In *Cucurbita*, (squashes, pumpkins, gourds) in which the pollination system has been thoroughly investigated, an intimate relationship exists between the cucurbit species and solitary bees of the genera *Peponapis* and *Xenoglossa*. Solitary bees are structurally and physiologically well adapted for pollination of the large *Cucurbita* flowers, much more so than domestic honey bees. As a group the solitary bees are entirely dependent upon *Cucurbita* for their pollen economy. Even more interesting, there are some examples of species to species relationship, suggesting reciprocal evolution of the bees and the genus *Cucurbita*.

Seed dissemination takes on bizarre forms in several genera of cucurbits. In *Ecballium* (the squirting cucumber), the mature fruit separates from the stem with explosive force, and the seeds are ejected from the fruit under pressure, travelling a distance of 10 to 20 feet (3 to 6 metres). In the genus *Momordica*, at maturity the fruit bursts, exposing the scarlet arils (structures attached to the seed, often enclosing it) that probably attract the attention of birds. In a similar fashion in the primitive forms of muskmelon (*Cucumis melo*), the fruit cracks longitudinally, exposing the seeds, which can be spread by birds or other animals.

There are several genera in the Cucurbitaceae that show unusual responses to environmental stimuli. The genera *Marah*, *Cucurbita*, and *Sechium* (chayote, mango squash, and other common names) respond to arid conditions by producing a large storage root. In *Marah* and the dry-habitat species of *Cucurbita*, the foliage, flowers, and fruits develop in a remarkably short period of time, while the huge, fleshy, storage root is perennial (*i.e.*, it lives through many growing seasons). The genera *Acanthosicyos* and *Dendrosicyos* respond to arid conditions by producing scalelike leaves and succulent stems. The only shrubby species of the cucurbits are in these genera. The female flower of *Cucumis humifructus* (the aardvark cucumber) responds to gravity, after being pollinated, by plunging into the soil, where the fruit is produced from 6 to 12 inches (15 to 30 centimetres) below the soil level. *Cucumis humifructus* is the only member of the Cucurbitales order having underground fruits. It has developed an interdependent relationship with a mammal, the ant bear or aardvark, *Orycteropus afer*, of southern Africa. The aardvark digs up and eats the fruit of *C. humifructus*, probably in part to quench its thirst. Some seeds pass through the alimentary canal of the animal intact, and are thus distributed with the dung in the vicinity of the aardvark lairs. The animal carefully prepares a hole for its dung and after defecation covers it with loose soil. This behaviour pattern of the animal results in the making of a perfect seed bed for the germination of the young plants of *C. humifructus*. Thus the association between the plant and animal contains the basic elements of a symbiotic relationship, at least to a limited degree. It is significant that the local name of the species is "aardvark cucumber."

Life history of the aardvark cucumber

Ecology. The prime ecological fact about members of the Cucurbitales order, one that governs their geographical distribution and area of cultivation, is their relative sensitivity to cool temperatures. There is not a single species that tolerates frost, and they are extremely sensitive to temperatures near freezing. As a group the cucurbits are intolerant of cold soils, and they are rapidly victimized by root rots and bacterial diseases in such an environment. Some of the cultivated species of *Cucurbita* (pumpkins, squashes, and gourds) exhibit great genetic plasticity, however, and plant breeders have developed cultivars with cold tolerance and a capability to grow rapidly in cool soils, but none will withstand frost.

The cultivated cucurbits require light, well drained soils with a neutral or slightly alkaline reaction for best growth. These requirements also seem to be generally true of the uncultivated species.

FORM AND FUNCTION

The gourd family, and hence order, has a distinct set of morphological (*i.e.*, structural) characters that easily distinguishes it from other plant families. Such characters include the prostrate or climbing habit, the mostly palmately lobed leaves on herbaceous (nonwoody) stems, that have long spiral tendrils at each node. Quite as distinctive are the flowers, which are mostly unisexual, with fused yellow or white petals. The ovary is inferior, that is, enclosed with the rest of the flower parts appearing to join it at its apex. There are other good morphological characters such as the basic number of stamens—five, which are usually highly modified. Seeds are usually present in large numbers, except in *Sechium*, which has a single seed. The fruit is fundamentally a berry (pepo), with a generally soft or sometimes hard skin (as in gourds). It sometimes splits open at maturity and is even explosive in *Ecballium* (squirting cucumber), but more generally it does not split open naturally.

Biochemical substances, mostly bitter, exist in the Cucurbitales order. These substances have been used as purgatives since ancient times. Bitter substances are wide-

ly distributed in the order. A systematic search for these substances in the Cucurbitales order indicates that the great majority of species investigated contain bitter principles in some portion of the plant (root, shoot or reproductive organs), at some stage of development. The most extensive search of this nature has resulted in the isolating of ten crystalline bitter substances, called cucurbitacins. Chemically, the cucurbitacins are tetracyclic triterpenes (*i.e.*, three ten-carbon terpene units joined in such a way that four cyclic or closed ringlike structures result) and occur in nature as glycocides (compounds that may be variously broken down into sugar and nonsugar components) in complicated mixtures. A comparative study of the nature and distribution of the cucurbitacins in various species provides useful information for determining relationships when combined with conventional taxonomic techniques.

EVOLUTION AND ETHNOBOTANY

Specific studies concerning the evolution of the Cucurbitales order are lacking, but those species often grouped together into the subfamily Cucurbitoideae are generally considered to be more primitive than those in the subfamily Zanonioideae. That is, the Cucurbitoideae subfamily has a predominance of characters that are considered to have appeared early in the evolution of the order. These include horizontally positioned ovules (developing seeds), stamen filaments attached to the floral tube, a single style (upper part of the ovary), and simple tendrils that spiral above the point of branching. The characters of the subfamily Zanonioideae, considered to be more advanced or to have appeared more recently in the evolution of the order, include pendulous ovules, filaments attached to or near the central part of the flower, three styles, and forked tendrils that spiral below the point of branching.

The ethnobotany, or study of the uses of plants by humans, of the squashes, pumpkins, and gourds, especially of the genera *Cucurbita* and *Lagenaria*, has become an exciting field of research within the past few years. Archeologists engaged in field work in the Americas have commenced to search out and catalogue botanical specimens as they would the common artifacts. One result has been an accumulation of authentically dated cucurbit material that can be identified at the species level. Studies of this material have provided much information about the origin and domestication of *Cucurbita* and *Lagenaria*, and the time and place where these events occurred. It is also possible by means of known archeological specimens of these genera to trace the migration of peoples and to assign to each culture the species of squash or pumpkin in common use. The white-flowered gourd (*Lagenaria siceraria*) was among the earliest of man's cultivated plants, and problems connected with its pre-Columbian distribution in the Americas have engaged the attention of geographers, archeologists, ethno-botanists and others.

CLASSIFICATION

Annotated classification.

ORDER CUCURBITALES

Annual or perennial herbaceous vines, usually with tendrils, rarely erect or suffrutescent. Leaves alternate, simple, usually with petioles and palmately or pinnately 5-lobed. Tendrils lateral, stipular in position, one at each node, simple or variously forked. Roots fibrous or fleshy, some enlarged into small or huge storage organs. Flowers mostly unisexual, monoecious or dioecious, occasionally perfect. Staminate or male flower with calyx and corolla united to form a tubular receptacle, corolla lobes usually yellow, white or cream-coloured; stamens often 3, occasionally 4 with one sterile stamen, filaments free or united, anthers free or coherent in a head, stamens basically 5, alternate with the petals, inserted on the receptacle tube or on the basal disk, but always modified in one or more ways. Pistillate or female flowers with calyx and corolla similar to the male flowers; ovary inferior, connate with receptacle to form a globose to fusiform tubular structure, often separated from the remainder of the flower by a constriction; usually with 1 locule and 3 or 4 carpels (primitively 5); style terminal with 2- or 3-lobed stigma (many lobes in some *Cucurbita*). Fruit a dry or fleshy berry or hard-shelled pepo, usually indehiscent, or occasionally dehiscent (explosive). Seeds large, numerous, rarely 1; with no endosperm, embryo with large cotyledons. One family, 90 to 100 genera (depending on the authority consulted) and about 700 species distributed mainly in the tropics and subtropics.

Family Cucurbitaceae

The only family in the order; it has the same characters as the order.

Critical appraisal. Prior to 1961 there were at least four well-known classifications of the Cucurbitaceae. In 1961, a new classification system was proposed because information from numerous recent studies permitted a re-evaluation of the relative merits of the characters used by earlier workers.

The disputed position of the Cucurbitales order within the classification system needs to be resolved. Some authorities include at least four families within the gourd order. In most of the modern treatments, however, only one family, the Cucurbitaceae, is assigned to this order. A comparative morphological study of those families obviously closely related to the Cucurbitaceae would be helpful. Such families as the Passifloraceae, Begoniaceae, and Caricaceae are good candidates for a study of this kind, along with the Campanulaceae. Comparative biochemical studies of these families might also yield data of significance for this problem.

To make progress in understanding the crucial problems of classification in the Cucurbitaceae more growth studies in experimental gardens and greenhouses are required. Such studies should include large collections of species of known origin. Collections of this kind are not easy to assemble because most species in the Cucurbitaceae are native to tropical areas. Also, the cucurbits are not the best of experimental subjects because the plants require a relatively large area to make normal growth. Thus the limitations of space, and the lack of suitable conditions for growth makes difficult the task of growing these plants for study. Nevertheless, a significant body of systematic knowledge is commencing to accumulate in several institutions particularly for the genera *Cucumis* and *Cucurbita*.

BIBLIOGRAPHY. A. COGNIAUX and H. HARMS, "Cucurbitaceae," in A. ENGLER, *Das Pflanzenreich* (1924), a classical treatment of this family; A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), an important modern work on the classification of the flowering plants; H.C. CUTLER and T.W. WHITAKER, "History and Distribution of the Cultivated Cucurbits in the Americas," *Am. Antiq.*, 26:469–485 (1961), a summary of the distribution and time sequence of archaeological cucurbits in the Americas; C. JEFFREY, "Notes on Cucurbitaceae, including a Proposed New Classification of the Family," *Kew Bull.*, 15:337–371 (1961), a modern classification of the Cucurbitaceae based upon careful study of the morphology of the various units of the family, but also taking into account the data of other disciplines, "On the Classification of the Cucurbitaceae," *Kew Bull.*, 20:417–426 (1966), an excellent statement of the history of the classification of the Cucurbitaceae up to the modern era, and *Flora of Tropical East Africa: Cucurbitaceae* (1967), one of the best studies of the Cucurbitaceae in an area where the family has filled a vast number of ecological niches; A.D.J. MEEUSE, "The Cucurbitaceae of Southern Africa," *Bothalia*, 8:1–111 (1962), a critical appraisal of the cucurbits in an area where they are comparatively numerous; C. NAUDIN, "Essais d'une monographie des espèces et des variétés du genre *Cucumis*," *Annls. Sci. Nat.*, sér. 4, *Bot.*, 11:5–87 (1859), "Revue des Cucurbitacées cultivées au Muséum en 1859," *ibid.*, 12: 79–164 (1859), two papers by the French botanist, Naudin, that were the forerunners of future biosystematic studies in the Cucurbitaceae; R.R. VON WETTSTEIN, *Handbuch der Systematischen Botanik*, vol. 1 (1901, rev. ed. 1933), a classical source of information on the classification of the flowering plants; T.W. WHITAKER and G.N. DAVIS, *Cucurbits: Botany, Cultivation and Utilization* (1962), a general discussion of the economically important species of the Cucurbitaceae.

(T.W.W.)

Cultural Areas, Theories About

A cultural area (or culture area) is an example of what human geographers call uniform or homogeneous regions. Similar concepts are found in anthropology and history although the actual terms used may vary ("culture realm," "culture province," "civilization"). The history of

the idea, if not of the term, reaches back into antiquity and typically involves three components: (1) a socially or culturally identifiable population, (2) a "natural" environment, and (3) some postulated, implied, or identified interaction between the two. Which of the three components is given the greatest emphasis, and the actual constitution of each, has varied. Areas and regions then are intellectual constructs and not real entities.

The design and purpose of cultural areas. The fundamental problem of attempting to regionalize cultural phenomena involves the identification of two basic types of regions, one called by geographers nodal and the other uniform. Nodal regions are areas of the earth's surface that are conceived of as being formed around a central point, or node. The focus of the spatial activity that occurs in this area is their node and the limits of the region are the limits of the activity or activities that are connected by various channels of circulation to the node. Every individual in the world forms a nodal region, as does every household, village, city, or nation. The individual is the focal point of his daily movements to or from work, his social or religious activities, his consumption, his communications, and much more. He may be envisioned as a point from which lines of interconnection radiate outward. A boundary drawn so as to include all or most of the external ends of these lines would encompass that particular nodal region. Uniform regions are regions defined by the existence of one or more particular traits or complex of traits within an area. An example would be a region formed by all the members of one individual's family, or kin group, or the region within which all or most of the individuals with a given surname appear. All nation-states, in this sense, are uniform regions by virtue of a common body of laws, and often a common language and religion as well. In the former instance, the uniform region is postulated in order to contain the stipulated phenomena—a kin or surname group. In the latter instance, the limits of the uniform region exist in the form of recognized political boundaries.

Nodal and uniform regions

It is immediately apparent that to establish regions, whether nodal or uniform, requires a series of careful definitions. First, the criteria to be used as defining variables must be carefully identified. For cultural features, this may be relatively simple (*e.g.*, a point where at least 50 percent of the people read and write) or more complex (*e.g.*, a social and political definition that requires the simultaneous existence of a number of phenomena, such as voting rights, a Parliament, etc.). Second, once the characteristic is identified, the intensity of the characteristic must be specified. Will literacy, for example, be defined as the ability of a person to write his name, the ability to read and write simple declarative sentences, or the ability to read and summarize a newspaper article? If the criteria used to determine an area are numerous, it must be decided if they all have equal weight and must all be present simultaneously and in equal intensity. It must be known, for example, whether in defining a region, the type of kinship grouping will be more or less important than other variables, such as economy or material culture, and whether mixtures of all will be considered. Finally, the time period must be specified. This is most evident in such ephemeral regions as, for example, "the American frontier," but it is also necessary for seemingly more permanent phenomena, such as "Chinese civilization" or "the arid regions of the world."

Problems of delimiting cultural areas

The artificial and temporary nature of regions is, thus, obvious. They are classification schemes designed by investigators in order to bring understanding to a chaotic mass of fact. They vary according to the purpose of the investigation, the time period involved, and the nature of the criteria used for delineation. This is as true for "natural" regions, those defined solely in terms of nonhuman criteria, as it is for cultural regions. Because of this artificiality it is valid to say that *all* regions are cultural; the delimitation of even natural regions is dependent upon a culturally biassed decision with regard to the specifying characteristics.

Theories from earliest times to the 19th century. The need to classify human populations in an areal fashion was recognized early in the expansion of the limits of the known world. The early Greeks, for instance, began to write about various non-Greek populations as their own sphere of discovery and activity expanded. As well as finding new peoples with different appearances and customs, they were discovering that the world had different climates. The frequent conjunction of new climates and new cultures suggested both a means of classifying the information and a causal relation between the two. This mode of reasoning stemmed in the first instance from early medicine and the "humour" theory, in which good humours or bad humours (that is, certain biological fluids) influenced the human body, and the body, in turn, influenced the mind. The origin of the humour was alleged to be in the "air, water, and places," and thus the characteristic of people could be deduced from the characteristic of place. While the first formulation of this belief is attributed to Hippocrates (flourished 400 BC), it is thought that he, in turn, was reflecting even earlier thinking. From this basic conceptualization there came a division of the known world into *klimata*, or "climates," which were held in most instances to correspond to racial and cultural differences between the human populations found in them. No new and comparable attempt at a logical and systematic classification of the culture areas of the world occurred again in the West until the middle of the 19th century. The intervening centuries saw the rise and decline of a number of regional classification systems, but none that implied or demonstrated a cause-and-effect relationship. Roman writers tended to compile or repeat the work of the earlier, more innovative Greek civilization. The resurgence of learning in the Western world, after its centuries-long lapse following the fall of Rome, brought little in the way of new ideas to the question of regional divisions. In the relatively static world that existed prior to the 16th century, knowledge of other people, places, and customs was limited to a handful of merchants, adventurers, and soldiers. The first pulse of change was faintly beating, however, and the intellectual island of European thought was about to become inundated by a wave of fact and fable from elsewhere, set into motion by the fleets of the Age of Discovery. In the 75 years between 1450 and 1525, a thriving sea trade was developed between Europe, West Africa, and Southeast and East Asia; a "New World" was discovered where none was thought to exist, and the earth was circumnavigated for the first time in history. These events and the attendant changes they brought served to call attention to the existence of great cultural and climatic diversity throughout the world. As the Greeks had done almost 2,000 years before them, men once again began to speculate on the apparent covariance of habit and habitat and to assume or imply a causal relation between them.

Early attention to climatic regions

In 1650 a German scholar, Bernhardus Varenius, set forth a regional delimitation of the known world, dividing geographic knowledge into two parts: general (what is now called systematic) and special. Special geography was to investigate areas or regions in the world that were identifiable because of specific forms of interaction between human and environmental processes. One hundred years later the philosopher Immanuel Kant was to follow and elaborate on this division. For Kant, geography, the spatial association of phenomena, was one of the few possible ways in which to organize knowledge. He himself lectured mostly in physical geography, and only briefly in what Varenius had called special geography, but his use of the dichotomy was to give it a philosophic foundation for many years to come.

Various other 18th-century writers concerned themselves with the abstract division of the subject matter of geography, but it was not until the 1790s and early 1800s that their programmatic suggestions were transformed into factual studies. Although they had valuable predecessors from whom they drew much of their inspiration, the two leading figures of this era were Germans—Alexander von Humboldt and Carl Ritter, both of whom saw the world as an organic whole. Humboldt was primarily concerned with physical and biological phenomena, Ritter with human problems in a regional context. Their

views were in most respects similar; and these similarities, in the intellectual context of the time, are of interest for the future development of the culture area concept. Both were empirically oriented; that is, whereas they believed that the world was an organic whole, functioning in accordance with fixed, if as yet unknown, natural laws, they saw the eventual unfolding of these laws as stemming from observation of fact—not from the development of a priori theory.

Ritter, whose work is shot through with teleological purpose, sought evidence of man's "progress," and this idea was reinforced in Western philosophy in 1859, when Charles Darwin's *Origin of the Species* was published. This work offered an explanatory natural law regarding the relation between organisms and their environment. Certain aspects of the thesis, particularly the idea of natural selection (a process that resulted in the "survival of the fittest," according to the social philosopher Herbert Spencer), were quickly adopted by students of human society, resulting in what has been called Social Darwinism—the doctrine that competitive struggle is as natural to social evolution as it is to biological evolution. It was primarily the materialistic aspect of Darwin that was influential in Social Darwinism. Man, by scientific "law," could now be viewed properly as just another part of nature—as an animal, highly evolved, to be sure, among other animals, but no longer as a special creation of God without precedent and unique in his very essence. He was now liable to investigation in his milieu just as were all other organisms; the interactions of man and environment differed only in quantity from the interactions of other animals with their environments.

Effects of Social Darwinism and environmental determinism

Thus the regional division of the newly discovered peoples of the world was viewed as a complex problem requiring not only the determination of regional distributions of cultural traits—such as language, religion, clothing type, pottery form, or any of the myriad other material and nonmaterial items of man's equipage—but also the determination of his cultural set in the nonhuman environment. Although man the divine might be studied in isolation from the profane world, man the animal was meaningful only in a context of soil, water, climate, and food supply. So-called natural regions had somehow to be combined with cultural regions.

The main thrust of this recombination was postponed until after the turn of the century. The natural sciences were still leading the way, and the social sciences were following after. Studies concerned with interrelations between man and environment became almost solely studies of environmental influence over man. This was so-called environmental or geographical determinism. One of the more influential of the studies of this kind was the German geographer Friedrich Ratzel's *Anthropogeographie*, in the first volume of which (1882), under the influence of Darwinian ideas, man was seen as the end product of a long evolution, during which the conditions of the natural environment had worked to shape him. Although the second volume (1891), in treating such matters as migration, diffusion of culture traits, and other social phenomena, did give culture a more active determining role, it was volume one that influenced social thought.

The year in which volume one of *Anthropogeographie* was first published saw the death of an influential French scholar, Frédéric Le Play, whose work also reflected an environmental-determinist cast. Le Play, building on the natural regions of the world that others had delineated, suggested his formula of place, work, and people. The place, identified as a natural region (steppe, tundra, forest, plain, and so on), determined, by its very nature, the livelihood, activity, or work, of the family. The basic social unit, the family, was thus structured by the natural environment. Family structure, in turn, determined the type of people who resulted, and it was these individuals who constructed suprafamily organizations, or societies. To point out that both Ratzel and Le Play confused occasional correlations with lawlike causation is not to deny their significant contributions for the time and place. They at least chose to deal with the complex whole and not the simplistic parts.

Twentieth-century theories and attitudes toward cultural areas. *The geographers.* The beginning of the 20th century saw the growing influence of a number of writers on the general subject of cultural areas. Among the most important were Sir Halford Mackinder in Great Britain, William Morris Davis in the United States, Paul Vidal de La Blache in France, and Alfred Hettner in Germany. Each of these men was concerned with the interplay of man and nature as it served to form the earth's surface into analytically recognizable areal units.

Mackinder differed from the other three in that what was to become his major contribution to geographical thought was (1) on a scale, both spatial and temporal, much grander than that of any of the others and (2) concerned with what may be viewed as a nodal region, rather than a uniform one. In 1904 he presented the concept of "the geographical pivot of history" to the Royal Geographical Society. In this paper, the main idea of which was to be widely quoted for almost 45 years, he argued that the "pivot region of the world's politics [is] that vast area of Euro-Asia which is inaccessible to ships, but in antiquity lay open to the horse-riding nomads, and is to-day about to be covered with a network of railways." He had argued that historically the steppe region of what was then Russia determined the development of a nomad way of life, which in turn was the pivotal force acting, by sporadic warfare, upon the marginal civilizations surrounding this "heartland." This pivot area was now being re-formed into a cohesive region by rail communication, and its peoples were moving into more and more southerly and easterly territories. This expanded resource base, if ever coupled with access to the sea and the development of sea power, would then form the heartland, or nodal focus, of a worldwide political system. He was to put his ideas into a succinct form in 1919 when he wrote:

The idea of the heartland

> Who rules East Europe commands the Heartland:
> Who rules the Heartland commands the World-Island:
> Who rules the World-Island commands the World.

Although the scope and purpose of his writing differed greatly from other geographers, MacKinder's initial postulate—that the steppelands of Russian Asia produced a certain form of livelihood, which in turn led to specific types of social and political structure and action—place him, like Le Play, in the general class of scholars concerned with the interaction of man and environment and with the regions thus formed. An American, William Morris Davis, who may be included in this broad group for similar reasons, was a physical geographer. His definition of geography, however, as the science that dealt with the interaction process, which he saw as environmental cause producing social effect, was to have profound impact on a whole generation of geographers.

Unlike Davis, and perhaps unlike most human geographers of his day, Paul Vidal de La Blache did not view man as passive in the combination of man and environment. On the contrary, he argued, man substituted goal-oriented action, action purposefully directed toward specific ends, for the "incoherent effects of local circumstances." In so doing, he believed, societies eventually established regions that were reflections of social purpose. He emphasized that the environment offered man a variety of possibilities rather than a fixed, determinate path, and he and his followers were consequently known as environmental possibilists or, more simply, possibilists. To be sure, there were restrictions on man, but these restrictions were cultural—the values of his culture, his social organization, and his technology. Roughly similar to what the anthropologists called culture, Vidal called this mix of attributes a society's *genre de vie*, or "way of life." Over long periods of time, man and environment would develop a series of unique regions, each reflecting the initial environmental conditions from which men choose, but also reflecting the effects of the cultural instruments involved in that choice. If the determinist saw society as conditioned by environment, Vidal and his school saw society and environment as mutually influencing, each cultural region formed by a unique combination of natural and cultural elements.

Vidal devised a regional division of all France, which provided for fifteen regions, each identified by "topography and human economic activities." These regions were nodal, each being centred upon a regional capital; and the most significant organizing principle of these nodal regions was the economic infrastructure—transportation, marketing, financing, and so forth—of a modern industrial state. The proposed regional boundaries, however, did take cognizance of older lines delineating a cultural area, an area that had developed a distinct *genre de vie*.

A major figure in geography in Germany during this period was Alfred Hettner, who denied the traditional division between physical and human geography. In asserting that environmental and cultural features were so complexly intertwined as to form a single subject, not a dual one, Hettner provided the synthesis needed between the concepts of Vidal and the natural or climatic concepts of other scholars.

The anthropologists. Within anthropology there developed four main concepts: the culture area, the age–area hypothesis, the area co-tradition, and the Kulturkreis. Whereas these are logically related in their basic purpose, they are distinct enough in underlying methodology and in substantive results as to require some comment.

By the beginning of the 20th century, archaeologists and anthropologists had already conceived of a need to design areal groupings reflecting the information derived from artifacts and fieldwork. Some museum exhibits were ordered in areal fashion, and there had been some initial attempts to divide American Indian populations into culture-area complexes. An anthropologist, O.T. Mason, had devised 11 culture areas for North America, designating them in part by locations (North Pacific coast, California, Oregon), in part by climate (Arctic), in part by physiography and drainage (Interior Basin, Columbia Drainage), in part by language (Athabascan, Algonkian), and so on. Appropriately for the time, the work in which these were set out was entitled "Influence of Environment upon Human Industries and Arts." His basic subdivisions, if not his names for them, were echoed more or less in the classifications of other anthropologists; but the significant debate inhered not in the classifications per se but in whether the divisions were real and objective or merely analytical constructs. In 1917 another anthropologist, Clark Wissler, recognized, at least in his division of the peoples of the North American continent, that he was offering only analytical constructs:

Culture areas

> It is customary to divide the continent into culture areas the boundaries to *which are provisional and transitional*, but which taken in the large *enable us to make convenient distinctions* (Italics added).

By 1923, however, another anthropologist, Alfred Kroeber, was claiming greater reality for his almost identical areal divisions. Justifying the emphasis on spatial distinctions on the grounds that for America there was little or no historical evidence available for developing temporal ones, he wrote:

> The result has been the recognition of a series of culture-areas or culture-centers. . . . These geographically defined types of culture are gradual and empirical findings. They are not the product of a scheme or imagination, nor the results of theory. . . . They do represent a consensus of opinion as to the classification of a mass of facts . . . in short, a non-philosophical, inductive, mainly unimpeachable organization of phenomena analogous to the "natural" classification of animals and plants on which systematic biology rests. (From *Anthropology;* Harcourt, Brace and Co., 1923.)

This passage connects a number of themes in the development of the culture-area concept. First, there is no distinction as yet between uniform and nodal regions (culture areas and culture centres are equated). Second, there is a strong insistence upon the scientific, inductive, nonspeculative base of the concepts (a probable reaction against the highly speculative evolutionary schemes of 19th-century writers, as well as a probable attempt to forestall criticism from those who insisted on extremely high standards of evidence and, partly as a consequence, rejected most contemporary attempts at generalization or theory formulation). Third, Kroeber continues to exhibit a fondness for using the biological sciences as the basis for analogy in anthropological study.

Wissler was to further the concept of the culture centre as well as that of the culture area. He fused the two in his so-called age–area hypothesis. In this view, culture elements diffused outward from an initial central point. The development of this central point, or node, as one would now call it, was accidental or "historical." Once established, however, culture and environment entered into an interaction, the dominant part of which was environmental because of the relative unchangeability of the natural environment. Nonetheless, argued Wissler, culture elements have a tendency to diffuse equally in all directions from this centre. From this it could therefore be inferred that, all things being equal, the element or elements having the widest distribution were the oldest —that is, had been in the process of diffusion the longest. This areal extent was the basis for inferring age, and this was called the age–area hypothesis.

Temporal theories: age–area and area co-tradition

Culture areas were time specific; that is, they attempted to organize cultural elements in a spatial framework frozen at a single instant in time. The age–area hypothesis was an attempt to introduce temporal depth by pure inference. In some parts of the New World, South America in particular, there existed a chronological sequence of culture areas. It was suggested that whenever such a sequence existed simultaneously for more than one culture area and wherever it could be shown or inferred that the cultures had been interacting through time, it should be called an area co-tradition. The application of the concept was limited to the New World.

In Europe, where there was much more archaeological and historical evidence upon which to base typological constructs, the spatially oriented culture-area concept got little attention, whereas the temporally oriented concept of Kulturkreis (culture circle) received a great deal. Whereas both approaches were attempts to explain the distribution of culture traits, they approached the phenomena from opposite philosophical and methodological directions. The culture-area concept suggested, among other things, that there existed a body of interrelated culture elements that, in interaction with the environment, produced a culture complex that may or may not have been unique, but that was produced *in situ*. The Kulturkreis group had its intellectual origins in the writings of Friedrich Ratzel, particularly in the second volume of *Anthropogeographie*. Ratzel and his group insisted that migration and diffusion accounted for the distribution of most culture traits, and, thus, that independent invention was extremely rare. There were, in fact, only a limited number of Kulturkreise, or fonts of diffusion. According to the Austrian Father Schmidt,

Kulturkreis

> One of the facts which has been established by culture history beyond all peradventure of doubt is that not only discrete culture elements or small groups of elements migrate and exert influence, but also whole compact culture complexes. If such a culture complex embraces all the essential and necessary categories of human culture, material culture, economic life, social life, custom, religion, then we call it a "culture circle" [*Kulturkreis*], because returning into itself, like a circle, it is sufficient unto itself . . . (From Wilhelm Schmidt, *The Culture Historical Method of Ethnology;* Eng. trans. by S.A. Sieber, 1939.)

The distribution of cultural traits, in this view, is generally the result of the diffusion of a trait, set of traits, or entire culture complex from one or more of these original circles.

In sum, with the exception of the concept of area co-tradition, which did not appear until after World War II, by the 1930s anthropology was beset with two conflicting schools. In America there were culture–area and age–area scholars. In Europe the Kulturkreis school was attempting to explain world history in terms of migration and diffusion.

The landscape school. A different trend had developed in geography. By the early 1920s it was generally held that the study of areas or regions constituted the basis of geographic research. Simplistic environmental de-

terminism was pretty well dead, but there was yet no commonly accepted subject matter that would serve to delimit the discipline. An attempt to fill this need was made by the American geographer Carl O. Sauer when he published a paper entitled "The Morphology of Landscape" in 1925. Rejecting a definition of geography as the study of relations between environment and man on the grounds that relations do not constitute a subject matter, he suggested instead the study of what he called landscape. By landscape he meant not (as in conventional usage) the totality of what the eye could see before it at any given time but rather something closely akin to what sociologist Max Weber had called an ideal type, "a summation of general characteristics" of a place, or a kind of purified image of a specified scene.

The concept of landscape

The study of cultural landscape became a genetic study of its formation and development. Cultural landscapes, for the geographer, were the result of cultural factors (populations, technology, etc.) working through time on the medium of the natural landscape. The base line from which such a study was to begin was the undisturbed natural environment that existed prior to the arrival of any human group. It was basically on this idea that the concept tended to founder, for such undisturbed landscapes are exceedingly rare anyplace on earth, and one could often only speculate on the character of such landscapes in past ages; research, in any case, tended to become removed further and further back in time and more and more remote in place. In a sense this landscape school provided an uneasy link between the culture areas of American anthropology and the Kulturkreise of the Europeans. In order to trace thoroughly the evolution of a given landscape, it became necessary to trace the origins and migrations of its successive inhabitants, and an interest in cultural origins and dispersals became a distinctive trait of this brand of cultural geography.

Decline of cultural area concepts. Although Darwin had provided some concepts for rigorous analysis, he had also provided an excuse for innumerable exercises in speculative history, most particularly when evolutionary ideas were applied to human society. In reaction to this wave of frequently groundless theorizing, geographers, anthropologists, and other social scientists sought refuge in an empiricism so strict as almost to deny the probability of the sun rising on any given day. Such terms as a priori and words like theory became anathema; and much that was written, particularly during the first decades of the 20th century, was prefaced by pronouncements of faith in empirical evidence and inductive reasoning, a denial of philosophical purpose or any pretense of establishing general principles of interaction. The observations in these disciplines, however, appeared to require explanation as well as description; and while frequently denying any attempt at generalization, a number of scholars attempted to imply or state the existence of general tendencies seemingly inherent in various phenomena. Such statements of generalizations or hypothetical tendencies are not theories, however, and they vary widely in their applicability or explanatory power.

Today the concepts of cultural areas, cultural landscapes, and the like are perhaps best known as teaching devices or as simple typological structures for ordering data. As such they remain useful analytical concepts. The fundamental areal assumptions of the concepts are still reflected in the curricular structure of courses in schools; and most professional scholars in geography and anthropology are identified as specialists in one, or perhaps two, culture areas. This is becoming less true in both disciplines, however, as systematic concepts are replacing what are now viewed as typological ones.

BIBLIOGRAPHY. For the pre-scientific era, R.H. LOWIE, *The History of Ethnological Theory*, ch. 1 (1937), provides a critical summary from the anthropological viewpoint; although the same author's "The Renaissance Foundations of Anthropology," *Am. Anthrop.*, 67:1–20 (1965), presents a new interpretation of Renaissance foundations. A critical re-evaluation and comparison of the Kulturkreis and culture-area schools will be found in MARVIN HARRIS, *The Rise of Anthropological Theory* (1968). In the geographical field, R.E. DICKINSON and O.J.R. HOWARTH, *The Making of Geography* (1933), similarly remains the standard text for developments prior to 1800. For 19th-century developments RICHARD HARTSHORNE, "The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past," *Ann. Ass. Am. Geogr.*, 29:173–412 (1939), is an exhaustive and standard guide that is indispensable for the consideration of the philosophical aspects of areal differentiation, particularly as seen in Germany. For the formative and classical periods of regional geography and culture-area theory in the first half of the 20th century, there are many studies. For the French school, a general survey and bibliography is given by R.J. HARRISON CHURCH in T.G. TAYLOR (ed.), *Geography in the Twentieth Century*, 3rd enl. ed. (1957). An excellent historical evaluation of Vidal de la Blache and his school is by E.A. WRIGELEY in RICHARD J. CHORLEY and PETER HAGGETT (eds.), *Frontiers in Geographical Teaching* (1965). For German geography under Hettner, see RICHARD HARTSHORNE, *op. cit.*, for an exhaustive treatment, with bibliography. For British regional studies, SIR C.F. FOX, *The Personality of Britain*, 4th ed. (1947, reprinted 1959); and C.B. FAWCETT, "Natural Divisions of England," *Geogrl. J.*, 49:125–141 (1917), are type studies. For the American cultural-area school, R.F. BENEDICT, "Configurations of Culture in North America," *Am. Anthrop.*, 34: 1–27 (1932), contains a useful introduction. E. BACON, "A Preliminary Attempt to Determine the Culture Areas of Asia," *SWest. J. Anthrop.*, 2:117–132 (1946), is an illustrative type study. There is a scattering of surveys on contemporary developments. Among them, R.T. ANDERSON, "Recent Trends in Ethnology," *Ann. Am. Acad. Pol. Soc. Sci.*, 369:141–148 (1967), is a brief survey; and SISTER MARY ANNETTE, "The Changing French Region," *Prof. Geogr.*, 17:1–5 (1965), reviews regionalist controversy in French geography. Alternative interpretations of trends in American geography may be found in B.J.L. BERRY, "Approaches to Regional Analysis: A Synthesis," *Ann. Ass. Am. Geogr.*, 54:2–11 (1964); and J.D. CLARKSON, "Ecology and Spatial Analysis," *ibid.*, 60:700–716 (1970); E.G. BOWEN, "Le Pays de Galles," *Publs. Inst. Br. Geogr.*, no. 26 (1959), represents recent British geographical and anthropological critiques of early regionalist studies. D.W. SCHWARTZ, "Culture Area and Time Depth: The Four Worlds of the Havasupai," *Am. Anthrop.*, 61:1060–1070 (1959), has a useful introduction; A. SPOEHR, "The Part and the Whole: Reflections on the Study of a Region," *Am. Anthrop.*, 68:629–640 (1966), surveys current thought relating to regions. J.H. STEWARD, *Theory of Culture Change: The Methodology of Multilinean Evolution* (1955), is an important critique of culture-area studies in America.

(J.D.Cl.)

Cuneiform Law

Cuneiform law embraces the body of laws revealed by documents written in cuneiform script, a writing invented by the ancient Sumerians and used in the Near East in the last three millennia BC. It includes the laws of the majority of the ancient inhabitants of the Near East—especially the Sumerians, Babylonians, Assyrians, Elamites, Hurrians, Kassites, and Hittites—who, in spite of many ethnic differences, were in sufficient contact with one another to develop similar civilizations. In certain periods this cultural community was reinforced by the diffusion of Akkadian, a diplomatic and scholastic language written in cuneiform. Thus to group together a certain body of laws and name them cuneiform, far from being arbitrary, is a scientific necessity. No other term would do: "Mesopotamian law" would cover only part of the range of laws involved; the notion of ancient "Near Eastern laws" is too vast, for it would also include, for instance, both Judaic and Egyptian laws, which were separate developments (see JUDAISM; EGYPTIAN LAW).

The law with which this article is concerned involves those peoples who inhabited Mesopotamia, the "land between the rivers" (*i.e.*, the Tigris and Euphrates), as well as adjacent areas. Three main peoples contributed to the civilization of Mesopotamia. The earliest, dating from before 3000 BC, were the Sumerians, who lived in a small area lying near the mouths of the rivers. Early in the 2nd millennium BC political power began to move north up the Euphrates to the city of Babylon. The brilliant 1st dynasty of the Babylonians lasted some 300 years and reached its greatest height about the 18th century BC under Hammurabi. Hammurabi spread the rule of the Babylonians south into Sumer and westward into Syria

The civilizations of Sumer, Babylonia, and Assyria

on the Mediterranean. After his death, the empire began to crumble, and wave after wave of tribes and nations swept down from the north and east—among them, the Hittites from Asia Minor, the Hurrians from the Caucasus, and the Kassites and the Elamites from Persia. Finally, a centre of power arose once again toward the end of the 2nd millenium BC, this time in Assyria, in the hill country of the upper Tigris. Lying at the nexus of a great trade and migratory route, Assyria was frequently attacked from both north and south, and the Assyrians developed as a warrior people governed by the strictest of laws. In their great periods of empire, from the 9th to the 7th century BC, their control extended over vast areas from the Mediterranean to the Persian Gulf.

Much of the evidence of these civilizations is relatively new. The nature of cuneiform law itself began to be revealed only about a century ago, more particularly since 1900, with the decipherment of inscriptions on stone and clay tablets unearthed in the Near East. These documents now number about half a million, three-quarters of them more or less directly related to the history of law—dealing, as they do, with contracts, acknowledgments of debts, receipts, inventories, and accounts, as well as containing records and minutes of judgments rendered in courts, business letters, administrative and diplomatic correspondence, laws, international treaties, and other official transactions. The total evidence enables the historian to reach back as far as the beginnings of writing, to the dawn of history. Unfortunately, however, the complexity of the writing system and the difficulties inherent in trying to understand an ancient mentality impose limits on present knowledge of cuneiform law. Moreover, because of the inconvenience of writing in stone or clay, Mesopotamians wrote only when economic or political necessity demanded it. There are no collections of juridical doctrines or historical and literary works of an interpretive nature comparable to those of the Greeks and the Romans, and thus the reconstruction of certain ideas and facts must be somewhat tentative.

THE COLLECTIONS OF CUNEIFORM LAWS

The general features of the law codes

The various collections of cuneiform laws developed by the several nations and kingdoms have certain features in common: (1) The text of several collections contains a prologue and an epilogue in which the prince emphasizes the importance of his actions, explains the object of his work, and commands its observance by blessings or threats. (2) Though written as if inspired by the gods, the legislation is secular, composed of dispositions fixed and codified by the temporal lord. (3) Though the laws may derive from different sources—custom, judicial decisions, or deliberate legislation—the fact that they are introduced by the prince gives them all the character of legislation or enactment. (4) In contrast to modern codes, these ancient Oriental "codes" do not systematically treat all the rules applicable to a given area of law; that is, they treat a variety of matters but often ignore many highly important rules simply because such rules were so grounded in custom that they went unquestioned. (5) Because legal customs were generally known, the collections focussed on explaining individual cases, setting them up as examples or precedents, and did not represent an attempt to devise general, abstract formulas. (6) Because of this absence of doctrinal intent, the order of presentation of the cases often seems, at least to modern eyes, erratic. Although sometimes it appears that the draftsman arranged the material according to some juridical criteria or some association of ideas or concrete subjects, the arrangements often defy modern interpretation.

Here it is possibly only to illustrate some of the major laws or codes extant. The most ancient legislator known is Ur-Nammu, the founder of one of the Sumerian dynasties at the city of Ur. His code, dating about 2100 BC, dealt with witchcraft, the flight of slaves, and bodily injuries. A more ample vestige of Sumerian law is the so-called Code of Lipit–Ishtar (*c.* 1934–24 BC), which contains the typical prologue, articles, and epilogue and deals with such matters as the rights of persons, marriages, successions, penalties, and property and contracts.

The Code of Hammurabi

Although earlier Babylonian codes are known, unquestionably the most perfect monument of Babylonian law is the Code of Hammurabi (*c.* 1758 BC), the main record of which was discovered on a stele, or stone monument, only in 1901–02. At the top of the stele a low relief represents the king in prayer before the god of justice; beneath it, engraved, is the text of laws, consisting, apart from the prologue and epilogue, of no less than 282 paragraphs. The fact that copies of parts of the code have been discovered elsewhere, in nations scattered over a millennium of time, confirms that the code had a lasting interest in the ancient Orient, even in countries where it no longer had force. Like some other Near Eastern codes, the Code of Hammurabi deals consecutively with penal law, the law of persons, family law, and price lists. It differs from earlier codes, as well as from the earliest laws of the Western world of Greece and Rome, with regard to the relative importance given to laws concerning property and other economic matters. Indeed, Babylonian society of the 1st dynasty is, in terms of its individualism, wealth of private property, and development of commercial exchange, far more "modern" than, for example, Roman society of the early republic.

Though developed later in time, the Assyrian laws summon up the image of a society of a lower cultural standard. The existing tablets, reflecting legal conditions of the 15th to the 13th century BC, before the Assyrian age of empire, deal with personal property, with questions of landed property, and with women and family law. The law is that of the patriarchal society, rather strict in its rules.

Among the laws of other peoples who invaded the area may be mentioned the Hittite Law Code, dating from about the 14th century BC. It reflects a society that had remained largely on the level of a rather closed rural economy, dominated by a feudal aristocracy. The penal laws, though less severe than those in Assyria, were nevertheless rather barbarous insofar as pecuniary compensations were concerned.

THE AREAS OF CUNEIFORM LAW

Law of persons and family law. At the dawn of the 3rd millennium BC the life and polity of Sumerian communities centred on the local temple, where citizens gathered in assembly not only to worship but also to decide matters relating to the creation and upkeep of the network of dykes and canals indispensable to life in the alluvial plain. Though there was a king, a kind of "president," society was founded largely on equality. Eventually, however, the king's power grew, and there developed privileged civil servants, and thus family property—the origin of economic distinctions. Stratification became even more noticeable when, as weaker cities were absorbed by stronger, territorial kingdoms took shape. In a typical principality, society came to be divided into three classes: (1) the free men who had high status, but whom one could hardly call a citizen, because increasing despotism had eroded his participation in governing public affairs; (2) the free men of inferior status, who paid or received lower sums in legal actions, as offender or victim; and (3) the slave, a human being who could be sold and was considered property. Slavery, unknown to the first societies founded on equality, developed in Mesopotamia when captives were taken in wars (the ideogram for "slave" meant "the man from a foreign land," or "the enemy"). Slaves increased in number as children inherited their parents' bondage, as delivery into servitude became a common means of punishment, and as free children began to be sold into slavery. In some respects the conditions of a Mesopotamian slave was more favourable than that, for instance, of a Roman slave: he could own property, appear in court, and contract marriage, not only with another slave but also with a free person. This extraordinary legal capacity reflects an absence of dogmatism: the Orientals never thought of artificially classifying the slave among "things."

Social stratification: free men and slaves

Marriage and family. The Mesopotamian family was patriarchal and established on marriage and adoption. Marriage was a union less of two individuals than of

two families. This concept, for example, accounts for the levirate, a custom obliging or authorizing a man to marry his brother's widow, a custom found in such nations as Assyria. The practice occurred only if the widow had no male issue and might be explained as a means of safeguarding the family succession by offering the widow the possibility of procreating sons derived from the deceased husband's bloodline. The Babylonians and the Hurrians, who did not practice levirate, had an analogous practice: the head of the family could guard against the unfortunate effects of the premature death of his son by "adopting" his daughter-in-law and uniting her to one of his other sons.

In Sumer, marriage took place before judges; in Babylon it was concluded by a convention, often written, between the future husband (or his father) and the father of the bride. In most Mesopotamian societies the bride brought with her a dowry, and the man paid a bride-price as well. Besides the normal marriage, based on a written document, the Assyrians knew marriage that involved raising a concubine to the rank of a legitimate wife and marriage that resulted after a widow had cohabited with a man for two years. Monogamy was moderated in several ways: in Babylon, certain husbands could have a second wife; in Assyria, a man sometimes had a second wife ranking below the principal one; and the levirate, of course, could result in a man's being united with his brother's widow as well as having his own chosen wife.

Paternal and husbandly authority

The authority of the head of the family appears to have been less severe in Mesopotamia than it was, for example, in Rome and apparently did not last throughout his lifetime and the lifetime of his children. As punishment, the father could alienate or disown his child, but it seems that he could sell his children only in case of need; that is, sworn poverty. Husbandly authority, though strict in Assyria, was compromised elsewhere by the wide freedom of action at least tacitly given to the wife; in fact, she was generally subordinate only in the sense that everyone was under the authority of the head of the family. When a young girl, she was under the authority of her father, when married under that of her husband.

Divorce. In Babylon a husband might repudiate his wife if she committed certain offenses; in such cases, he could send her away without any compensation or even reduce her to servitude in her own house. He could also divorce her for reasons of alleged sterility, but he had to give her parting compensation. If he repudiated her for no reason, she was entitled to take back her dowry (if there were a dowry), and the judge could confer upon her the lifetime use of certain properties belonging to the husband, as well as the guardianship of the children. She could sometimes even succeed to his inheritance, even if she remarried. For the wife's part, she could obtain divorce for serious wrongs committed by her husband, but her risk was that a claim deemed to be unjustified exposed her to a death sentence. The Assyrian system was even less favourable to the wife; she was apparently never entitled to compensation, regardless of the nature of the divorce.

Adoption. Adoption served a variety of purposes: to create descendants or increase their number for the continuance of a family or business, to legitimate a bastard son, to liberate a slave, to establish an heir, to give a wife to a son, or to grant someone an inalienable property. One could adopt a son, a brother, a daughter-in-law, even a father.

Inheritance. The patrimony went first to all the sons or their male descendants (primogeniture, or preferential inheritance on the part of an eldest son, was known only in Assyria and a few Babylonian communities), but the exclusion of daughters was more apparent than real, since the dowry was considered an advance on the inheritance; in any case, if there were no sons, the daughters inherited. In the absence of descendants, the inheritance went to the brothers of the deceased.

A father might favour one of his sons, but a complete disinheritance was possible only if it were judicially established that a child had committed a serious offense.

Property law and contract law. Three kinds of property existed: state property, family property, and private (or individual) property. In certain ages and nations—such as the 3rd dynasty of Ur and among the Kassites—there existed a variety of state socialism, with most property owned collectively; in the earliest Sumerian period all land or soil belonged to the god of the city; that is, it was collectively owned. On the other hand, during the great 1st dynasty of Babylonia, private property assumed prime importance in the nation's economic wealth. At all times, records of real estate indicate the continuing importance of jointly owned family property.

Varieties of property ownership: state, familial, and individual

In all Mesopotamian societies there existed land tenures, which were inalienable, though transmissible by inheritance, and which were granted by the prince to those whom he wished to reward for personal services. Among some of the more warlike nations of Mesopotamia, this system resembled feudalism; that is, as in many military societies of the past, the maintenance of expensive armament and chariots obliged a prince to grant his warriors the economic means to acquire and keep such equipment in good condition.

Among all Mesopotamian peoples, the different varieties of contracts—exchange, sale, loan, deposit, surety—were practiced without there having appeared the abstract notion that a contract engendered an obligation. A sale always remained an operation in cash. To obtain the effects of a sale on credit, for example, the buyer or seller resorted to fictions: the seller "loaned" the buyer the object in question or "deposited" it with the buyer until payment was made. Conversely, when payment was made in a purchase for future delivery, the seller might be considered a temporary "tenant" of the property, or he might be considered to have retained it on "loan."

Penal law. The most ancient societies sanctioned private revenge and the execution of severe punishments within the household. Although such practices survived for some acts (notably in Assyria), the state tended to take over more and more jurisdiction of crimes and other injuries and wrongs, introducing fines as well as corporal punishment.

Lex talionis

In the Code of Hammurabi and the Assyrian laws, the ruling principle of punishment was the lex talionis, or "eye for an eye," though the death penalty was also freely inflicted not only on murderers but also on a variety of thieves, brigands, persons guilty of criminal negligence, false accusers, and so on. Corporal punishment was inflicted according to the nature of the offense: for example, the hand that struck a father was cut off. The notion of a "life for a life" extended to such extremes as the execution of a builder's son for negligence in construction causing the death of a purchaser's son.

Even in its most barbarous manifestations, however, penal law showed a concern for impartiality. The intention of the offender was taken into consideration: the man who was unaware that his sexual partner was a married woman was not guilty of adultery; a shepherd was not liable for damages inflicted by wild beasts on the cattle in his care. Penalties or rewards must be equitable; co-authors must suffer the same punishment or profit equally from an action.

BIBLIOGRAPHY. The chief bibliographical work is the *Introduction bibliographique à l'histoire du droit*, specifically vol. A/2, *Droits cunéiformes*, ed. by G. CARDASCIA and J. KLIMA (1966), and vol. A/3, *Hethitisches Recht*, ed. by R. HAASE (1967). Recent interpretive works include: J. GAUDEMET, *Institutions de l'antiquité*, pp. 13–95 (1967); R. HAASE, *Einführung in das Studium keilschriftlicher Rechtsquellen* (1965); and V. KOROSEC, "Keilschriftrecht," in *Handbuch der Orientalistik*, pt. 1, suppl. vol. 3 (1964). A.L. OPPENHEIM, *Ancient Mesopotamia: Portrait of a Dead Civilization* (1964), describes the social, economic, and political structures. The best legislative sources are the following: R. HAASE, *Die keilschriftlichen Rechtssammlungen in deutschen Übersetzung* (1963); S.N. KRAMER, "Ur-Nammu Law Code," *Orientalia*, 23:40–48 (1954); F.R. STEELE, "The Code of Lipit-Ishtar," *American Journal of Archaeology*, 52:425–450 (1948); R. YARON, *The Laws of Eshnunna* (1969); G.R. DRIVER and J.C. MILES (eds.), *The Babylonian Laws*, 2 vol. (1952–55), *The Assyrian Laws* (1935); and E. NEUFELD, *The Hittite Laws* (1951).

(G.Ca.)

Curie Family

For 60 years, two generations of the French Curie family played a prominent role in the development of modern physics, beginning with studies of crystal symmetry and the discovery of polonium and radium, and extending to the building of nuclear reactors. Pierre and Marie Curie, their daughter Irène, and son-in-law Frédéric Joliot, all recipients of the Nobel Prize, provided fundamental insights into the constitution of the atom. Their research on radioactivity led to an increasingly close alliance between atomic physics and theoretical chemistry.

Pierre and Marie Curie. *Pierre Curie.* Born in Paris on May 15, 1859, Pierre was educated by his father, who was a physician. At 14 he developed a keen interest in mathematics and revealed a particular aptitude for spatial geometry. Matriculating at the Sorbonne at 16, he obtained his *licence ès sciences* (bachelor of science) at 18 and in 1878 became a laboratory assistant there. In the same year Pierre performed his first work on the calculation of the wavelength of heat waves. He then studied crystals with his elder brother Jacques. The Curie brothers associated the phenomenon of pyroelectricity, explained by Lord Kelvin—the appearance of electric charges due to heating—with a change in the volume of the crystal on which it appears; following this work, they discovered piezoelectricity—the generation of electric charges in a nonconducting crystal that is subjected to pressure and, conversely, the change in volume in certain crystals when subjected to an electric field. Later Pierre formulated the principle of symmetry, which states that it is impossible to bring about a specific process in an environment lacking a certain minimal dissymmetry characteristic of the process, and holds that this dissymmetry cannot be found in the effect if it does not exist in the cause. He then defined the symmetry of different physical phenomena, such as the relation between electrical field and electrical current in conducting crystals.

Discovery of piezoelectricity and principle of symmetry

After his appointment in 1882 as supervisor at the École de Physique et de Chimie Industrielle at Paris, Pierre continued his research. He perfected an aperiodic, analytical balance in which the last decimal weight can be read promptly and directly, the oscillations of the beam being rapidly damped. Then, for his doctoral thesis he began the study of magnetism with the aim of discovering if transitions exist between the three types: ferromagnetism, paramagnetism, and diamagnetism. In order to measure how much force a magnetic field exerted on a sample, he constructed a very accurate torsion balance, which, in a simplified version, is now called the magnetic balance of Curie and Chèneveau. He discovered Curie's law—that the magnetic susceptibility of paramagnetic bodies varies in inverse proportion to the absolute temperature. He established an analogy between paramagnetic bodies and perfect gases and between ferromagnetic bodies and condensed fluids, suggesting that the phenomenon of ferromagnetism could be represented by the same type of equation used for fluids, as it was later established by the physicist Pierre-Ernest Weiss. The demonstration given by Curie of the totally different character of paramagnetism and diamagnetism was later also explained theoretically by Paul Langevin, the French physicist. In 1895 Curie defended his thesis on magnetic properties at different temperatures and was granted a doctorate of science.

Marie Curie. Manya Skłodowska (familiarly known as Mania and later by her French name Marie) had been born on November 7, 1867, in Warsaw. From childhood she had been remarkable for her prodigious memory and at 16 had won a gold medal on completion of her secondary education at the Russian lycée. Because her father, a teacher of mathematics and physics, had lost his savings through bad investment, she had been obliged to help the family. While working as a teacher she had clandestinely taken part in a nationalist "free university" (*l'université volante*) by reading in Polish to women workers; and at 18 she had taken a post as governess, contributing her earnings to her sister Bronia's medical studies in Paris, on the understanding that Bronia later would help her.

In 1891 Marie Skłodowska had gone to Paris, where she attended the lectures of Paul Appell, Gabriel Lippmann, and Edmond Bouty at the Sorbonne. Marie often worked far into the night in her garret in the students' quarter, where she lived on a diet of bread, butter, and tea. She had achieved the *licence ès sciences physiques*, ranking first place, in 1893 and began to work in Lippmann's laboratory. In 1894, the year she met Pierre Curie, she had obtained the *licence ès sciences mathématiques*, ranking second.

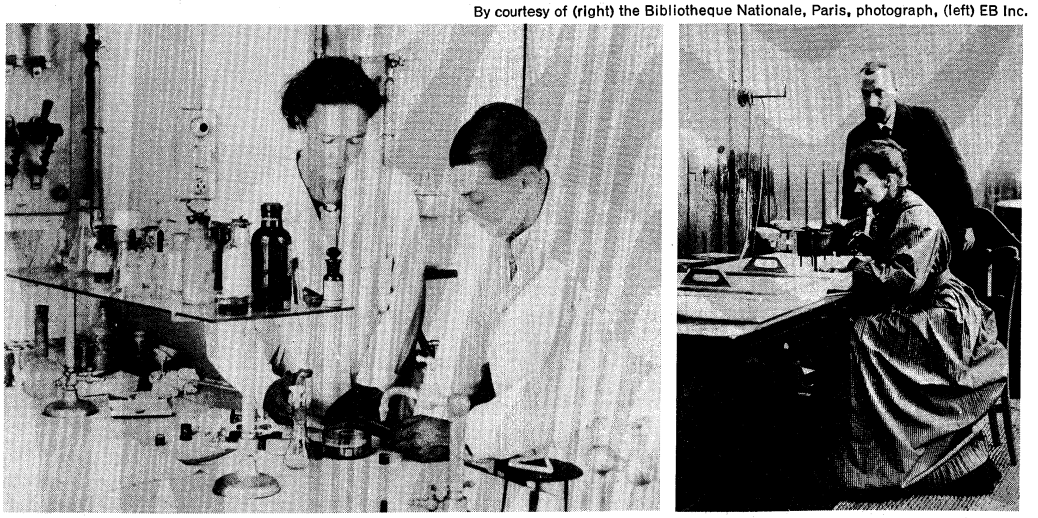
Joint work. The marriage of Pierre and Marie on July 25, 1895, marked the start of a partnership that soon achieved world significance. Their first daughter, Irène, was born in 1897.

Looking for a thesis subject, Marie became intrigued by recent discoveries in physics: Henri Becquerel had discovered in 1896 a new phenomenon, the spontaneous emission by salts of uranium of rays similar to X-rays, and it was known that electrically charged bodies were discharged by gases submitted to those rays. Deciding to find out if the property discovered in uranium was to be found in other matter, she discovered that it was true for thorium at the same time as did the German chemist Gerhard Carl Schmidt.

Analysis of pitchblende

As she began to study minerals, her attention was drawn to the mineral pitchblende, the emission of rays from which is superior to that of pure uranium; the activity of pitchblende could be explained only by the presence in the ore of an unknown substance. Pierre Curie joined her in the work she had undertaken to resolve the problem.

By courtesy of (right) the Bibliotheque Nationale, Paris, photograph, (left) EB Inc.



(Left) Irène and Frédéric Joliot-Curie. (Right) Marie and Pierre Curie.

By using an ionization chamber and a piezoelectric quartz electrometer, he could compensate for the electric current produced by the active sample by adjusting the quartz. Thus the electric charges produced by the radioactivity were measured by weights applied to the quartz. While working with Marie to extract pure substances from ores, which they achieved using relatively primitive conditions, Pierre also concentrated on the physical study, including the luminous and chemical effects of the new radiations. By measuring the action of magnetic fields on the rays given out by the radium, he proved the existence of varying amounts of three types of particles—electrically positive, negative, and neutral ones that Sir Ernest Rutherford later called alpha, beta, and gamma rays. The varying amounts of particles were caused by radioactivity, a term first used by Marie in 1898, in the thorium. Moreover, the measurement techniques of Marie and Pierre in the same year led them to announce the discovery of two new elements, polonium (in honour of her country) and radium. Pierre studied then the radiations by calorimetry and also observed the physiological effects of radium.

Discovery of polonium and radium

Refusing a chair at the University of Geneva in order to continue his joint work with Marie, Pierre was appointed lecturer in 1900 and professor in 1904 at the Sorbonne. In 1900 Marie was appointed a lecturer in physics at a school for girls at Sèvres where she introduced a method of teaching science based on experimental demonstrations. In 1903 the Curies and Becquerel shared the Nobel Prize for Physics for the discovery of radioactivity; and in the same year Marie received her doctorate of science and Pierre and Marie were awarded the Davy Medal of the Royal Society of London. The Curies' second daughter, Eve, was born in 1904. In December 1904, Marie was appointed chief assistant in the laboratory directed by Pierre, who was elected to the Académie des Sciences in 1905. Pierre was run over by a dray in Paris on April 19, 1906, and died instantly.

The work continued. The sudden death of Pierre Curie, although it was a bitter blow to Marie, was also a decisive turning point in her career; henceforth she devoted all her energy to completing the scientific work they had undertaken. On May 13, 1906, she was appointed to the professorship that had been left vacant on her husband's death, becoming thereby the first woman to teach in the Sorbonne. In 1908 she became titular professor. In 1910 her fundamental treatise on radioactivity was published; the same year (with the help of the French chemist André-Louis Debierne), she obtained pure radium in the metallic state. In 1911 she was awarded the Nobel Prize for Chemistry, for the discovery of radium and polonium and the isolation of pure radium.

Medical application of radium

Throughout World War I, Marie Curie, with the help of her daughter Irène, worked on the application of X-rays for medical diagnosis, particularly in the treatment of wounded soldiers. In 1918 the Radium Institute, the staff of which Irène had joined, began to function in earnest. It was to become a major centre for nuclear physics and chemistry. Marie Curie, then at the height of her fame, in 1922 became a member of the Académie de Médecine; she devoted her researches to the study of the chemistry and medical applications of radioactive substances.

In 1921, accompanied by her two daughters, Marie made a triumphant journey to the United States, where President Harding presented her with a gram of radium bought as a result of a fund drive among American women. Continuing her travels overseas, she gave lectures and was made a member of the International Commission on Intellectual Cooperation by the Council of the League of Nations. She also had the satisfaction of seeing the development of the Curie Foundation in Paris and the inauguration in 1932 in Warsaw of the Maria Skłodowska-Curie Radium Institute, of which her sister Bronia became director. But her chief joy was the success of the new team formed by her daughter Irène and the brilliant Frédéric Joliot.

Irène Curie and Frédéric Joliot. *Irène Curie.* Born on September 12, 1897, Irène from 1912 to 1914 prepared for her *baccalauréat* at the Collège Sévigné. During World War I, while working as an X-ray technician in the army, she began work toward a degree in physics. As Marie Curie's assistant in 1918, she served an apprenticeship in the manipulations and precision measuring of radioactive elements and proved to have her father's aptitude for using ingenious devices. She used the new cloud chamber, devised by the Scottish physicist C.T.R. Wilson, in which ionizing radiation induces the formation of droplets of moisture, giving an image of the trajectory of the rays in the chamber. In 1925 she presented her doctoral thesis on the alpha rays of polonium. In the same year she met Frédéric Joliot in her mother's laboratory; she was to find in him a mate who shared her infatuation with science, sports, humanism, and the arts.

Frédéric Joliot. Born on March 19, 1900, Jean-Frédéric Joliot was the last of six children. His father had passed on to him a fondness for hunting, fishing, and music. Frédéric's mother (*née* Roederer) held liberal republican opinions that greatly influenced him, and Frédéric had acquired a liking for painting from his gifted sister Marguerite. As a boarding student at the Lycée Lakanal, the young man had distinguished himself more in sports than in studies. Reversals of family fortune had then forced young Joliot to choose a free public education at the Lavoisier municipal school in order to prepare for the entrance competition at the École de Physique et de Chimie Industrielle, where Pierre Curie had been supervisor. Frédéric had been accepted in 1919 but because of illness had not entered until 1920. There the director of studies, Paul Langevin, had influenced him in scientific, pedagogical, and social questions. Frédéric, who had acquired a great admiration for the work of the Curies, had been graduated, ranking first place, with a degree in engineering from the École. After completing his military service, he had accepted a research scholarship and, on the recommendation of Langevin, had been hired in October 1925 as Marie Curie's assistant.

Joliot simultaneously pursued new studies to obtain his *licence ès sciences* in 1927, his teaching activities at the École d'Électricité Industrielle Charliat in order to augment his finances, and the learning of laboratory techniques under the guidance of Irène Curie. His first work, which dealt with dilute solutions of radioactive salts, was used in his doctoral thesis, presented in 1930, on the electrochemical properties of radioactive elements.

Joliot had seen beneath the rough and cold exterior of Irène Curie "an extraordinarily sensitive and poetic creature" in whom he was to find "the purity, common sense and tranquility of Pierre Curie." They were married on October 9, 1926. Their daughter Hélène was born in 1927 and their son Pierre in 1932. Beginning in 1928 they signed their scientific work jointly.

The Joliot-Curies. Irène and Frédéric Joliot-Curie manifested a great interest in the study of atomic nuclei, along the lines of the pioneer work of Ernest Rutherford, who in 1911 and in the 1920s had devised techniques for bombarding nuclei with alpha particles (helium ions). They were aided in their research by the intensity of the radioactive sources at their disposal. One of the outstanding achievements of Marie Curie was to understand the need to accumulate intense radioactive sources, not only for the treatment of illness but also for research in nuclear physics. The resulting stockpile was an unrivalled instrument until the appearance after 1930 of particle accelerators. The existence at the Radium Institute of 1.5 grams of radium in which, over a period of several years, radium D and polonium had accumulated was a decisive contribution to the success of the experiments undertaken in the years around 1930. Frédéric and Irène were highly skilled in the use of radiation detectors. Frédéric used the ionization chamber. But the Wilson cloud chamber, which Irène had taught him to use, was his favourite instrument; it permitted him to obtain the first photograph, in 1932, of the simultaneous production of an electron–positron pair.

Their contribution to the discovery of the neutron was also decisive. Having studied in 1931 the strongly penetrating radiation described in 1930 by W.W. Bothe and H. Becker, the Joliot-Curies published on January 18,

1932, the surprising result that such a radiation is able to project heavy nuclei and is not composed of photons, the discrete particles of which light is composed. On February 27, 1932, Sir James Chadwick, having reviewed the experiments in the Cavendish Laboratory, interpreted the particles, in the journal *Nature*, as the neutrons postulated as early as 1923 by Rutherford.

In 1934, the year Frédéric Joliot-Curie was appointed lecturer at the Sorbonne, Frédéric and Irène directed an experiment in which a thin sheet of aluminum was bombarded with alpha particles from an intense source of polonium; neutrons and positrons were emitted, the latter continuing to appear after the withdrawal of the source of polonium at a rate that decreased exponentially with time. Radioactive phosphorus was thus formed. In the same way, radioactive nitrogen could be formed by bombarding boron, and radioactive silicon was formed when magnesium was the target. The radioactive elements were in fact produced also with particle accelerators in other laboratories, but their artificial formation was discovered in Marie Curie's laboratory. Thus, Marie Curie had the pleasure of seeing the first sample of artificial radioactive material a few months before she died of leukemia caused by radiation on July 4, 1934, near Sallanches.

Artificial radioactivity

Radioactive isotopes were developed, and their use soon became widespread; the Joliot-Curies saw their applications in medicine and biology. In 1935 they were awarded the Nobel Prize for Chemistry for the synthesis of new radioactive elements. The Joliot-Curies then moved into a home at the edge of the Parc de Sceaux. They left it only for visits to their house in Brittany at Pointe de l'Arcouest, where university families had been meeting together since the time of Marie Curie. And, for the sake of Irène's lungs, they visited the mountains of Courchevel during the 1950s.

Frédéric, appointed professor at the Collège de France in 1937, devoted a part of his activities to preparing new sources of radiation. He then supervised the construction of electrostatic accelerators at Arcueil–Cachan and at Ivry and a cyclotron of seven million electron volts at the Collège de France, the second (after the Soviet Union) installation of equipment of this type in Europe.

Irène then devoted her time largely to the upbringing of Hélène and Pierre. But both she and Frédéric had a lofty idea of their human and social responsibilities. They had joined the Socialist Party in 1934 and the Comité de Vigilance des Intellectuels Antifascistes (Vigilance Committee of Anti-Fascist Intellectuals) in 1935. They also took a stand in 1936 on the side of Republican Spain. Marie Curie had already favoured granting to women the right to vote. Irène was also convinced that women did not have the place in society they deserved. Considering it her duty, therefore, to accept the high offices that were offered her, she was one of three women to participate in the Popular Front government of 1936. As undersecretary of state for scientific research, she helped to lay the foundations, with Jean Perrin, for what would later become the Centre National de la Recherche Scientifique (National Centre for Scientific Research). She appeared two times in vain before the Académie des Sciences to affirm the right of women to become members.

Political activities

Nuclear fission. Appointed professor at the Sorbonne in 1937, Irène continued her work on radioactive elements produced with neutrons. In 1938, along with Pavle Savić, she demonstrated on a target of uranium the transitory appearance of an element analogous to lanthanum. The result was interpreted in December 1938 by Otto Hahn and Fritz Strassmann, who conceived the process of nuclear fission. Frédéric immediately undertook to demonstrate fission by physical means. He succeeded in this as early as January 1939 while Otto Robert Frisch was independently obtaining the same result. Then came experiments carried on by Frédéric Joliot-Curie, with Hans von Halban, Lew Kowarski, and Francis Perrin, to prove that many neutrons are produced during fission, that the possibility exists of developing nuclear reactions in explosive chains, and that such nuclear reactions can be controlled in order to release great quantities of energy. These discoveries were then protected by five patents, which Joliot-Curie and his collaborators turned over to the Centre National de la Recherche Scientifique.

The war years. Pierre and Marie Curie had decided to publish everything. This was also the attitude adopted by the Joliot-Curies for the discovery of artificial radioactive elements. But anxiety resulting from the rise of Nazism and the awareness of the dangers that could result from the application of chain reactions led them to cease publication. On October 30, 1939, they recorded the principle of nuclear reactors in a sealed envelope, which they deposited at the Académie des Sciences; it remained secret until 1949. Joliot-Curie became the director of a group of scientific researchers at the beginning of the war in 1939; continuing the work, he convinced Raoul Dautry, minister of armaments, to purchase in Norway the world stockpile of 130 litres (140 quarts) of heavy water. Then, in June 1940 Joliot-Curie sent the stockpile of heavy water to England; the stockpile of uranium entrusted personally to him by the Union Minière du Haut Katanga was hidden near Toulouse.

Uranium and heavy water

Joliot-Curie chose to remain in occupied France with his family and to make certain that the Germans who came into his laboratory could not use his work or his equipment, whose removal to Germany he prevented. The Joliot-Curies continued their research, notably in biology; after 1939, Frédéric demonstrated, with Antoine Lacassagne, the use of radioactive iodine as a tracer in the thyroid gland. He became a member of the Académie de Médecine in 1943.

But the struggle against the occupying forces began to require more and more of his attention. In November of 1940 he denounced the imprisonment of the French scientist Paul Langevin. In June of 1941 he took part in the founding of the National Front Committee, of which he became the president. In the spring of 1942, after the execution by the Nazis of the theoretical physicist J. Solomon, Joliot-Curie joined the French Communist Party, of which in 1956 he became a member of the central committee. He created the Société d'Études des Applications des Radio-éléments Artificiels, an industrial company that gave work certificates to scientists and thus prevented their being sent to Germany. In May 1944, Irène and their children took refuge in Switzerland and Joliot-Curie lived in Paris under the name of Jean-Pierre Gaumont. His laboratory at the Collège de France, at which he organized the production of explosives, served as an arsenal during the battle for the liberation of Paris. In recognition, he was designated a commander of the Legion of Honour with a military title and was decorated with the Croix de Guerre.

Applications of atomic energy. In France, after the liberation in 1944, Joliot was elected to the Académie des Sciences and was entrusted with the position of director of the Centre National de la Recherche Scientifique. He helped to establish with G. Teissier a system that created for the centre a structure of democratic management and assigned to it the mission of coordinating research on a national scale.

Then, in 1945 General de Gaulle authorized Joliot and Dautry to create the establishment of the Commissariat à l'Energie Atomique to ensure for France the applications of the discoveries made in 1939. Irène Curie devoted her scientific experience and her abilities as an administrator to the acquisition of raw materials, the prospecting for uranium, and the construction of detection installations. In 1946 she was also appointed director of the Radium Institute. Frédéric's responsibilities also took him twice in 1946 to the United Nations. In spite of the difficulties encountered in France, where industry was only beginning to be reconstructed, his efforts culminated in the deployment, on December 15, 1948, of ZOE (*zéro, oxyde d'uranium, eau lourde*), the first French nuclear reactor, which, though only moderately powerful, marked the end of the Anglo-Saxon monopoly. Already Joliot-Curie had chosen at Saclay, near Versailles, the location for a large research centre envisaged as the

French nuclear reactor

next step in the development of the Commissariat à l'Energie Atomique. In 1949 the first weighable sample of plutonium produced in France was presented to Joliot, 15 years after the discovery of artificial radioactivity. But in April 1950, in the climax of the cold war, Prime Minister Georges Bidault removed him without explanation from his position as high commissioner, and a few months later Irène was also deprived of her position as commissioner in the Commissariat à l'Energie Atomique. They devoted themselves henceforth to their own laboratory work and to teaching.

The danger that radioactivity, "in criminal hands," according to the expression used by Pierre Curie in his Nobel speech in 1903, could create a burden for the human race had been felt by the Curies, but for them it was a distant possibility. For the Joliot-Curies, it was a matter of conscience, which Frédéric had already expressed in his Nobel speech of 1935. They were committed to the struggle for the free circulation of ideas and discoveries, for using atomic energy in the service of peace, and for checking the nuclear arms race and the testing of atomic and hydrogen bombs. In 1946 Frédéric renewed contacts with his British colleagues to found the World Federation of Scientific Workers, became even more involved as president and founder of the World Peace Movement in 1949, and launched the "Stockholm Appeal" against nuclear armaments in 1950. In 1955, along with others, he helped to initiate the Pugwash conferences, named for the town in Nova Scotia where scientists from around the world first met to discuss international politics.

During the 1950s, following several operations, Irène's health began to decline. In May of 1953 Frédéric had the first attack of hepatitis from which he was to suffer for five years, with a serious relapse in 1955. In 1955 Irène drew up plans for the new nuclear physics laboratories at the Université d'Orsay, south of Paris, where teams of scientists could work with large particle accelerators under conditions less cramped than in the Parisian laboratories. Early in 1956 Irène was sent into the mountains, but her condition did not improve. Wasted away by leukemia as her mother had been, she again entered the Curie Hospital, where she died on March 17, 1956.

Ill, knowing that his days were also numbered, Frédéric decided to carry on Irène's unfinished work. In September 1956 he accepted the position of professor at the University of Paris left vacant by Irène, at the same time occupying his own chair at the Collège de France. He also successfully completed the establishment of the Orsay laboratories and saw the start of research there in 1958. Making one final gesture for peace, he took the chair at the inauguration in July of the International Congress of Nuclear Physics, the first one organized in France since World War II. He died on August 14, 1958, at Paris.

BIBLIOGRAPHY. The basic biographies are MARIE CURIE, *Pierre Curie* (Eng. trans. 1923); and EVE CURIE, *Madame Curie* (Eng. trans. 1937), both eyewitness accounts. EUGENIE COTTON, *Les Curies* (1963), contains recollections of the four Curies and an assessment of their scientific accomplishments. The author was a student of Marie Curie at the École Normale Supérieure at Sèvres, and remained a friend of the family for the rest of her life; she met Irène, as a child of five; and, toward the close of her life, she was associated with Frédéric Joliot in the Movement for Peace. See also PIERRE BIQUARD, *Joliot-Curie* (1961; Eng. trans., 1965), for an account of the friendship between the author and Joliot, which began when they were students at the École de Physique et Chimie; Biquard was Joliot's principal secretary at the French Atomic Energy Commission, and was closely associated with his activities for peace. A scientific evaluation is given in the address by FRANCIS PERRIN on the "Oeuvre scientifique de Marie Skłodowska-Curie et son influence sur les grandes conquêtes de la physique moderne" in *Centenary Lectures*, given in Warsaw 1967 (1968).

(F.Ne.)

Curzon, Lord

Viceroy of India from 1898 to 1905 and foreign secretary from 1919 to 1924, George Nathaniel Curzon, 1st Marquess Curzon of Kedleston, was the last English statesman to belong in spirit to the 18th-century aristocracy.



Lord Curzon.
Radio Times Hulton Picture Library

Though gifted with brilliant talents, Curzon possessed an aloofness and aristocratic disdain that belonged to the politics of a previous century and prevented him from achieving the premiership he felt he so richly deserved.

Curzon was born at Kedleston Hall on January 11, 1859, eldest son of the 4th Baron Scarsdale, rector of Kedleston, Derbyshire. His early development was strongly influenced by the benign neglect of his parents and the dominating character of his governess (whom he termed "a brutal and vindictive tyrant") and of his first preparatory schoolmaster (a firm believer in corporal punishment). At Eton, where he proved a wayward and emotional pupil, he clashed with his tutors but developed an extraordinary gift for assimilating the contents of books; by studying hard in private, he surprised everyone by winning more prizes (for French, Italian, and history, among other subjects) than had ever been carried off before.

Early life

Just before entering Oxford in 1878, he was struck down by a devastating pain in his back, the aftermath of a riding accident of four years previous. He refused to accept medical advice to rest and instead donned a leather harness, which he wore for the rest of his life. The back pain was to plague him from that time on, robbing him of sleep, forcing him to take drugs, and often making him querulous and unbalanced at some of the most vital moments in his career and in the affairs of the British Empire. It should be added that the pain sharpened his mind and never kept him from achieving remarkable feats of physical and mental endurance.

Curzon was elected president of the Oxford Union in 1880 and made a fellow of All Souls College in 1883. He had a gift for making friends in high places, and this was apt to be resented by his contemporaries. About this time a verse was circulated at Oxford of which he was to write later: "Never has more harm been done to one single individual than that accursed doggerel has done to me." It went as follows:

My name is George Nathaniel Curzon,
I am a most superior person,
My cheek is pink, my hair is sleek,
I dine at Blenheim once a week.

(Blenheim is the residence of the dukes of Marlborough.) Two years later he was dining even more frequently at Hatfield House, ancestral home of Lord Salisbury, Conservative leader in the House of Lords, for whom he was now doing research and drafting speeches. His reward was Salisbury's recommendation of Curzon to the Tories of Southport, Lancashire, who agreed to adopt him as their candidate at the next election. It was a safe Tory seat, and in 1886 Curzon became a member of Parliament for the first time. With Salisbury's approval he neglected his parliamentary duties to embark on a world tour and came back infatuated with Asia. From this and sub-

sequent journeys emerged three books: *Russia in Central Asia* (1889); *Persia and the Persian Question* (1892), by far the most successful of his works; and *Problems of the Far East* (1894).

Rise to political eminence. On November 10, 1891, Curzon took his first step up the political ladder by accepting Salisbury's invitation to become undersecretary of state for India in the Tory government. The financial worries that beset him at the time (for he had developed extravagant tastes) were solved when he married Mary Victoria Leiter, daughter of Adolphus (Levi) Leiter, a Chicago millionaire. The marriage took place in Washington, D.C., on April 22, 1895, and the union involved marriage settlements of several millions of dollars. There was also a present from Lord Salisbury: the newly married couple returned from their honeymoon to find him waiting with an offer to Curzon of the job of undersecretary of state, Salisbury having just been appointed foreign secretary. Curzon accepted on the condition that he was also to be made a privy councillor, and on June 29, 1895, he was duly sworn in by Queen Victoria at Windsor Castle. From this moment his rise to political eminence was swift.

Viceroy of India

In 1898 it was announced that he would succeed Lord Elgin as viceroy of India, and, in September of that year, he was created Baron Curzon of Kedleston. He was the youngest viceroy of India in history, and he cherished the prospect of it, for it was an office filled with the pomp and ceremony he loved. India was the most treasured jewel in Queen Victoria's crown, and, after the pageantry of his arrival in Calcutta on January 3, 1899, Curzon wrote: "I suddenly saw what had come into my hands, and what prodigies of energy and inspiration would be needed on my part to guide them." He demanded obedience and a bent knee from the rajas, maharajas, and provincial governors he now ruled, but there was no doubt of his sense of mission toward the Indian people. He initiated commissions of inquiry into education, police, and civil services; he reduced taxes; he ordered immediate punishment of any Briton (including members of the army) who ill-treated Indian nationals. In external affairs he paid special attention to India's frontiers, toured the Persian Gulf, and sent a successful mission to Tibet to frustrate Russian ambitions there. On his return from what he called a "triumphal" tour of the Indian provinces, he ordered the restoration of the Tāj Mahal, which was decaying, and thereafter took a personal interest in India's artistic and cultural heritage.

At the end of his first five years in India, his successes were recognized by the government at home by a renewal of his term; but, in fact, the period of blazing glory was over, and now came political tragedy. At Curzon's personal request, the job of commander in chief of the Indian army and military member of the viceroy's Cabinet had been given to England's military hero of the day, Lord Kitchener of Khartoum. Curzon believed that to have such an illustrious soldier on his staff would embellish his own image, though his friends in England warned him repeatedly that Kitchener was, in Lord Esher's words, "an uncouth and ruthless man." It was a clash of personalities, and the two were soon unscrupulously intriguing against each other. A final confrontation between the two men, which ended in Curzon's bursting into tears, brought matters to a climax. Curzon, confident that the government would take his part, cabled that either his views must be accepted or he would go. On the morning of August 16, 1905, he received a cable from King Edward VII telling him that his resignation had been accepted. He delayed his return to England; and, by the time he was once again in London, the Tories were out of office, and his Indian achievements had been forgotten. He was not even given the earldom usually awarded to retiring viceroys.

Temporary retirement

In the period of political eclipse that followed, he became an excellent and enlightened chancellor of the University of Oxford and filled many other important offices. But his temporary retirement from politics was marred by the death of his beloved wife, Mary. Her death affected him deeply, but the money that now came to him enabled him to indulge in his passion for the collection of art treasures and old buildings. In 1911 he bought his first castle, Tattershall, in Lincolnshire, which he restored; and later he did the same to Bodiam Castle, Sussex, eventually presenting both of them to the nation.

His political ambitions had been damped down but never extinguished, and his hopes were stirred anew in 1911. In that year, after the coronation of King George V, Curzon received an earldom, along with the viscountcy of Scarsdale and the barony of Ravensdale. He showed his gratitude to the Tories who had elevated him by persuading his fellow peers (against his own and their feelings) to abstain from voting against the Parliament Bill, which curtailed their powers, thus avoiding a constitutional crisis the government had feared. He joined the coalition Cabinet of H.H. Asquith in the summer of 1915, and, when Lloyd George took over that December, he became leader of the House of Lords with the office of lord president. From then on Curzon was one of the members of the inner Cabinet concerned with the policies and pursuits of World War I.

There had been a time when all fashionable London imagined that Curzon would marry the flamboyant redheaded novelist Elinor Glyn, but to everyone's surprise—not least that of Miss Glyn—he announced his engagement in December 1916 to Mrs. Alfred (Grace) Duggan, widow of a rich Argentinian rancher and daughter of J. Monroe Hinds, an American diplomat. They were married on January 2, 1917. His first wife had given Curzon three daughters. He hoped that his second would produce the son to inherit his title, and for both of them the years ahead were filled with hopes and disappointments.

Disappointments. There were disappointments in politics, too. Curzon had decided that the one lesson he must learn from his bitter experience in India was: never to resign. But in his case, it was a poor one. In the postwar government led by Lloyd George, he was appointed foreign secretary, a position for which he was eminently fitted. But time and again he was overruled or pushed aside by his boisterous leader, and his carefully planned policies thwarted. It was a time when resignation might well have gained him the overwhelming support of the Tories (who despised the Liberal coalition leader, Lloyd George) and taken him to the top. Instead he clung to office, and it was not until Tories took over in 1922 that he came into possession of the full powers of his office. He served with distinction until 1923, painstakingly dealing with the chaotic problems of postwar Europe and the Near East. When the Tory prime minister Bonar Law, a dying man, prepared to relinquish office, Curzon had good reason to believe that his efforts would be rewarded by the premiership that he felt he so richly deserved. It was not to be. Backstairs political intrigue (and the fear that a premier in the House of Lords would be "out of touch") resulted in the appointment of a House of Commons man, Stanley Baldwin, as prime minister.

It was a bitter blow to Curzon's hopes, and he might have been forgiven if he washed his hands of politics there and then and retired to his academic and artistic pursuits. Instead, he insisted on presiding at the meeting at which Baldwin was elected to the job for which he was so much better equipped. He hung on to his job as foreign secretary until 1924, when Baldwin replaced him with Austen Chamberlain, yet even then he did not resign but accepted the emasculating political role of lord president of the council. He kept silent in public but confided his grief and humiliation to his private diaries and notes.

He had been created a marquess in 1921, and more than ever he hoped for a son to inherit his title, but in this, too, he was to be frustrated. On March 9, 1925, he was operated on for an internal condition and, for the next eight days, seemed to be recovering. But then complications set in, and his doctor advised his wife, Grace, to tell him the truth, that he was dying. He took the news philosophically, repeating the Lord's Prayer; and then Curzon, who all his life had been a despairing insomniac, racked by the pain in his back, suddenly discovered the knack of it and went off into a deep and peaceful sleep. It was in his sleep that he died on March 20, 1925.

Death

With him died his marquessate and his earldom. The viscountcy subsequently passed to his nephew and the barony of Ravensdale to his eldest daughter, Lady Irene Curzon.

BIBLIOGRAPHY. LORD RONALDSHAY, *The Life of Lord Curzon*, 3 vol. (1928), is the official biography. LEONARD O. MOSLEY, *Curzon: The End of an Epoch* (U.S. title, *The Glorious Fault*, 1960), made use of the collected Curzon Papers for the first time. HAROLD G. NICOLSON, *Curzon: The Last Phase, 1919–1925* (1934), is an acute and sympathetic study of Curzon during one of the post-World War I conferences. PHILIP M. MAGNUS, *Kitchener: Portrait of an Imperialist* (1959, reprinted 1968), gives a good account of the clash between Curzon and his Military Member during his Indian viceroyalty.

(Le.M.)

Customs Unions and Trade Agreements

Many efforts have been made in modern times to promote trade among nations. The ways in which this may be attempted range from agreements among governments to reduce or eliminate trade barriers to more ambitious attempts to harmonize economic policies, as in the European Economic Community or the Council for Mutual Economic Assistance.

The article is divided into the following sections:

I. Trade agreements
- Bilateral trade agreements
- Modern commercial policies
- Multilateral agreements since World War II

II. Economic integration
- Integration between the states of federations
- The economic integration of colonial empires
- The Zollverein
- The Benelux economic union
- The European Coal and Steel Community
- The European Economic Community
- The European Free Trade Association
- Economic Integration in Latin America
- The Council for Mutual Economic Assistance
- An appraisal of the European Economic Community

I. Trade agreements

The term "trade agreement" or "commercial agreement" can be used to describe any contractual arrangement between states concerning their trade relationships. Trade agreements may be bilateral or multilateral—that is, between two states or between more than two states.

BILATERAL TRADE AGREEMENTS

The making of trade agreements

A bilateral trade agreement usually includes a broad range of provisions regulating the conditions of trade between the contracting parties. These include stipulations governing customs duties and other levies on imports and exports, commercial and fiscal regulations, transit arrangements for merchandise, customs valuation bases, administrative formalities, quotas, and various legal provisions. Most bilateral trade agreements, either explicitly or implicitly, provide for (1) reciprocity, (2) most-favoured-nation treatment, and (3) "national treatment" of nontariff restrictions on trade.

Reciprocity. In a trade agreement, the parties make reciprocal concessions to put their trade relationships on a basis deemed equitable by each. The principle of reciprocity is extremely old, and in one form or another is to be found, implicitly at least, in all trade agreements. The concessions may, however, be in different areas. In the Anglo-French Agreement of 1860, for example, France pledged itself to reduce its duties to 20 percent by 1864. In return, Britain granted duty-free imports of all French products except wines and spirits. The principle of reciprocity implies only that the gains arising out of foreign trade are distributed fairly.

The most-favoured-nation clause. The most-favoured-nation clause binds a country to apply to its partner country any lower rate of import duties that it may later grant to imports from some other country, in place of the rate specified in the agreement. The clause may cover a list of specified products only, or specific concessions yielded to certain foreign countries. Alternatively, it may cover all advantages, privileges, immunities, or other favourable treatment granted to any third country whatever. The clause is intended to provide each signatory with the assurance that the advantages obtained will not be attenuated or wiped out by a subsequent agreement concluded between one of the partners and a third country. It also guarantees the parties against discriminatory treatment in favour of a competitor.

Provisions against discrimination

The effect of the most-favoured-nation clause on customs duties is to amalgamate the schedules of duties annexed to the successive trade agreements concluded by a state. If the rates in different agreements are fixed at varying levels, the clause reduces them to the lowest rate specified in any agreement. Thus goods imported from a country benefitting from most-favoured-nation treatment are charged the rate of duty applicable to imports from another country which, in a subsequent trade agreement, has negotiated a lower rate of duty.

The coverage of the most-favoured-nation clause can be considerably reduced by a minute definition of a particular item so that a concession, while general in form, applies in practice only to one country. The best-known illustration of this technique is to be found in the German Tariff of 1902, which admitted at a special rate

> large dappled mountain cattle, reared at a spot at least 300 metres above sea level, and which have at least one month's grazing each year at a spot at least 800 metres above sea level.

The advantages granted under the most-favoured-nation clause may be conditional or unconditional. If unconditional, the clause operates automatically whenever appropriate circumstances arise. The country drawing benefit from it is not called on to make any fresh concession. By contrast, the partner invoking a conditional most-favoured-nation clause must make concessions equivalent to those extended by the third country. A typical wording was that of the 1911 treaty between the United States and Japan, which stated that

> in all that concerns commerce and navigation, any privilege, favour or immunity . . . to the citizens or subjects of any other State shall be extended to the citizens or subjects of the other Contracting Party gratuitously, if the concession in favour of that other State shall have been gratuitous, and on the same or equivalent conditions, if the concession shall have been conditional.

The conditional form of the clause may at first sight seem more equitable. But it has the major drawback of being liable to raise dispute each time it is invoked, for it is by no means easy for a country to evaluate the compensation it is being offered as in fact being equivalent to the concession made by the third country.

The effect of the unconditional form of the most-favoured-nation clause is, finally, to wipe out any relevance the principle of reciprocity may have had to the purely bilateral preoccupations of the negotiating parties, since the results of the bargaining process, instead of being limited to the participants, influence their relationships with other states. In practice, therefore, a country negotiating a trade agreement must measure the advantages it is willing to concede in terms of the benefits these concessions will provide collaterally to that third country which is the most competitive. In other words, the concessions that may be granted are determined by the minimum protection that the negotiating state deems indispensable to protect its home producers. This sets a major limitation on the scope of bilateral negotiations.

Protagonists of free trade consider that the unconditional most-favoured-nation clause is the only practical way by which to obtain the progressive reduction of customs duties. Apologists for protectionism are resolutely against it, preferring the conditional form of the clause or some equivalent mechanism.

The conditional most-favoured-nation clause was generally in use in Europe until 1860, when the so-called Cobden–Chevalier Treaty between Great Britain and France established the unconditional form as the pattern for most European treaties. The United States used the conditional most-favoured-nation clause from its first trade agreement, signed with France in 1778, until the passage of the Tariff Act of 1922, which terminated the practice.

Decline of the most-favoured-nation clause. The Genoa conference in May 1922 and the International Economic Conference in May 1927 both recommended that trade agreements should include the most-favoured-nation clause whenever possible. But the depression of the 1930s led instead to an upsurge of restrictions in world trade. Imperial or regional systems of preference came into being: the Ottawa Agrement of 1932 for the British Commonwealth; similar arrangements for the French empire; and a series of tariff and preference agreements negotiated in eastern and central Europe from 1931 on.

Henceforth, a major objection to the clause in its unconditional form was that it appeared to hinder the creation of preferential systems and customs unions. Paradoxically enough, a clause designed to promote the lowering of tariff barriers was found to have the opposite effect; those in favour of lowering or abolishing customs duties within regional groupings of countries therefore opposed it, and their opposition was an important factor when postwar arrangements were being worked out in 1947.

The "national treatment" clause. The "national treatment" clause in trade agreements is designed to ensure that internal fiscal or administrative regulations are not used to introduce discrimination of a nontariff nature. It forbids discriminatory use of the following: taxes or other internal levies; laws, regulations, and decrees affecting the sale, offer for sale, purchase, transport, distribution, or use of products on the domestic market; valuation of products for purposes of assessment of duty; legislation on prices of imported goods; warehousing and transit regulations; and the organization and operation of state trading corporations.

MODERN COMMERCIAL POLICIES

Mercantilism. Much of the modern history of international relations concerns efforts to promote freer trade among nations. The 17th century saw the growth of restrictive policies that later came to be known as "mercantilist." The mercantilists held that economic policy should be nationalistic and aim to secure the wealth and power of the state. The concept was based on the conviction that national interests are inevitably in conflict—that one nation can increase its trade only at the expense of other nations. Thus, governments were led to impose price and wage controls, foster national industries, promote exports of finished goods and imports of raw materials, and prohibit the exports of raw materials and the import of finished goods. The state endeavoured to provide its citizens with a monopoly of the resources and trade outlets of its colonies.

A typical illustration of the mercantilist spirit is the famous English Navigation Act of 1651, which reserved for the home country the right to trade with the colonies and prohibited the import of goods of non-European origin unless transported in ships flying the English flag. This law lingered on until 1849. A similar policy was followed in France.

The growth of free trade

Liberalism. A strong reaction against mercantilist attitudes began to take shape toward the middle of the 18th century. In France, the economists known as Physiocrats demanded liberty of production and trade. In England, Adam Smith demonstrated in his *Wealth of Nations* (1776) the advantages of importing goods when articles could be obtained more cheaply in this way than by domestic manufacture. Economists and businessmen voiced their opposition to excessively high and often prohibitive customs duties, and urged the negotiation of trade agreements with foreign powers. This change in attitudes led to the signature of a number of agreements embodying the new liberal ideas, among them the Anglo-French Treaty of 1786, which ended what had been a real economic war between the two countries.

After Adam Smith, the basic tenets of mercantilism were no longer considered to be defensible. This did not, however, mean that nations abandoned all mercantilist policies. Restrictive economic policies were now justified by the claim that, up to a certain point, the government should keep foreign merchandise off the domestic market in order to shelter national production from outside competition. To this end, customs levies were introduced in increasing number, replacing outright bans on imports, which became less and less frequent.

In the middle of the 19th century, customs walls effectively sheltered many national economies from outside competition. The French tariff of 1860, for example, charged extremely high rates on British products: 60 percent on pig iron; 40 to 50 percent on machinery; and 600 to 800 percent on woollen blankets. Apart from the duties proper, transport costs between the two countries provided further protection.

A triumph for liberal ideas was the Anglo-French trade agreement of 1860, which provided that French protective duties were to be reduced to a maximum rate of 25 percent within five years, with free entry of all French products except wines into Britain. This agreement was followed by other European trade pacts, usually containing most-favoured-nation clauses.

Resurgence of protectionism. A reaction in favour of protection spread throughout the Western world in the latter part of the 19th century. Germany adopted a systematically protectionist policy and was soon followed by most other nations. Shortly after 1860, during the Civil War, the United States raised its customs sharply; the McKinley Tariff Act introduced in 1890 was ultra-protectionist. England was the only country to remain faithful to the principles of free trade.

The decline of free trade

But the protectionism of the last quarter of the 19th century was mild by comparison with the mercantilist policies that had been common in the 17th century and were to be revived between the two world wars. It is easy today to ignore how much economic liberty prevailed in 1913. There were no political barriers to international trade. Quantitative restrictions were unheard of; customs duties were low and stable. Currencies were freely convertible into gold, which in effect was a common international money. Balance of payments problems were nonexistent. People who wished to settle and work in a country could go where they wished almost without restriction; they could open businesses, enter trade, or export capital freely. Equal opportunity to compete was the general rule, the sole exception being the existence of limited customs preferences between certain countries, most usually between a home country and its colonies. Trade was freer throughout the Western world in 1913 than it was in Europe in 1970.

The new mercantilism. World War I wrought havoc with these orderly trading conditions. By the end of hostilities, world trade was in a straitjacket that made recovery very difficult. The first five years of the postwar period were marked by the dismantling of wartime controls proper. The 1920 crisis and the commercial advantages accruing to countries whose currencies had depreciated, as had Germany's, rapidly led to fresh measures in restraint of trade. The protectionist tide engulfed the world economy, not because policy makers consciously adhered to any specific theory but because of nationalist ideologies and the pressure of economic conditions. In an attempt to end the continual raising of customs barriers, the League of Nations organized the first World Economic Conference in May 1927. Twenty-nine states, including the main industrial countries, subscribed to an international convention that was the most minutely detailed and balanced multilateral trade agreement ever approved until then. It was a precursor of the arrangements made under the General Agreement on Tariffs and Trade in 1947.

The 1927 agreement remained practically without effect. During the Great Depression of the 1930s, unemployment in major countries reached unprecedented levels and engendered an epidemic of protectionist measures. Countries attempted to shore up their balance of payments by raising their customs duties and introducing a range of import quotas or even import prohibitions, accompanied by exchange controls.

From 1933 onward, the recommendations of all the postwar economic conferences based on the fundamental postulates of economic liberalism were ignored. The

planning of foreign trade came to be considered a normal function of the state. Mercantilist policies dominated the world scene until after World War II.

MULTILATERAL AGREEMENTS SINCE WORLD WAR II

Attempts at trade liberalization

When World War II ended, the lessons learned from the growth of protectionism since 1871, and most of all from the resurgence of trade restrictions in the interwar years, spurred the development of multilateral trade agreements and other forms of international economic cooperation. These developments culminated in the General Agreement on Tariffs and Trade (GATT).

The General Agreement on Tariffs and Trade. The General Agreement on Tariffs and Trade was signed at Geneva on October 30, 1947, by 23 countries, which among them accounted for four-fifths of world trade. On the same day ten of them, including the United States, the United Kingdom, France, Belgium, and The Netherlands, signed a protocol bringing the agreement into force on January 1, 1948.

The GATT takes the form of a multilateral trade agreement setting forth the principles under which the signatories, on a basis of "reciprocity and mutual advantage," shall negotiate "a substantial reduction in customs tariffs and other impediments to trade, and the elimination of discriminatory practices in international trade." With the adherence of additional countries (there were 80 signatories as of November 1971), GATT has become a charter governing almost all world trade except for that of the Communist countries.

The main principles underlying GATT are as follows: (1) There shall be no trade discrimination of any kind. The most-favoured-nation clause is regarded as fundamental. (2) As a rule, there is to be no protection other than that provided by the customs tariff (the "national treatment" principle). (3) Customs unions and free trade groupings are considered legitimate means of trade liberalization, provided that, *taken as a whole*, such arrangements do not discriminate against third countries. (4) Members of GATT are entitled to levy the following charges on imports: (*a*) an import tax equal in amount to internal taxes on the product concerned, subject only to the general principle embodied in (2) above; (*b*) "antidumping" duties in the case of imported products that are being sold at a loss or are benefitting from an export subsidy; (*c*) fees and other proper charges for services rendered.

These, however, are only the basic principles. The agreement also contains a variety of clauses providing exceptions from the rules in special situations. These include balance of payments disequilibrium; serious and unexpected damage to domestic production; the requirements of economic development or, subject to very broad reservations, of agricultural policy; the need to protect domestic raw material production; and the interests of national security. In the overall prosperity that reigned during the first 20 years of GATT's existence, members needed to make only limited use of these escape clauses. Were there to be a general economic downturn, however, all of GATT's accomplishments could be endangered.

Under GATT a series of conferences have been organized for the purpose of lowering trade barriers: at Geneva (1947), Annecy (1949), Torquay (1950–51), Geneva (1956, 1960–61, and 1964–67). The agreement provides an indispensable framework for simultaneous negotiation of bilateral tariff reductions, which thus take multilateral effect.

The formula for multilateral tariff bargaining applied in negotiations held under GATT auspices is a major innovation in intergovernmental cooperation. In appraising the concessions that they could afford to make, governments have been able to take account of the indirect advantages that they could expect to accrue to them from the full set of bilateral negotiations. Over the years since its inception, it has been successful far beyond its creators' expectations. It has made a major contribution to the growth of world trade.

The Organization for Economic Co-operation and Development. On April 16, 1948, sixteen European countries responded to a United States offer of economic aid under the European Recovery Program by setting up the Organization for European Economic Co-operation. Although the immediate aim was to coordinate the distribution of U.S. credits, the Organization for European Economic Co-operation convention was also designed to foster free trade between the members and allow their participation in customs unions or similar institutions. The members by 1955 consisted of Britain, France, West Germany, Italy, Spain, the Benelux countries, Austria, Denmark, Sweden, Norway, Switzerland, Portugal, Greece, Ireland, Turkey, and Iceland.

The stimulus of the European Recovery Program

The Organization for European Economic Co-operation did much to facilitate the recovery of intra-European trade, and particularly to abolish most of the quantitative restrictions on imports within the area. On September 30, 1961, it was converted into a new institution, the Organization for Economic Co-operation and Development (OECD), and its membership expanded to include the United States and Canada. Japan became a member in 1964, Finland joined in 1969, and Australia joined in 1971.

The three fundamental aims of the OECD are to promote the economic growth of member countries in conditions of financial stability, to contribute to the economic growth of less developed countries, and to foster the growth of world trade on a multilateral, nondiscriminatory basis. The organization proved to be a most useful vehicle for joint action on the many economic problems of the 1960s, particularly in the liberalization of economic relationships. Although its 23 members comprise only one-fifth of the world's population, they account for two-thirds of world trade.

Tariff negotiations

The Kennedy Round. As the economic integration of western Europe progressed, opinion in the United States became concerned at the prospect of remaining outside. Pres. John F. Kennedy pursued the goal of an Atlantic partnership, and secured special negotiating powers under the Trade Expansion Act of 1962. The act authorized tariff reductions of up to 50 percent, subject to reciprocal concessions from the European partners. This marked a fundamental shift away from the traditional protectionist posture of the United States, and led to the so-called Kennedy Round negotiations in the GATT, held in Geneva from May 1964 to June 1967.

The Kennedy Round negotiations concerned four types of problems: (1) progressive reduction, to amount finally to 50 percent, in the duties on all but a few products, in place of the item-by-item bargaining that had prevailed in earlier GATT conferences; (2) inclusion of agricultural as well as industrial products in the scope of the negotiations; (3) discussion of nontariff obstacles as well as of customs duties; and (4) nonreciprocity for economically less developed countries. Fifty-four countries participated in the negotiations, which covered 400,000 tariff headings.

The final result was an average reduction of 35 percent in the duties levied on industrial products, to be implemented over a five-year period. This was less than the 50 percent originally envisaged. Further, the reductions were not geographically uniform: United States, European, and Japanese duties were to fall by an average of 35 percent, British scales by 38 percent, and Canadian by 24 percent. Little change was made in steel and textile tariffs, since the participants felt that reductions in those industries would create intolerable political and social tensions in most of the industrial countries. Problems arose with regard to chemicals because of a so-called American Selling Price which was used for appraising the dutiable value of some products (mainly derivates of benzene), based on prices ruling in the U.S. market; in return for a reduction of 50 percent in the rates of duty charged by the U.S., Great Britain and the European countries agreed to lower their scales by 22 percent, with a further 24 percent reduction to become effective upon abrogation of the American Selling Price. Rather less spectacular results were achieved for agricultural commodities. These included the setting of a minimum price for wheat and a weighted reduction of between 15 and 18 percent in the duties charged on other agricultural and

food products. In the area of nontariff barriers to trade the most significant result was the adoption of a uniform antidumping code.

The Kennedy Round all but completed the process of tariff reduction begun two decades earlier by the industrial countries. By the end of the transition period in 1972, it was anticipated, duties on industrial imports would range between 5 and 15 percent. While developing countries drew little immediate advantage from the Kennedy Round negotiations, they were able to obtain the addition of a new part titled "Trade and Development," to the GATT charter, calling for stabilization, as far as possible, of raw material prices; reduction or abolition of customs duties or other restrictions that differentiate unreasonably between products in their primary state and the same products in finished form; and renunciation by the advanced countries of the principle of reciprocity in their relations with less developed countries.

II. Economic integration

Forms of economic integration

The economic integration of several countries or states may take a variety of forms. The term covers preferential tariffs, free trade zones, customs unions, common markets, economic unions, and full economic integration. The parties to a system of preferential tariffs levy lower rates of duty on imports from one another than they do on imports from third countries. For example, Great Britain and its Commonwealth countries operated a system of reciprocal tariff preferences after 1919. In a free trade zone, no duty is levied on imports from members, but different rates of duty may be charged freely by each member on its imports from the rest of the world. The European Free Trade Association is an example. A further stage is the customs union, in which free trade among the members is sheltered behind a unified schedule of customs duties charged on imports from the rest of the world. The 19th-century German Zollverein (see below) was a customs union. A common market is an extension of the customs union concept, the additional feature being that it provides for the free movement of labour and capital among the members; an example was the Benelux common market until it was converted into an economic union in 1959. The term economic union denotes a common market in which the members agree to harmonize their economic policies generally, as is the case with the European Economic Community (often referred to incorrectly as the "Common Market"). Finally, total economic integration implies the pursuit of a common economic policy by the political units involved; examples are the United States of America, or the cantons of the Swiss confederation.

Economic integration may be brought about by the political will of a state powerful enough to impose it, as under the Roman Empire or the European colonial systems of the 19th century, or it may result from freely negotiated agreement between sovereign states.

The attempts at economic integration made after World War II can be appraised only by reviewing them against the background of the long process through which, over the centuries, the nations of the world have progressively achieved economic integration. Thus, for instance, the world's greatest power in the 17th century, France, was divided into a number of provinces separated from one another by various customs barriers involving a multitude of duties, tolls, and prohibitions. Trade regulations and fiscal charges differed from one region to the next; there was not even a single system of weights and measures. Not until after the Revolution did the economic integration of France really get under way.

INTEGRATION BETWEEN THE STATES OF FEDERATIONS

The United States. The economic integration of the United States was not achieved all at once, but as the result of a long process during which the powers of the federal authorities were constantly reinforced. The Constitution empowered the federal government to regulate the conditions of trade with other countries and to set up a single system of duties. It also abolished the right of individual states to maintain separate customs legislation and to issue their own currencies. It authorized the federal government alone to issue currency, and established the principle of free movement of persons, merchandise, and capital between the federated states. But the conflict of interest between North and South was settled only by the American Civil War. Almost two hundred years were to elapse before the economies of the states could be considered as integrated for practical purposes, and even in the 1970s many economic and fiscal disparities still existed among them.

The difficulties faced by the 13 original states should not be underestimated. During the years prior to the adoption of the Constitution there were bitter trade disputes among the states, which imposed tariffs against each other and refused to accept each other's currencies. Everything seemed to justify the words of a contemporary liberal philosopher, Josiah Tucker, Dean of Gloucester (England):

> As to the future grandeur of America, and its being a rising empire under one head, whether republican or monarchical, it is one of the idlest and most visionary notions that ever was conceived even by writers of romance. The mutual antipathies and clashing interests of the Americans, their differences of governments, habitudes, and manners, indicate that they will have no centre of union and no common interest. They never can be united into one compact empire under any species of government whatever; a disunited people till the end of time, suspicious and distrustful of each other, they will be divided and sub-divided into little commonwealths or principalities, according to natural boundaries, by great bays of the sea, and by vast rivers, lakes, and ridges of mountains.

Switzerland. The Swiss example is no less instructive. Although the Helvetic Confederation emerged as a political entity in the 14th century, its economic integration was achieved, only after many vicissitudes, with the constitution of 1848, which established a common currency, set forth the principle of a common protective system for the cantons, and provided for free movement of goods and Swiss citizens throughout the national territory. Swiss economic integration is all the more remarkable in that it comprises peoples who speak four different languages.

THE ECONOMIC INTEGRATION OF COLONIAL EMPIRES

When the great colonial powers of Europe founded their empires from the 16th century onward, they attempted to monopolize trade with the colonies and to turn it to their own profit. This policy involved four main restrictions: (1) The colonies were to trade exclusively with the mother country. (2) They were not to undertake manufacturing; transformation of raw materials into finished goods remained a monopoly right of the mother country. (3) Imports and exports of the colonies were to be carried only in ships flying the mother country's flag. (4) The mother country exempted colonial products from duty, or imposed lower rates.

This system, although progressively attenuated, applied in various forms from the 16th to the 19th century. Based on force, it was to the benefit of the home countries and detrimental to the economic growth of their colonies.

THE ZOLLVEREIN

The German customs union

The best known of all customs unions is the German Zollverein (literally, "customs union"). Even though Napoleon had reduced the number of German states from 300 to 40 at the beginning of the 19th century, those that remained were isolated from each other by their own customs systems. In addition, numerous internal customs barriers hampered trade within each state. At the same time there was no single external tariff, and the German industries that had sprung up during the Napoleonic Wars were being crushed by English competition. These difficulties were at the root of the creation of the Zollverein.

The starting point was Prussia's abolition of all internal duties and its adoption of an external tariff in 1818. In the next few years a number of other German states followed the Prussian example. Bavaria and Württemberg

set up a customs union in 1828, and by 1830 four separate customs unions were in existence. Prussia then sought to break up the local customs unions and attach them to a general customs union, the Zollverein. The coverage of the Zollverein increased until, by 1871, it included all the German states.

In its first phase, from 1834 to 1867, the Zollverein was administered by a central authority, the Customs Congress, in which each state had a single vote. A common tariff, the Prussian Tariff of 1818, shielded the member states from foreign competition, but free trade was the rule internally.

During a second phase, from 1867 to 1871 (following Prussia's victory over Austria at Sadowa), executive power was wielded by a federal council (Bundesrat) composed of governmental delegates, in which decisions were taken by an absolute majority. Prussia was entitled to 17 of the 58 votes, and held the chair of the council. Legislative power lay with a "customs parliament" (Zollparlament) composed of deputies directly elected by popular vote, and, like the council, taking decisions by a majority vote. This arrangement transformed what had been a confederation into a federal state.

After the victory over France and the proclamation of the German empire in 1871, the customs parliament and the federal council were replaced by the parliament and the executive council of the empire. The federal state had become a nation.

The progressive destruction of a tangled maze of regulations, prohibitions, and controls set the stage for the subsequent rapid development of the German economy. Although economic integration occurred before political unification, it would not have been possible had not many difficulties been swept away by irresistible pressure from Prussia with its military victories.

THE BENELUX ECONOMIC UNION

Belgium–Luxembourg Economic Union

In 1921 Luxembourg, a former member of the Zollverein, signed the Convention of Brussels with Belgium creating the Belgium–Luxembourg Economic Union. Since 1921 Belgium and Luxembourg have, then, had the same customs tariff and a single balance of payments.

The union was expanded after World War II to include The Netherlands. At the beginning of 1948 most import duties within the Benelux area were abolished, and a common external tariff was put into operation. Exceptions were made, nevertheless, for a few agricultural products, and it was also felt necessary to introduce a system of quotas.

It was rapidly perceived that a simple customs union was inadequate, and a treaty on October 15, 1949, set as its target the progressive and complete liberalization of trade between the partners, systematic coordination of their international commercial and monetary policies, and the adoption of a joint bargaining position in negotiations with other countries. Though the experiment was optimistically viewed everywhere as the precursor of a wider European economic integration, it faced difficulties arising from the very different postwar situations of Belgium and The Netherlands. The two economies were competitive rather than complementary. Other problems arose in connection with the free access of Dutch agricultural products to the Belgian market. Moreover, the Belgian economic system was more liberal than the Dutch, where rigorous price control had long been a standard practice.

The development of Benelux received strong impetus from the formation of the European Economic Community in the 1950s. The Treaty of Rome in 1957 creating the EEC, or Common Market, spurred the members of Benelux to confirm and strengthen their own integration in the Benelux Treaty of Economic Union signed at The Hague on February 3, 1958. The Hague treaty, however, contained little that was new, and in outline was no more than the codification of results already achieved.

The Benelux Economic Union is an intergovernmental union associating three sovereign states. All of its decisions must be unanimous. The union has executive organs (a committee of ministers, the council of the economic union, a number of commissions, and a general secretariat); consultative organs (the interparliamentary council and the economic and social council); and legal organs (the arbitration tribunal and its registry).

The progress of the European Economic Community in the 1960s attenuated the practical importance of Benelux, which seemed certain to diminish as the integration of western Europe proceeded.

THE EUROPEAN COAL AND STEEL COMMUNITY

Efforts toward further integration in Europe

An important step in European integration was taken in May 1950 when the French foreign minister, Robert Schuman, proposed that a common market for coal and steel be set up by countries willing to delegate powers over these sectors of their economies to an independent authority. The motive behind the plan was the belief that a new economic and political framework was needed if European unity was to be achieved and if the threat of a future Franco-German conflict was to be avoided. In April 1951 France, West Germany, Italy, and the three Benelux countries signed a treaty in Paris setting up the European Coal and Steel Community.

The signatories bound themselves to abolish all customs barriers and other restrictions on the movement of coal and steel between their countries; to renounce all discriminatory practices among producers, purchasers, or users (with respect to price and delivery conditions, transport charges, selection of suppliers, etc.); to end government subsidies or grants-in-aid; and to eliminate all practices interfering with the operation of markets.

The constitution of the community. The Coal and Steel Community was to be governed by a high authority, assisted by a consultative committee, a common assembly, a special council of ministers, and a court of justice.

The high authority was the permanent executive organ of the community. Its decisions, taken by a majority vote, were fully binding on all member countries, each of which was pledged to respect the "supranational character" of the high authority. The authority was to refer important substantive matters to the consultative committee before taking a decision. The latter was composed of coal and steel industry representatives, including producers, workers, users, and traders.

The assembly was empowered to exercise only parliamentary control, but could overrule the high authority by a two-thirds majority. Its delegates were composed of deputies of national parliaments.

The function of the council of ministers was to "harmonise the actions of the high authority and the governments responsible for the economic policy of respective countries." It was composed of representatives of member countries, each of which delegated a member of its government. Most decisions of the council were valid if voted by a majority of representatives of member countries, providing that the majority included at least one from a country accounting for one-sixth or more of the total value of the community's production of coal and steel—France or Germany. Unanimous agreement was required only on decisions concerning production questions and shortages.

Taken as a whole, the treaty was similar to a federal constitution, embodied in a long and complex document. It was inspired by three guiding ideas. The first was the concern to pursue a common policy for basic industries. The second was the belief that once the key steel and coal industries had been merged under one international authority, any major Franco-German conflict would be impossible. The third was the desire to prepare the ground for European political integration.

There is, however, a basic incompatibility between the community's provenance, limited to the coal and steel industries, and the sovereignty of the member countries, each of which is responsible for its own general economic policy. As a practical matter, during the first 17 years of the community's existence, authority on all substantive issues remained vested in the national governments. The high authority was autonomous only in matters of secondary importance. Thus the coal crisis of 1958—when West German, Belgian, and French stocks of unsold coal

rose to unmanageable proportions—was resolved at the national level. All the high authority could do was to confirm the measures taken, even when they were contrary to the provision of the treaty. Similarly, the reduction of the labour force in coal mining from 650,000 persons at the end of 1957 to 300,000 ten years later was effected by measures taken by the individual countries pursuing their own national policies, and there was no community-wide action.

The treaty reserved for member countries responsibility for their own trade policies toward third countries. This hindered the establishment of an effective common market since a common market requires a unified system of protection from foreign competition. At the height of the coal crisis, for example, when stocks of coal rose in Belgium, West Germany, and France, Italy nonetheless continued to buy cheap supplies from the United States.

Later developments. An assessment of the Coal and Steel Community is therefore difficult to make, since the political and economic realities with which it must work are so much more complex than the language of the treaty suggests. Much has been accomplished. The markets for steel and coal have been liberalized to a considerable degree, perhaps partly because of the general movement toward liberalization in the economies of the member countries. The community has been a useful forum in which questions of common interest could be examined. It fostered the growth of an international spirit and did much to facilitate the negotiation of the Treaty of Rome establishing the European Economic Community and the European Atomic Energy Community (Euratom).

In 1957 the Coal and Steel Community's assembly and court of justice were replaced by parallel institutions established by the European Economic Community. In 1967 its executive organs were merged with those of the European Economic Community and Euratom. The other provisions of the treaty remained unchanged.

THE EUROPEAN ECONOMIC COMMUNITY

The Common Market

The European Coal and Steel Community was only the initial step in the movement for European integration. On March 25, 1957, its six member governments signed the Treaty of Rome, under which they agreed to establish the European Economic Community, or "Common Market," which came into being on January 18, 1958. In 1973 it was enlarged with the entry of Great Britain, Ireland, and Denmark. The EEC is the most far-reaching attempt at economic integration among sovereign countries. Its founding treaty has been the model, in whole or part, for all subsequent attempts at economic integration.

The Treaty of Rome aimed to "establish a common market" and "progressively bring the economic policies of members into alignment" so as to

> promote the harmonious growth of economic activity in the Community as a whole, regular and balanced expansion, augmented stability, a more rapidly rising standard of living, and closer relations between the participating states.

The treaty pledged the signatories to

> abolish customs duties and quantitative restrictions on the entry and outflow of merchandise, to abrogate all other measures having an equivalent effect, and to fix a common customs tariff for imports from non-member states.

They also agreed to "abolish, as between members, all barriers to the free movement of persons, services and capital." This was to be accomplished during a transition period of 12 years. The transition period ended on January 1, 1970, and the community then entered into its definitive phase.

Formation of a customs union. The treaty set a timetable for the abolition of customs duties between member states. On balance, this timetable was met and in some areas exceeded so that, by the middle of 1968, tariff barriers had been abolished for agricultural as well as industrial products. By that date also, most quota restrictions had been lifted. The customs posts had not disappeared, however; they were still needed for such tasks as assessing and collecting the compensatory taxes that equalized the differences in taxes between member countries.

Tariffs on imports from outside the community were gradually brought closer, and on July 1, 1970, a common community tariff was put into effect. The treaty had envisaged that the common tariff would be a simple arithmetical average of the rates at the beginning of 1957 in the four customs territories of West Germany, Benelux, France, and Italy—about 14 percent. But international tariff negotiations under the auspices of GATT brought the EEC members' tariffs down to an average level of 10.7 percent by the middle of 1968, and they were scheduled to fall to 7.5 percent in 1972.

Development of a common agricultural policy. When the treaty took effect at the beginning of 1958, agriculture was subsidized in all six member countries. The various price-support mechanisms differed substantially, as did foreign trade policies and tariff levels. The cumulative impact of governmental intervention of various kinds over the years had led to major differences among the members in agricultural price levels. Taking the average price of wheat in the six countries in 1959 as 100, the relative price levels in individual countries were: Germany, 108; France, 78; Italy, 108; Belgium, 101; Luxembourg, 119; and The Netherlands, 86. The achievement of common policies in agriculture appeared to be so difficult that the treaty limited itself to setting forth a number of general provisions on which agreement seemed feasible. Despite this, in 12 years a common agricultural policy had been achieved: all tariff and quota restrictions on trade in farm products among member countries had been abolished; a common set of tariffs on agricultural imports from non-EEC countries had been established; and a common system of price supports had taken the place of the former national systems.

The harmonization of agricultural policies

The price supports required difficult compromises among the member governments because of the differences in their domestic price levels for farm products. The EEC wheat price, for example, was set roughly halfway between the prices of the lowest-cost suppliers in the community, France and The Netherlands, and those of Germany, which was the highest. France exerted considerable political pressure to persuade Germany to accept a substantial lowering of the returns to its wheat producers. The community prices are supported by purchases from a common fund. The fund begins buying in any area where the price of a crop drops to a fixed intervention level. The cost of the price support program is financed by contributions from the members. The net cost for the Six of the price support was about $2,000,000,000 in 1970, in relation to an agricultural output valued at approximately $35,000,000,000.

Toward an economic union. A second fundamental aim of the Treaty of Rome was to achieve an economic union among the signatories. This would require a general harmonization of national economic policies. The treaty envisaged the working out of common rules covering such matters as competition, taxation, and economic legislation. It also called for the development of common policies in such areas as foreign trade, agriculture, and transportation. Members were asked to concert their economic policies in the fields of fiscal and monetary policy, balance of payments policy, and social welfare. A great deal of work had been done along these lines by the end of the 1960s.

Relations with other countries. *Privileged association with overseas countries and territories.* The treaty provides that overseas countries and territories having a special relationship with Belgium, France, The Netherlands, and Italy may be granted associate membership. The purpose of this was to eliminate preferential tariff arrangements of any of the six members with outside countries by extending the same preferential terms to all of the Six. The treaty also provides for a European Development Fund to assist the economic development of these countries. In December 1970 some 31 countries were associated with the community through this provision. There were 18 African states: Burundi, Cameroon, Central African Republic, Chad, Congo (Brazzaville), Dahomey, Gabon, Ivory Coast, Malagasy Republic, Mali, Mauritania, Niger, Rwanda, Senegal, Somalia,

Togo, Upper Volta, and Zaire. There were nine overseas territories: Netherlands Antilles, Comoro Islands, New Caledonia, French Polynesia, St. Pierre and Miquelon, Surinam, French Territory of Afars and Issas, French Southern and Antarctic Territories, and Wallis and Futuna Islands. There were four overseas French departments: Guadeloupe, French Guiana, Martinique, and Réunion.

New members. Any European state may request membership in the EEC. Acceptance requires a unanimous decision by the present members after "the conditions for entry and the modifications to be made to the Treaty as a result" have been agreed upon by the member states and the would-be entrant. Great Britain was a candidate for membership in 1961 and again in 1967. The first attempt failed in 1963, following a veto by France. Negotiations on the second application began in 1970 (see below *The issue of British entry*). Other countries that requested membership were Denmark (1961 and 1967), Ireland (1961 and 1967), and Norway (1962 and 1967).

Associates. The EEC may also conclude an agreement of association with a nonmember country, a union of states, or an international organization. The associate is entitled to special terms in its trade with the EEC and can send representatives to its meetings. Associate status providing for a custom union after a 12-year transitional period was granted to Greece in 1962 and Turkey in 1964. Partial association agreements have been signed with Nigeria (1966); the East African states of Kenya, Uganda, and Tanzania (1968); Tunisia (1969); Malta (1970); and Israel (1970). Negotiations for associate status have been undertaken with Austria (1961), Sweden (1961), Switzerland (1961), Cyprus (1961), Portugal (1962), Spain (1962), Algeria (1963), United Arab Republic (1970), and Lebanon (1970). In the case of Spain, association was expected to be a first step toward full membership.

Constitution of the EEC. The execution of the tasks to be undertaken by the European Economic Community is entrusted to a council, a commission, an assembly (the European Parliament), and a court of justice. The council and commission are assisted by an economic and social committee with advisory functions.

The council has decision-making power in all matters falling within the sphere of competence of the community. Each member sends one delegate. During the transition period before 1970, all important decisions had to be taken unanimously; they may now be taken by either absolute majority or "qualified" majority (*i.e.*, with weighted voting rights) or unanimously. Either a qualified majority or unanimity are required on certain matters by the treaty. Decisions must be taken unanimously on a very large number of issues.

The commission is a body of 13 members "selected by reason of their general competence and offering the utmost guarantees of independence." They are appointed by agreement among the member countries, not more than two being of the same nationality. Members serve a four-year term and may be reappointed. They are expected to act in the general interest of the community, without deference to any government or other organization.

The commission's basic function is to watch over the application of the treaty and to assist the council with recommendations or advice. Its powers for the most part are delegated to it by the council. The commission maintains a staff of several thousand employees in Brussels.

The assembly is formed of 198 delegates selected from the parliaments of the members. France, West Germany, Italy, and the U.K. each have 36 seats, Belgium and The Netherlands 14, Denmark and Ireland 10, and Luxembourg 6. In principle the assembly meets once a year. It passes recommendations and resolutions, and discusses the commission's annual report. It may pass a resolution of no confidence in the commission by a two-thirds majority, resulting in immediate dismissal of the commission members.

The court of justice is charged with the interpretation and application of the treaty. It is composed of nine judges appointed by mutual agreement of the governments of the member states for six-year terms. Actions may be brought in the court by any member state or by any physical or legal person. The court's powers are considerable since community law takes precedence over the national laws of each member country.

The economic and social committee is a consultative body composed of representatives of various economic and social strata, including manufacturing, agriculture, transportation, trade, handicrafts, the liberal professions, wage earners, and the public. Its members are appointed for a four-year term by unanimous decision of the council, and may be reappointed. Membership on the committee is fixed as follows: Belgium 12, Germany 24, France 24, Italy 24, the U.K. 24, The Netherlands 12, Denmark 9, Ireland 9, and Luxembourg 6.

THE EUROPEAN FREE TRADE ASSOCIATION

A response to the Common Market

When the European Economic Community was being organized, Great Britain sought to organize a free trade area that would include 17 member countries of the Organization for European Economic Co-operation. This would have given Britain access to the benefits of the industrial common market on the continent while avoiding possible infringements of British sovereignty. The effort failed, mainly because of French opposition. Britain then undertook the formation of a free trade area in association with Austria, Denmark, Norway, Portugal, Sweden, and Switzerland.

The convention setting up the European Free Trade Association (EFTA) was signed at Stockholm on January 4, 1960. The preamble stated that one of the main purposes of the organization was to "facilitate the future establishment of a wider multilateral association for abolition of customs barriers." More specifically, it was intended as a mechanism for freeing trade with the six Common Market countries without subscribing to the commitments of political character embodied in the Treaty of Rome. In the meantime, EFTA gave its seven members a stronger bargaining position vis-à-vis the other six, as well as the means of creating a large market of their own.

The EEC and the EFTA compared. The European Free Trade Association and the European Economic Community differed in a number of important respects. Whereas the EEC included agricultural products in its scope, EFTA excluded them. Whereas the Treaty of Rome establishing the EEC provided for free movement of persons, services, and capital, the Treaty of Stockholm in general did not. The EFTA, unlike the EEC, did not establish a customs union and a common trade policy. Although tariff barriers were to be abolished between member states of EFTA, each maintained its own schedule of duties vis-à-vis outside countries. Thus, although the EEC was much more than a simple customs union, since it sought to develop common economic policies, the EFTA was much less than a customs union. The EFTA was, in effect, an association formed to deal with temporary concerns; it would cease to exist should Britain become a member of the EEC.

Operation of the EFTA. The European Free Trade Association was governed by a council composed of one member from each participating state. The council found it useful to set up a joint consultative committee composed of representatives of industry, business, and labour; a set of six permanent technical committees (on customs, trade, economic development, agriculture, economics, and budget); and some 40 working parties dealing with special topics.

The EFTA treaty, like that of the EEC, provided for a transitional period, set forth rules governing competition, and called for the abolition of all indirect protection and trade discrimination.

The EFTA had one special problem arising from its nature as a free trade area. Since the duties charged on imports from outside countries might differ from one member to another, traders could take advantage of the differences by channelling imports through the country levying the lowest rates and delivering them to customers in another member country. Rules were established to prevent this by classifying merchandise according to

whether it was produced or fabricated in one of the member countries. In the case of goods made from imported raw materials, the rules required that the import content not exceed 50 percent of the export price of the finished product.

The EFTA's record. Although a ten-year transitional period was originally envisaged, internal customs barriers were eliminated on January 1, 1967, three years ahead of schedule.

The EFTA passed through two grave crises in the 1960s. The first was in 1961 when Britain, acting unilaterally, informed its partners that it had applied for membership in the EEC. The upshot was a joint declaration in which the EFTA members committed themselves to "coordinate their action and remain united throughout the negotiations." The second crisis occurred in October 1964, when, to shore up the pound sterling, Britain suddenly introduced a surcharge of 15 percent on all its industrial imports—an act that was in violation of the treaty. But the crises in EFTA were not as acute as those in EEC when France vetoed the British application for membership in January 1963, or during the conflict over majority voting (which eliminates the national veto) in the EEC council of ministers in July 1965.

ECONOMIC INTEGRATION IN LATIN AMERICA

Progress toward economic integration in Europe encouraged the Latin American republics to make similar attempts. By the 1970s four organizations had been established to work toward such integration: the Central American Common Market; the Latin American Free Trade Association; the Andean Development Corporation; and the Caribbean Free Trade Association. The Punta del Este declaration of 1967 set forth the principle that the various regional units should be integrated.

The Central American Common Market. On June 10, 1958, El Salvador, Guatemala, Honduras, Nicaragua, and Costa Rica signed a multilateral treaty aiming at free trade and economic integration. It provided for the establishment of a free trade area within ten years. The participating countries also agreed to the industrial integration of the region. These arrangements were completed by the signature on December 13, 1960, of the Treaty of Managua. Its aims were similar to those of the EEC in Europe, namely, the establishment of a common market within five years and the organization of integrated industrial development. The Central American Common Market included a total population of 14,000,000 persons in 1969. By November 1968, a series of overall tariff cuts had freed 98 percent of the region's internal trade, and a single customs tariff had been introduced for most products imported from outside the area.

The Latin American Free Trade Association. On February 18, 1960, Argentina, Brazil, Chile, Mexico, Paraguay, Peru, and Uruguay signed a treaty setting up the Latin American Free Trade Association. By 1970 the seven signatories had been joined by Ecuador, Colombia, Venezuela, and Bolivia. The treaty provided for a 12-year transition period during which all obstacles to trade were to be eliminated. It was based on the principle of reciprocity and most-favoured-nation treatment. Member states also committed themselves to progressive coordination of their industrialization policies. Special treatment was provided for agriculture and for the relatively least-developed member countries.

Liberalization of trade between the member countries was carried out initially through negotiation of product-by-product concessions. In 1967, however, the negotiations failed; they were postponed to 1968, when agreement was reached on a system of across-the-board automatic tariff reductions similar to those of the European Economic Community. But the Latin American Free Trade Association, which had been intended as the first step in a process that was to lead to a common market, had yet to become more than a series of preferential arrangements based on a large number of bilateral trade, industrial, and financial accords.

The Andean Development Corporation. In 1966 Bolivia, Chile, Colombia, Ecuador, Peru, and Venezuela, all members of the Latin American Free Trade Association, agreed to form a regional subgroup. The Andean Group finally began its official existence in June 1969 without Venezuela, which had withdrawn.

The Caribbean Free Trade Association. In 1968 five former British colonies (Antigua, Barbados, Guyana, Jamaica, and Trinidad and Tobago) signed the Treaty of Antigua creating the Caribbean Free Trade Association. This organization has not yet developed any definite institutional link with the general movement toward economic integration in Latin America.

Toward further integration. The Declaration of the American Presidents at Punta del Este on April 14, 1967, recommended unification of the various movements toward economic integration. Seventeen countries, including the five members of the Central American Common Market, the 11 members of the Latin American Free Trade Association, and Panama, committed themselves to create a common market over a 15-year period by progressively integrating the Caribbean Free Trade Association and the Latin American Free Trade Association.

So far, attempts at economic integration in Latin America have encountered a number of obstacles, some of which are peculiar to the political and social conditions of that region. These include a heterogeneous and rapidly expanding population, very high rates of price inflation in some countries, a legacy of national antagonisms, and major differences among the countries in size and economic development. In consequence, economic integration is much more difficult to achieve in Latin America than in Europe.

THE COUNCIL FOR MUTUAL ECONOMIC ASSISTANCE

Integration in eastern Europe

A Soviet-sponsored effort to integrate the economies of eastern Europe began as early as January 25, 1949, in response to the Marshall Plan. The founding states were Bulgaria, Hungary, Poland, Romania, Czechoslovakia, and the Soviet Union. Albania joined in 1949, the German Democratic Republic in 1950, and the Mongolian People's Republic in 1962. Albania ceased to participate after 1961. Mainland China, Cuba, North Korea, North Vietnam, and Yugoslavia participate in the organization as observers with associate membership. In its early years the activities of the Council for Mutual Economic Assistance (or Comecon) were limited mainly to the registration of bilateral trade and credit agreements among the member countries. After Stalin's death in 1953 it made efforts to promote industrial specialization and to reduce "parallelism" in the economies of its members. In 1956 and 1957, when most of its standing commissions began to operate, attempts were made to harmonize the long-term plans of the members.

The establishment of the European Economic Community in 1958, together with pressures from the eastern European countries for a greater degree of independence, induced the Soviet leadership to rethink the organization. A new charter was signed by the members in Sofia on December 14, 1959.

Economic objectives. The economic aims of the Council for Mutual Economic Assistance are

> through unification and coordination of the efforts of its member countries to contribute to the balanced growth of their economies, more rapid economic and technical progress, increased industrialization of those countries which are least advanced in this area, the continued growth of labour productivity, and the regular improvement of the welfare of the peoples of member countries.

Its pursuit of these objectives has been hindered by certain political and economic constraints. One of the most serious is the absence of flexible and realistic price systems in the member countries. This has made it impossible to base trade on relative prices; instead it has been conducted mainly on a barter basis through bilateral agreements between governments. In negotiating such agreements, the parties have been led to use "world prices"; *i.e.*, the prices prevailing in the trade of countries outside the Council for Mutual Economic Assistance.

Another hindrance to economic integration is the highly centralized economic planning that is the rule in the

member countries, which so far have had only limited success in coordinating their plans.

There have also been serious nationalistic tensions within the council. The Romanian government, for example, announced its intention to pursue all-around industrialization, including the development of its heavy industries, in opposition to the policy of specialization in raw materials and agricultural products that was said to have been the Council for Mutual Economic Assistance's policy for Romania.

Among the practical achievements of the Council for Mutual Economic Assistance, however, have been: the organization of railroad coordination (1956); construction of a high-voltage electricity grid (1962); creation of the International Bank for Economic Cooperation (1963); the pooling of 93,000 railway freight cars (1964); and construction of the "Friendship" oil pipeline from the Soviet Union's Volga region to the eastern European countries.

The constitution of the Council for Mutual Economic Assistance. The highest authority of the organization is the council-in-session. It is composed of delegations from all member countries, the composition of each delegation being fixed by the government concerned. The conference of representatives of member countries, composed of one representative of each member country, may issue recommendations and decisions. It may also submit proposals for examination by the council-in-session.

Various permanent commissions are composed of experts and official representatives of member countries. Some are general economic commissions; others deal with issues affecting specific industries. The headquarters of the various commissions are located in the capital cities of member countries. The central secretariat is in Moscow.

The Council for Mutual Economic Assistance is often called the eastern European counterpart of western Europe's EEC. Although the general aims are indeed the same, the two organizations differ radically in their approach to the problems involved. The EEC aims to achieve integration on a decentralized basis by means of an economic market in which goods, services, capital, and persons have full freedom of movement—a market regulated by uniform economic legislation. The Council for Mutual Economic Assistance seeks to achieve cooperation among national economies each of which is centrally planned and administered.

AN APPRAISAL OF THE EUROPEAN ECONOMIC COMMUNITY

The accomplishment in western Europe

Since the EEC has, in varying degree, served as the model for every subsequent attempt at economic integration, and since the problems it raises are of broad scope, some analysis of its accomplishments and difficulties will be useful.

The Common Market. By July 1970 a practically complete customs union had been achieved. All internal trade restrictions, whether tariffs or quotas, had been removed; a unified schedule of customs duties was in force on industrial and agricultural products imported from outside countries; and most quantitative restrictions on imports from outside the community had been abolished. There remained, nonetheless, six primary areas in which the Common Market was still incomplete. These had to do with prices, customs legislation, technical legislation, government contract letting, fiscal legislation, and trade policy.

1. Although internal customs barriers had been removed, there were still considerable differences in the prices of similar products among the member countries. In 1969 the prices of radio and television sets in France, for example, were almost 50 percent higher than they were in Germany. The price of food was 15 percent higher in Italy than in Luxembourg. Textiles cost 33 percent more in France than in The Netherlands.

2. The operation of a true customs union requires harmonization of customs legislation on such technical matters as bonded warehouses and systems for payment of duty. This is indispensable if trade is to flow on equal terms among the members.

3. Different technical norms among countries are an obstacle to true liberalization of trade. For example, in 1970 automobile windshields could not be exported from some countries of the community to others because of differences in specifications. It was estimated that the total volume of trade could be increased by one-third if homogeneous technical standards were adopted.

4. Government purchases and contracts for public works were administered on a highly discriminatory basis in each member country, preference being given to national suppliers.

5. Harmonization of fiscal legislation is also essential for the achievement of an effective common market. When taxes fall more heavily on some producers than on others, the effect is to make them less competitive. Thus it has been argued that a producer in a country in which the turnover tax rate is 25 percent cannot meet competition from another country in which the rate is only 10 percent. For this reason the EEC allows members to make drawback payments on exports up to the amount of direct and indirect taxation borne by the products concerned, and to levy taxes on imports up to the amount of direct and indirect taxation borne by similar national products. This is easy to do with respect to indirect taxes (sales and turnover taxes), but there is no practical way to assess the relative burden of direct taxes (income and property taxes) in different countries. Only similar tax laws in the countries can eliminate this distortion.

6. Although the six members of the EEC had adopted a common trade policy, there were still important gaps to be bridged, especially concerning trade with the Communist world. One issue in this was the desire of West Germany to have a free hand in its negotiations with eastern Europe.

From the Common Market to economic union. Further progress toward economic union would require new legislation in each of the member countries, the merger of the three treaties governing the Coal and Steel Community, the Economic Community, and Euratom (meaning that they must be revised), far-ranging agreement on industrial and agricultural matters, and, most important of all, a single monetary policy.

Although a common agricultural policy had been developed, it was based on the maintenance of relatively high prices. In 1967–68 the guaranteed prices for corn, wheat, sugar beets, and butter ranged from 1.6 to 4 times the level of world market prices. Although the labour force employed in agriculture fell from 20 percent of the total labour force to 13 percent between 1960 and 1970, the high prices, in association with rapid productivity growth, generated massive commodity surpluses. The mounting cost of the price support program caused increasing dissatisfaction in West Germany and Italy, the two countries on which the burden mainly fell. In addition, EEC was accused of dumping its commodity surpluses on the world markets at disastrously low prices. Despite its cost, moreover, the agricultural policy of the 1960s failed to provide all farmers with incomes equivalent to those enjoyed by the rest of the population.

Economic union would also require a unified policy on money and credit, leading to the introduction of a single currency. There were two major obstacles to be overcome. The first obstacle was political: a single currency would require the members to renounce some of their national sovereignty. The second obstacle was social: even if a common monetary policy could be applied, it would be viable only if the wage level in each country was fixed by the market, a policy difficult for the national trade unions to accept. A permanent liberalization of trade and investment within the EEC would in fact require that the national currencies be made fully convertible at an unvarying rate of exchange—which would mean the effective existence of a single currency.

Toward political union. The statesmen who drew up the Treaty of Rome regarded the EEC as a stage in the political unification of Europe. This view was made explicit in the provision for election of deputies to the European Assembly by universal suffrage. It also found expression in those clauses providing that at the end of the transition

Problems of sovereignty

period many decisions were to be taken by absolute or qualified majority vote and were to be binding on every member state.

The members of the EEC had not, by the beginning of the 1970s, shown themselves ready to give up their political independence. No member could, in fact, allow vital decisions to be imposed on it by majority vote of the council. To do so would mean allowing a treaty to override its constitution, and a vote taken by five foreign ministers to overrule a government elected by universal suffrage. This was the root of the 1965 crisis, when France left its seat on the council from June 30, 1965, to February 7, 1966. The nominal cause of the rupture was agricultural policy. The crisis was resolved at an extraordinary council meeting at which the six states tacitly agreed not to invoke the rule of a majority on issues in which the vital interests of a member were at stake.

Despite these difficulties, the cohesion of the EEC did become progressively more concrete. During the Kennedy Round tariff negotiations (1964–67), the community delegation spoke for the six member countries throughout, even though the transitional period had not yet expired and EEC's common tariff had not yet been put into force. Again, in the sensitive area of agriculture, the members relinquished autarkic national policies and adopted a liberally oriented common policy.

The issue of British entry. In August 1961 Great Britain applied for membership in the EEC. Denmark, Ireland, and Norway entered their applications shortly afterward. Negotiations began in November 1961, and were terminated in February 1963 by a French veto. Britain applied again in 1967, and negotiations resumed in July 1970; in June 1971 agreement was reached by which Britain would enter the EEC on January 1, 1973. The terms called for Britain's adoption of the common EEC tariff, the common agricultural policy, and the rules of the Community, by transitional stages over a period of five years. Britain's share of the Community budget would start at 8.64 percent in the first year, rising to 18.92 percent in the fifth year.

Four major economic considerations were involved: the interests of the British Commonwealth countries, the cost to Britain of importing European agricultural products, the matter of London's sterling balances, and the interests of Britain's partners in the European Free Trade Association. The transition period was expected to protect the Commonwealth countries from undue hardship in the gradual loss of their preferential trading advantages with Britain. Although the common agricultural policy would raise food prices in Britain, this would be offset by the opportunities opened to British industry on the Continent. The sterling balances, however, remained a barrier in the path of British financial integration with the EEC. And of Britain's partners in EFTA, only Denmark chose to enter the Community in 1973. In Norway a referendum in September 1972 rejected the country's proposed entry. The possibility remained that the other EFTA members, Sweden, Switzerland, Austria, and Portugal, would be granted associate membership in the EEC. Ireland, not an EFTA member, joined EEC along with Britain.

BIBLIOGRAPHY

Historical works: F.W. TAUSSIG, *Free Trade, the Tariff and Reciprocity* (1920); LEAGUE OF NATIONS, *Commercial Policy in the Inter-War Period: International Proposals and National Policies* (1942); HERBERT HEATON, *Economic History of Europe*, rev. ed. (1948); G. LACOUR-GAYET, *Histoire du commerce* (1955); ROBERT SCHNERB, *Libre-échange et Protectionnisme* (1963); GARDNER PATTERSON, *Discrimination in International Trade: The Policy Issues, 1945–1965* (1966).

International organizations: H.K. JUNCHERSTORFF, *Common Market* (1963); FRANCOIS VISINE, *ABC de l'Europe*, 4 vol. (1967–69); FRANCOIS CLERC, *Le Marché Commun Agricole* (1970); F.A.M. ALTING VON GEUSAU, *Economic Relations after the Kennedy Round* (1969); GRADUATE INSTITUTE OF INTERNATIONAL STUDIES, *The European Free Trade Association and the Crisis of European Integration* (1968); *Britain and the European Communities: An Economic Assessment Presented by the Prime Minister by Command of Her Majesty, February 1970* (1970); ROLAND HILTON (ed.), *The Movement Toward Latin American Unity* (1969); WALTER KRAUSE and F. JOHN MATHIS, *Latin America and Economic Integration* (1970); JOHN S. LAMBRINIDIS, *The Structure, Function and Law of a Free Area* (1965); F.L. PRYOR, *The Communist Foreign Trade System: The Other Common Market* (1963); K.R. SIMMONDS, "The Central American Common Market," *The International and Comparative Law Quarterly*, 16:911–945 (1967); M.S. WIONCZEK (ed.), *Latin American Economic Integration: Experiences and Prospects* (1966).

Trade liberalization and economic integration: MAURICE ALLAIS, *L'Europe unie, route de la prospérité* (1960), *Le Tiers-Monde au Carrefour* (1962), "Toward an Integrated Atlantic Community," in *Nato in Quest of Cohesion* (1965), and *La Libéralisation des relations économiques internationales, accords commerciaux ou intégration économique* (1971); J.D. BELA BALASSA, *The Theory of Economic Integration* (1961), and *Trade Liberalization Among Industrial Countries* (1967); W.M. CORDEN, *Recent Developments in the Theory of International Trade* (1965); EUROPEAN FREE TRADE ASSOCIATION, *The Effects of EFTA on the Economies of Member States* (1968); GOTTFRIED HABERLER, *The Theory of International Trade with Its Applications to Commercial Policy* (1933, reissued 1959); JAMES E. MEADE, *Problems of Economic Union* (1953), and *The Theory of Customs Unions* (1955); LIONEL ROBBINS, *Economic Planning and International Order* (1936); JACOB VINER, *The Customs Union Issue* (1950). (Ma.A.)

Cutlery and Tableware

Cutlery is the general collective term applied to cutting implements, such as knives, razors, and scissors, used for industrial, commercial, and domestic purposes. Tableware comprises both hollow ware, including such items as dishes, platters, teapots, and coffeepots, and flatware, consisting of the various forms of spoons and forks used at the table.

ORIGINS AND DEVELOPMENT

Prehistoric implements used for cutting, hunting, and defense were fashioned from stone, especially flint; from obsidian, a volcanic glass; and from bones and shells. Cutting edges were formed by rubbing the implement in the hollow of a stone, a method still employed by aborigines of central Brazil, Australia, and New Guinea. In the earliest spoons, baked clay formed both the bowl-shaped receptacle portion and the supporting stem or handle. Later, spoons were made from suitably shaped bone or wood pieces. By 1500 BC bronze cutting implements were being used from the British Isles to China. Scissors with blades connected by a C-shaped spring at the handle end also originated at about this time. As various metals became known, the art of forging blades developed in China, India, and Europe. Pivoted scissors of bronze or iron, connected by a rivet or screw between the handles and blades, were known in ancient Rome and in China, Japan, and Korea.

Early scissors

The Egyptians fashioned cutting implements from flints chipped to form serrated edges and then glued into slots in wood that had been appropriately shaped for the intended purpose. Knives served mainly for hunting and as weapons, but the wealthy used small ornamental eating knives. Spoons were frequently made of bronze, some having spiked handles to extract snails from their shells. Elaborate cosmetic spoons had carved handles representing human or animal forms; long incense spoons served ceremonial functions.

The Greeks produced bronze knives, and the Romans spread blade-making techniques throughout the Mediterranean and Europe. As in Egypt, small ornamental eating knives were used by the wealthy. Both Greeks and Romans employed bronze and sometimes silver for spoons. Some Roman spoons, made of bone, had small holes in the centres of their bowls; the purpose of these holes is not known.

In western Europe the Celts used short bronze spoons with broad shanks formed to fit the hand. Knives were usually pointed, allowing food to be speared. Steel-bladed eating knives dating from the Roman period have been found in Italy and Britain.

As knowledge of techniques spread, cutlery production was established in areas able to offer plentiful timber to heat furnaces and provide charcoal, in addition to soft water for the hardening and tempering of steel. Medieval

grindstones were sometimes hand-operated, but animal or water power was frequently employed to revolve treadmills or wheels. From about 1200 cutlery manufacture began to settle in London and Sheffield in England; in Thiers and Paris in France; in Solingen, Germany; and in many other places where craft guilds were founded. Craftsmen produced elaborately ornamented blades and fashioned handles of such fine materials as gold, silver, ivory, ebony, agate, amber, and marble.

Table cutlery was not provided by innkeepers, and the affluent possessed elegant travelling sets. Others used plain knives with handles of bone or wood and crude molded forks and spoons made by tinkers from an alloy of lead and antimony. In the homes of the wealthy it became usual to provide knives for guests, though most men still carried their own. Serving knives made in pairs, sometimes called *présentoirs*, were used only for passing food. Sets known as "wedding knives," consisting of a pair of knives in a sheath, were common gifts from bridegrooms to their brides.

The invention of forks

Forks, which originally had a single point, were made with two prongs by the Romans. In the Middle Ages large forks with two flat prongs were used for serving. Smaller eating forks were gradually developed, replacing the traditional pair of pointed table knives that were part of the transition to knife and fork. Handles were sometimes made of precious or semiprecious materials.

Silver spoons originally had long, pointed bowls, but by the later Middle Ages the bowls were frequently fig-shaped, while the stems were often topped with decorative knobs. Matching sets of spoons and forks in standard patterns were common by the mid-18th century. The modern tablespoon, with its stem ending in a rounded curve and turned downward, was adopted about 1760. Table knives of the 18th century frequently had pistol-shaped handles and curved blades like those of scimitars. Although individual eating knives were no longer carried for ordinary use by the late 17th century, sets consisting of knife, fork, spoon, and drinking vessel were still being made for travellers well into the 19th century.

The rise of Sheffield as a manufacturing centre

By the 18th century Sheffield, England, had become an international centre of the industry. In the early 1700s Sheffield cutlers and silversmiths were making forks having two or three prongs and knives with hollow silver handles that were stamped in two halves, soldered together, and filled with pitch into which the tang, the projecting portion of the knife blade, was inserted. Large-scale production of pivoted scissors and shears began in 1761, when Robert Hinchliffe of Sheffield first used crucible cast steel for their manufacture. By the 19th century scissors with elaborately designed hand-filed and polished bows and shanks were made in Europe.

Steel razors were made with ornamental handles, and blades were individually hollow-ground, producing a concave surface behind the cutting edge. The forerunner of the modern safety razor, with a guard along one edge, was introduced in 1828. In 1880 a hoe-shaped safety razor was manufactured in the U.S., and early in the 20th century King C. Gillette began to manufacture a model with double-edged replaceable blades.

MATERIALS USED FOR CUTLERY AND TABLEWARE

Plated materials. Sheffield plate, produced by a process developed about 1743 by Thomas Boulsover of Sheffield, consisted of thin sheets of silver fused onto small copper ingots. The ingots were then rolled into sheets, and a thin coating of silver remained on the surface. Between 1750 and 1880 this material was employed for such items as knife handles, serving dishes, tea urns, and candelabra, manufactured mainly in Sheffield but also in Birmingham, England.

By about 1860 a new process of electroplating, the application of metallic coatings to base metals by means of an electric current, superseded Boulsover's fusion process. Electroplating of silver onto alloys of nickel and copper was soon common, followed by the plating of nickel onto brass. Sheffield plate was no longer commercially manufactured, and surviving pieces eventually became valuable antiques (see also ELECTROPLATING).

Cutlery steel. Cutlery steel consists of iron to which from 0.35 to 1 percent carbon has been added. Early methods involved hammering charcoal into red-hot iron bars. In the 18th century Benjamin Huntsman built new types of furnaces in Sheffield for making highly refined steel in clay vessels called crucibles. His process greatly increased both the availability and quality of steel during the first part of the Industrial Revolution. Mass production of steel was achieved in the 19th century by the open-hearth process.

Stainless steel. Harry Brearley of Sheffield, England, discovered one form of this steel in 1913 and others were developed during the following decade in Germany and the United States. Although not strictly "stainless," these steels are highly resistant to corrosion from common household acids and alkalis. Corrosion resistance depends upon such factors as the amount of chromium they contain, satisfactory heat treatment, and the quality of the grinding and finishing processes (for a description of the different types of steel and their methods of production, see STEEL PRODUCTION).

Martensitic stainless steels, widely used for both table knives and trade knives, contain from 12 to 18 percent chromium, imparting corrosion resistance, and from 0.12 to 1 percent carbon, permitting a great degree of hardening by heat treatment. Edge retention increases with higher carbon content; corrosion resistance is increased by higher chromium content but reduced with additional carbon.

Austenitic stainless steels contain about 18 percent chromium and 8 percent nickel. They can be hardened to a limited extent by cold-rolling or stamping but not by the application of heat. Such types have corrosion resistance superior to that of martensitic steels but cannot be adequately hardened and tempered to allow grinding to a cutting edge. They are, however, easily drawn out without breaking, making them suitable for both flatware and hollow ware production.

By 1928 the manufacture of carbon-steel blades was limited to commercial knives and some carving, hunting, and pocket knives.

Flatware materials. Though since about 1860 much flatware has been silver-plated by the electroplating method, the use of stainless steel has grown steadily since 1920. New designs for stainless-steel flatware developed rapidly in Europe and the U.S. By 1955, similar growth had taken place in Japan, Taiwan, Korea, and Hong Kong. In the early 1970s most of the world's cutlery was made of martensitic stainless steel and most flatware was made of austenitic varieties. Ferritic stainless steel, containing 12 percent chromium, was being used for less costly flatware, particularly in the Far East. Spoons and forks manufactured for use in food preparation were frequently made of stainless steel.

Other flatware materials include gold for luxury services and unplated nickel alloys, aluminum, tin-coated iron, and plastics for inexpensive ones. Wood and natural horn are popular for salad servers. Aluminum is especially useful where lightness and low cost are desired; lightweight plastic eating implements, made in various colours, are produced for picnic sets, ice cream spoons, and airline food service. The least expensive materials for metal flatware are regular steels electroplated with copper, nickel, or chromium.

Silver plate and sterling

Silver-plated flatware is manufactured by electroplating silver onto a base metal such as finely buffed nickel silver (an alloy consisting mainly of copper, zinc, and nickel) or stainless steel, its quality being determined by the strength and composition of the base metal, the standard of finish, and the thickness of the silver deposit.

Solid-silver flatware, utilizing essentially pure silver, is a luxury item. Standards for silver purity vary, the principal one being not less than 925 parts of fine silver in 1,000 parts, established by the British assay offices for silver hallmarked as "sterling." The balance is copper or other base metals that add strength to the finished piece. Similar controls exist in Austria, Czechoslovakia, Finland, France, Hungary, The Netherlands, Poland, Portugal, the Soviet Union, Sweden, and Yugoslavia. Some

European nations accept a lower standard of 800 parts of silver in 1,000 parts; others employ voluntary marking systems. In Europe, silver articles usually bear hallmarks indicating that the metal contains a prescribed amount of silver. Other marks record the year of manufacture and the maker. United Kingdom hallmarks include the lion passant, a symbol in heraldry, guaranteeing that the silver is not less than 92.5 percent pure; the assay office (town) mark; the monarch's head; the date letter recording the year of manufacture; and the maker's full name, initials, or trademark. In the United States, the word sterling when used by a reputable supplier is accepted as a sufficient guarantee, and there are no fixed standards.

MODERN FABRICATING TECHNIQUES

In the early 1900s, machine processes were rapidly adopted both in the United States and in the major manufacturing centres of Europe. Between 1920 and 1930, forging techniques were revolutionized and stainless steels rapidly supplanted carbon steels.

Cutlery. The many kinds of cutlery all have distinctive uses, the blade being the major determining factor. Retention of sharpness depends on the type of steel employed and the skill with which it is processed. Cost is determined by the quality of the blade steel, workmanship, material used for the handle, and ornamentation. In the early 1970s hand-fabrication methods predominated in such developing areas as India, China, and Africa. In industrialized nations, including the United States, Canada, Britain, Germany, France, Australia, and Japan, and also in Hong Kong, hand labour had been reduced to a minimum.

Table cutlery. In the production of table cutlery described below, processes include: (1) forging the steel into the desired blade shape; (2) hardening and tempering it correctly; (3) grinding the blade to a cutting edge and removing all traces of forging and heat treatment; (4) polishing the blade; and (5) making, fitting, and polishing the handle, a process known as cutling.

Forging knife blades

High-quality knife blades (Figure 1) are forged by mechanical hammers from bars of steel. After heating, a bar is placed between forging dies that rapidly hammer out the rough blade shape desired. The final blade shape is obtained by trimming the forgings. Some knives are made as complete forgings of blade and handle by one stroke of a large drop-forging hammer; others are cut from sheet steel so that a short flat tang without a forged bolster (the part of the knife blade that abuts upon the handle) is left. After forging, the blades are hardened by heating and then quenching in a cooling liquid or between metal plates cooled by an internally circulating liquid. The blades are then tempered by reheating to the correct temperature to give them flexibility and toughness.

By courtesy of the Granville College of Further Education, Sheffield, England

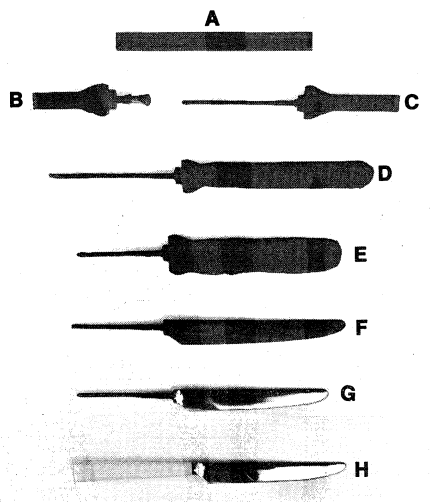


Figure 1: *Stages in the manufacture of a table knife.*
(A) The length of stainless steel bar for two blades. (B) Bolster stamped. (C) Tang drawn out. (D) Blade forged out to approximate length. (E) Blade cross-rolled for widening. (F) Blade cropped to final shape, hardened and tempered. (G) Blade ground and partly polished. (H) Blade fully polished and sharpened with handle fitted.

Grinding consists of applying the blades to the rapidly revolving periphery of an abrasive wheel, removing the steel until the desired tapers from the back to the cutting edge and from bolster to point are attained. The blades are kept cool with water or a "cutting fluid" to maintain their temper.

After either machine or hand grinding, the surface of a blade is given a finer finish in successive operations known as glazing and buffing, followed, if desired, by mirror polishing, or "satin" finishing. The bolsters are also ground, glazed, and polished to fit the desired handle. The maker's name is then applied by acid or electric-arc etching.

Materials used for knife handles

Natural materials used for handles include animal horns and tusks, various woods, mother-of-pearl, and bone; manufactured handles range from gold, silver, and porcelain to stainless steel, silver plate, nickel alloys, compressed wood, and plastics. Some of these materials are processed mechanically and others by the cutler. Natural-stag-horn handles must be matched to each blade individually because no two are alike. Cellulose adhesives, cements, or resin mixtures are used to fix the tangs securely in the handles, except for hollow metal handles, which are secured by hard soldering or welding.

Scissors and shears. By the end of the 19th century mechanical methods of production had begun to simplify styles and patterns. Blanks (unfinished pieces) or scissors are now made by high-speed forging of red-hot steel bars between the dies of drop hammers. For ordinary scissors, steel containing 0.55 percent carbon is mainly used; for the finest scissors and shears for trade use, such as tailors' shears and trimmers, a harder steel containing 0.75 percent carbon is preferred. For cutting man-made fibres, harder and more blunting than natural fibres, the blades of tailors' shears are sometimes made of a composite material consisting of an even harder steel on the cutting side (1.03 percent carbon), backed by tough iron. Many scissors, including surgical varieties, are made of stainless steel, and some specialized scissors and shears are made of nonferrous alloys that will not produce a spark or interfere with magnetism (*e.g.*, for cutting cordite and magnetic tape). Low-cost scissors are made from relatively soft steel wire that is pressed cold and is not hardened.

If scissors are to cut well, the blades must touch in two places only—at the joint and at a single spot along the blades wherever the cutting takes place. The blades are made to twist or curve toward one another, and, when completely closed, their points should touch. Both blades must be accurately tempered to equal hardness. In the finest scissors and shears, the two blanks and the screw, as well as the nut in large shears, are coded with an identification mark early in manufacturing to ensure being treated as a set.

Modern razors. In the early 1960s manufacturers in several countries produced stainless-steel blades for safety razors, allowing longer usage than the older steel blades. Further efforts to improve life and cutting edge led to the use of chromium and even platinum in the early 1970s. Electric razors, operating on a principle similar to that of the barber's hair clipper, were introduced as early as 1931. Steel razors with hollow ground blades continue to be made in most countries for use by barbers.

Flatware. *Forming blanks.* Modern flatware (Figures 2 and 3) is produced in all the cutlery centres of the world. During the 20th century, the processes used in its manufacture reached a high degree of mechanization. The metal, carefully refined, is formed into sheets of proper thickness and cut into strips of the required width. These processes involve the strictest control of metal behaviour and correct annealing to remove ex-

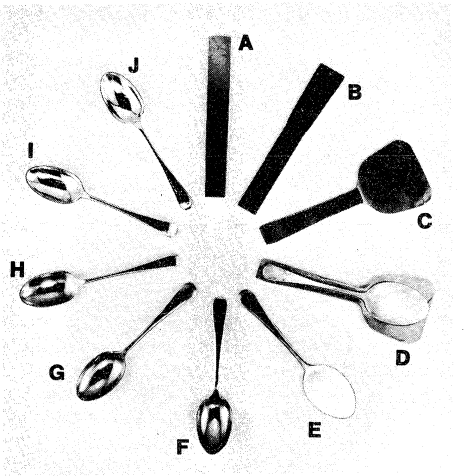


Figure 2: *Stages in the manufacture of a silver-plated spoon.*
(A) Blank of nickel silver alloy for one spoon. (B) Blank cross-rolled to proper thickness and width, which also hardens it. (C) Spoon end cross-rolled thinner than handle. (D) Shape of spoon blanked. (E) Blank handle stamped with pattern. (F) Bowl formed. (G) Spoon set and buffed. (H) Fine buffing. (I) Plating. (J) Polishing.
By courtesy of the Granville College of Further Education, Sheffield, England

Roughing out flatware blanks

cessive strains. The strips are fed into machine presses that cut out each spoon or fork in its rough shape, one end being at first almost square for a spoon and rectangular for a fork. The ends of these "blanks" are rolled again in a direction at right angles to the centre line, reducing the thickness at this point without altering the thickness of the handle. The bowls of the more expensive spoons are no more than half as thick as their handles.

Stamping. After being trimmed, the blanks are stamped in alloy-steel dies that hollow the bowls and stamp a pattern on the handles. In the case of forks,

By courtesy of the Granville College of Further Education, Sheffield, England

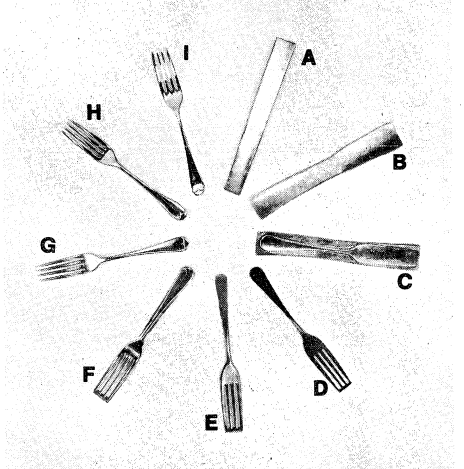


Figure 3: *Stages in the manufacture of a silver-plated fork.*
(A) A blank of nickel silver alloy. (B) Blank cross-rolled to increase its width and hardness. (C) Shape of fork cut out. (D) Fork prongs blanked. (E) Fork buffed (partly polished). (F) Handle pattern stamped on and handle bent to shape. (G) Prong points dressed and bent to shape. (H) Fork silver plated. (I) Final polishing.

slots are cut out to form the prongs, which are then stamped in dies to the required curvature, tapered, and pointed on abrasive belts. These processes are approximately the same whatever metal is used, although in manufacturing cheaper products, made from thinner sheets, cross-rolling can be omitted and the stamping can be performed in one operation.

Finishing. Subsequent finishing processes vary according to the metal used. In the case of silver, successively finer stages of buffing prepare the surfaces for final polishing or satin finishing. In the case of alloys that are to be plated, the articles, after being buffed, are wired individually on frames; quantities of 100 or more can be immersed simultaneously in the series of cleaning baths and plating vats. In most factories the complete frames of articles are transferred automatically from baths to vats and finally to washing and drying. The thickness of deposit is increased by some makers at the points of maximum wear; for example, on the centre of the convex surface of spoon bowls. Although the deposit of silver is specified in grams or pennyweights per dozen pieces and sometimes in actual thickness in millimetres or thousandths of an inch, the more popular method of indication is use of the terms "30 years," "25 years," or "20 years" plate. The designation A1 is considered satisfactory as a guarantee of quality if given by a manufacturer of good repute.

Final polishing. After the pieces have been plated, their surfaces are dull and require polishing. Hand polishing is performed by holding the articles upon rapidly rotating mops dressed with an aluminum compound or rouge. The least expensive plating process is "bright plating," in which a very thin coating of silver or chromium is deposited bright, thus eliminating final polishing. Such coatings are of short duration, and the process is therefore restricted to the cheaper grades of flatware. Stainless steel is more difficult to polish than silver, silver plate, or unplated nickel alloys. Techniques have been developed for stamping the cheaper varieties of stainless-steel spoons and forks from prepolished sheet. In some countries stainless steel is polished electrolytically.

In the most modern finishing operations, various mechanisms stack, feed, and transport the pieces from one process to the next without damage. Edges are smoothed by stacking many spoons and forks together and pressing them onto a series of progressively finer abrasive belts or wheels. To polish the main faces, many spoons and forks are placed side by side, gripped at one end in a long clamp, and fed longitudinally, with a racking motion, into and out of the polishing rolls, so that every portion of the surface is reached. A series of machines is used to achieve progressively finer polishing.

Hollow ware. Many manufacturers of cutlery and flatware also produce hollow ware, in the production of which the principle of the potter's wheel has long been applied; thin sheets of metal are gripped between clamps and rapidly rotated. Pushing a wooden former against the side of the disk bends the metal over, producing the body of a jug or vase. Before soldering was developed, handles and mounts were made separately and rivetted to the body. Shallow dishes and bowls were stamped directly from sheet metal by dropping a large concave stone onto a sheet that had been placed upon a corresponding stone with a convex surface. By the late 17th century, such dies were widely used. Ornamentation, such as embossing or indentation, was added by hand.

In the 19th century the advent of press-tool machinery enabled the initial sheet or blank to be accurately formed by pressure methods. In the modern process stainless-steel blanks are pushed through a die hole by hydraulically operated shaping punches. After annealing, the process is repeated with a smaller but longer punch, producing a seamless drawn body of accurate dimensions.

Forming lids and spouts

Lids are stamped and spouts pressed in halves, subsequently combined by brazing or welding. Handles of low heat conductivity, such as wood or plastic, are attached to handle mounts joined in the same way. Polishing is similar to that used for flatware, with sand and pumice fed to polishing mops to produce smooth surfaces on nickel alloys prior to silver- or chromium-plating. Silver is worked similarly, and many handicraft methods continue in the production of trophies.

WORLD PRODUCTION

Cutlery production statistics are provided by only a limited number of nations. In the early 1970s, the estimated

overall value of Europe's cutlery and flatware production ranged from $450,000,000 to $600,000,000, with West Germany, the United Kingdom, and France the leading nations. The cities of Sheffield, Thiers, and Solingen continued to be major manufacturing centres. During the same period, Japanese production of table knives and flatware in stainless steel amounted to an estimated $60,000,000, while United States production of cutlery was about $270,000,000. Cutlery and flatware of every kind and quality were being manufactured on all continents. In technically advanced countries, electric carving knives were successful; stainless-steel, wafer-type, safety-razor blades and electric shavers had largely supplanted forged-steel, hollow-ground razors; electric scissors, clippers, and shears were widely used in industry and elsewhere.

BIBLIOGRAPHY. C.T.P. BAILEY, *Knives and Forks* (1927), is a profusely illustrated treatise on the uses of cutlery and the work of cutlers from the 15th to the 18th century. H. RAYMOND SINGLETON, "A Chronology of Cutlery" (1965), is a short, illustrated pamphlet that traces the development of table cutlery from the 10th through the 20th century. J.B. HIMSWORTH, *The Story of Cutlery* (1953), provides a comprehensive account of the development and manufacture of various types of cutlery and flatware. D.S. DUGDALE, *The Cutting Tool Story* (1970), traces the development of cutting implements from the Stone Age to modern times. CAMILLE PAGE, *La Coutellerie depuis l'origine jusqu'à nos jours*, 6 vol. (1896–1904), is possibly the most comprehensive work of its type, profusely illustrated and covering cutlery throughout the world. HENRI LANDRIN, *Manuel du coutelier* (1835), discusses in detail the work of cutlery craftsmen, including processes and materials. Some of the world's major cutlery manufacturing centres are treated in: R.E. LEADER, *The History of the Company of Cutlers in Hallamshire in the County of York*, 2 vol. (1905), a comprehensive account of the activities of the cutlers of Sheffield, England (*c.* 1624–1905); G.I.H. LLOYD, *The Cutlery Trades* (1913, reprinted 1968), a work describing and comparing the cutlery trades of Sheffield (England), Solingen (Germany), and Thiers (France); and HARRY TOWNSEND, *The Cutlery Trade* (1958), a survey of the cutlery industry in Sheffield. FREDERICK BRADBURY, *History of Old Sheffield Plate and Silver* (1912, reprinted 1968), is an early standard work treating this particular ware and silver.

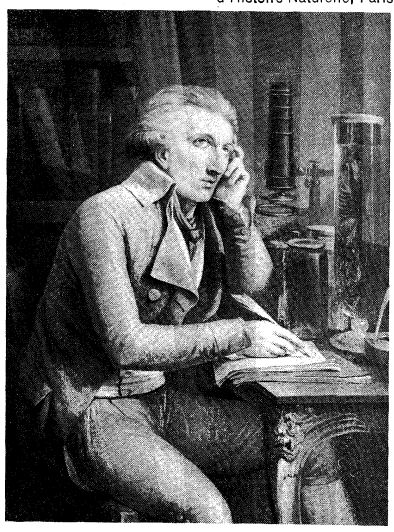
(W.G.Ib.)

Cuvier, Georges, Baron

The French zoologist Georges Cuvier established comparative anatomy and paleontology as sciences. By systematically comparing the structure of living animals with fossil remains, he showed that the past must be taken into account when studying present-day life. Although he believed that species did not change, his work provided important evidence on which the doctrine of evolution was later based.

Cuvier was born on August 23, 1769, in Montbéliard, France. He was named Jean-Léopold-Nicolas-Frédéric Cuvier. In 1784–88 he attended the Académie Caroline (Karlsschule) in Stuttgart, Germany, where he learned to dissect and studied comparative anatomy.

By courtesy of the Musee National d'Histoire Naturelle, Paris



Cuvier, portrait by Van Brae, 1798.

After graduation Cuvier served in 1788–95 as a tutor, during which time he wrote original studies of marine invertebrates, particularly the mollusks. His notes were sent to Étienne Geoffroy Saint-Hilaire, a professor of zoology at the Museum of Natural History in Paris, and at Geoffroy's urging Cuvier joined the staff of the museum. For a time the two scientists collaborated, and in 1795 they jointly published a study of mammalian classification, but their views eventually diverged.

Cuvier refused an invitation to become a naturalist on Napoleon's expedition to Egypt in 1798–1801, preferring to remain at the museum to continue his research in comparative anatomy. His first result, in 1797, was *Tableau élémentaire de l'histoire naturelle des animaux* ("Elementary Survey of the Natural History of Animals"), a popular work based on his lectures. In 1800–05, he published his *Leçons d'anatomie comparée* ("Lessons on Comparative Anatomy"). In this work, based also on his lectures at the museum, he put forward his principle of the "correlation of parts," according to which the anatomical structure of every organ is functionally related to all other organs in the body of an animal, and the functional and structural characteristics of organs result from their interaction with their environment. Moreover, according to Cuvier, the functions and habits of an animal determine its anatomical form, in contrast to Geoffroy, who held the reverse theory—that anatomical structure preceded and made necessary a particular mode of life.

Cuvier's principle of "correlation of parts"

Cuvier also argued that the anatomical characteristics distinguishing groups of animals are evidence that species had not changed since the Creation. Each species is so well coordinated, functionally and structurally, that it could not survive significant change. He further maintained that each species was created for its own special purpose and each organ for its special function. In denying evolution, Cuvier disagreed with the views of his colleague Jean-Baptiste Lamarck, who published his theory of evolution in 1809, and eventually also with Geoffroy, who in 1825 published evidence concerning the evolution of crocodiles.

Cuvier advanced rapidly. While continuing his zoological work at the museum, he brought about major reforms in education. He served as imperial inspector of public instruction and assisted in the establishment of French provincial universities. For these services he was granted the title "chevalier" in 1811. He also wrote the *Rapport historique sur les progrès des sciences naturelles depuis 1789, et sur leur état actuel* ("Historical Report on the Progress of the Sciences"), published in 1810. These publications are lucid expositions of the European science of his time.

Meanwhile, Cuvier also applied his views on the correlation of parts to a systematic study of fossils that he had excavated. He reconstructed complete skeletons of unknown fossil quadrupeds. These constituted astonishing new evidence that whole species of animals had become extinct. Furthermore, he discerned a remarkable sequence in the creatures he exhumed. The deeper, more remote strata contained animal remains—giant salamanders, flying reptiles, and extinct elephants—that were far less similar to animals now living than those found in the more recent strata. He summarized his conclusions, first in 1812 in his *Recherches sur les ossements fossiles de quadrupèdes* ("Researches on the Bones of Fossil Vertebrates"), which included the essay "Discours préliminaire" ("Preliminary Discourse"), as well as in the expansion of this essay in book form in 1825, as the *Discours sur les révolutions de la surface du globe* ("Discourse on the Revolutions of the Globe"). Cuvier assumed a relatively short time span for the Earth but was impressed by the vast changes that undoubtedly had occurred in its geologic past. His work gave new prestige to the old concept of catastrophism according to which a series of "revolutions," or catastrophes—sudden land upheavals and floods—had destroyed entire species of or-

ganisms and carved out the present features of the Earth. He believed that the area laid waste by these spectacular paroxysms, of which Noah's flood was the most recent and dramatic, was sometimes repopulated by migration of animals from an area that had been spared. Catastrophism remained a major geologic doctrine until it was shown that slow changes over long periods of time could explain the features of the Earth.

The problem of extinction of species

Just before Napoleon abdicated, in 1814, Cuvier was elected to the Council of State, and in 1817 he became a vice president of the ministry of the interior. In 1817 he also published *Le Règne animal distribué d'après son organisation* ("The Animal Kingdom, Distributed According to its Organization"), which, with its many subsequent editions, was a significant advance over the systems of classification established by Linnaeus.

Cuvier showed that animals possessed so many diverse anatomical traits that they could not be arranged in a single linear system. Instead, he arranged animals into four large groups of animals (vertebrates, mollusks, articulates, and radiates), each of which had a special type of anatomical organization. All animals within the same group were classified together, as he believed they were all modifications of one particular anatomical type.

Although his classification is no longer used, Cuvier broke away from the 18th-century idea of the scale of nature. This scheme saw all living things as arranged in a continuous series from the simplest up to man.

The increasing theoretical differences between Geoffroy and Cuvier culminated in 1830 in a public debate in the Academy of Sciences over the degree to which the animal kingdom shared a uniform type of anatomical organization, in particular, whether vertebrates and mollusks belonged to the same type. Geoffroy thought they did and that all animals, in fact, were representatives of only one type, whereas Cuvier insisted that his four types were completely distinct. At issue in their controversy was how to explain similarity and diversity in animals. Darwin's doctrine of evolution eventually clarified this question by showing that similar animals were descended from common ancestors and that diversity meant that hereditary changes had occurred.

Assessment

Cuvier's lifework may be considered as marking a transition between the 18th-century view of nature and the view that emerged in the last half of the 19th century as a result of the doctrine of evolution. By rejecting the 18th-century method of arranging animals in a continuous series in favour of classifying them in four separate groups, he raised the key question of why animals were anatomically different. Although Cuvier's doctrine of catastrophism did not last, he did set the science of palaeontology on a firm, empirical foundation. He did this by introducing fossils into zoological classification, showing the progressive relation between rock strata and their fossil remains, and by demonstrating, in his comparative anatomy and his reconstructions of fossil skeletons, the importance of functional and anatomical relationships. Cuvier died on May 13, 1832, during the first cholera epidemic to strike Paris.

BIBLIOGRAPHY. WILLIAM COLEMAN, *Georges Cuvier, Zoologist: A Study in the History of Evolution Theory* (1964), is a lucid analysis of Cuvier's zoological ideas, with a partial bibliography. A general introduction to Cuvier's work is found in ERIK NORDENSKIOLD, *Biologins historia*, 3 vol. (1920–24; Eng. trans., *The History of Biology*, new ed., 1935). RENE DUJARRIC DE LA RIVIERE (ed.), *Cuvier, sa vie, son oeuvre, pages choisies* (1969), contains a short biography, with selections from his works, and a partial bibliography. A collection of important papers on Cuvier's life and work was published for the centenary of his death in the *Archives du Muséum National d'Histoire Naturelle*, series 6, vol. 9 (1932).

(Ed.)

Cycling

Cycling is the use of the bicycle for sports. The bicycle, improved throughout the years from the time of the first official contests set up in 1868, and the sport of cycling have both profited from progress in design and technology. For the history of the invention and development of the bicycle and descriptions of various types, see the article BICYCLE.

Cycling has become a common means of transport, and, as a sport, with its rules and its classic races, it has continued to grow throughout the world. Its popularity, in spite of the constant improvement of public and private means of motorized transport, remains great, in the United States as well as in Europe. In the flat countries, especially, cycling is a common means of transport. In countries with varied terrain, such as France, cycling still remains popular and has become a traditional sport, in spite of the increasing use of motorbikes.

History of the sport. Cycling as a sport is more than 100 years old. Its exact birthdate could be fixed as May 31, 1868, when the first velocipede race was organized by the Oliver brothers, managers of the Michaux factory, which made the first real bicycles by attaching pedals to the axle of the front wheel. The race, held in the Saint-Cloud Park near Paris, was won by James Moore. The next day, according to the records, the first cycle race in England was held at Hendon, Middlesex. That same year Moore also achieved victory riding a *grand-bi*, or boneshaker, which has a front wheel much greater in diameter than the rear one, in the first town to town cycle race ever held, 135 kilometres (84 miles) from Paris to Rouen. Also in 1868 the first cycling periodical, *Le Vélocipède*, appeared in Paris; and the following year the magazine *Le Vélocipède Illustré* was founded.

The first races

Almost two decades later, in 1888, the pneumatic tire for bicycles was invented and bicycle riding became more comfortable and stable and thus a more popular sport.

The *grand-bi* vogue disappeared little by little. Although it had evolved into the more graceful wire-wheeled "ordinary," the vehicle was dangerous because of its instability, and it required an acrobat's skill to ride. Toward 1880 the rear-wheel-propelled bicycle, using a chain transmission from pedals between the wheels, was successfully produced in England. The new machine was called the "safety" because it was much more stable than the *grand-bi*. In 1891 the first races on bicycles with pneumatic tires were held. The first race, from Bordeaux to Paris (580 km [360 mi]), was reserved for British amateurs and won by G. Mills. Later the same year a race from Paris to Brest and back to Paris (1,200 km [750 mi]) was won by a Frenchman, Charles Terront, in 72 hr 22 min.

At the same time, the first cycle-racing tracks were built in New York and Paris. New York City in 1891 held the first six-day race, which was won by Charlie Miller, who covered 3,368 km (2,093 mi) in six days. Not only were *grand-bi* cycles ridden but the same person had to ride the entire race—a really superhuman effort. In 1894 the six-day race was transformed into a team event (American-type relay).

Soon the great modern, classic contests began: the Tour de France (1903) initiated by Henri Desgrange, manager of *L'Auto* (forerunner of the sports daily *L'Équipe*); and such town-to-town races as Paris–Robaix, Paris–Tours, and Milan–San Remo.

Organization and development. The cycling sport, born in France, rapidly spread to Italy, Belgium, Spain, Switzerland, Germany, Great Britain, and The Netherlands and then to Australia and the United States.

Cycling unions

In 1878 the Bicycle Union was founded in England; it was only in 1900 that a single international cyclist union was formed, the Union Cycliste Internationale, which gave cycling its modern rules and races.

In 1965 two different organizations were created: the Fédération Internationale du Cyclisme Professionnel and the Fédération Internationale Amateur de Cyclisme. This professional–amateur separation had been demanded by the International Olympic Committee so that cycling could remain in the Olympic Games. More than 100 countries have joined the amateur federation. Only about ten, all in western Europe, belong to the professional federation. The eastern European countries do not recognize professionalism in any sport and have not joined.

The World Championships are organized each year by the International Cyclist Union, which also is in charge

of Olympic Games competition; national championships are controlled by the various national federations. All other cycling contests are organized by private firms or promoters. Most often they are financed by the press, and most of the promoters belong to the Association Internationale des Organisateurs de Courses Cyclistes (International Association of Organizers of Cycle Competitions).

Outstanding performers and performances. Since the beginning of the 20th century, legends have grown up around the champions of the great international events. Among the outstanding performers in touring races have been M. Garin (France), winner of the first Tour de France, in 1903; P. Thys (Belgium), who won that event three times (1913, 1914, 1920); and Fausto Coppi (Italy), who won the Tour de France twice (1949 and 1952) and the Tour of Italy five times (1940, 1947, 1949, 1952, 1953). Coppi also was a world champion in touring and track racing, and for a long time he held the one-hour distance record. Louison Bobet of France had three victories in the Tour de France (1953, 1954, 1955), and Jacques Anquetil, also of France, had a record five victories in this race (1957, 1961–64). In the 1960s and 1970s the Belgian Eddy Merckx dominated the international scene as no one had before.

But no doubt the greatest prestige still belongs to Italy's Coppi, who died prematurely in 1960. He excelled in any branch: track pursuit races, touring races, and against-the-clock races (one racer at a time runs the event and is clocked). Merckx, who took Coppi's place as a popular figure, was most of all a road racer. He won all the classic races within three years—the Tour de France, the Tour of Italy, Paris–Roubaix, and others.

Outstanding track performers in distance covered in 60 minutes include the Frenchman Henri Desgrange, first one-hour distance record (35.325 km [21.95 mi] in 1893); the American W.W. Hamilton, first to surpass 40 km (40.781 [25.346 mi] in 1898); the Italian G. Olmo, first over 45 km (45.090 [28.02 mi] in 1935); and the Danish Ole Ritter (48.65 km [30.24 mi] October 16, 1968, in Mexico City). Famous sprinters (speed racers) also deserve mention: the Dane T. Ellegaard, six times the world champion from 1901 to 1911; the American Frank Kramer, who held the U.S. title from 1901 to 1916 and won it again in 1918 and in 1922; the Belgian J. Scherens, seven times world champion; and the Italian Antonio Maspes, eight times world champion. The fastest speed achieved on a bicycle was 204.73 kph (127.243 mph), by Jose Meiffret (France), July 16, 1962, on the German Autobahn from Freiburg, behind a car (see below *Events*).

Competition. The national federations in each country organize cycling competitions as they wish, according to the international rules. Most of them set up separate contests for men and women. The most widely used age categories for amateurs are seniors (over 20), juniors (18–20), younger juniors (16–17), adolescents (14–15), and youngsters, or benjamins (12–13). Competitions usually are categorized as road races, track races, or cyclo-cross (cross-country) races. Road races may include town to town or circuit-team races, individual or team races against the clock, and stage races (run in stages over several days) such as the classic tours. Track events include sprints, or speed races; pursuit races; middle-distance races behind a motorcycle; and races against the clock. Cyclo-cross or cross-country races usually cover rough terrain—often crossing ditches, streams, or other water hazards—in which racers may change off walking and cycling.

World cycling championships are organized each year for amateurs as well as professionals. The Olympic events, strictly for amateurs, include an individual road race and a 100-km (62-mi) team race, a 4,000-metre team pursuit race, a 1,000-m time trial, a 1,000-m match sprint (or scratch sprint), and a 2,000-m tandem race. A cyclo-cross world championship is also organized each year (20 to 24 km [12 to 15 mi]).

Stage races and the classics. The best known stage races are the Tour de France (race around France), created in 1903; the Tour of Italy (1909); the Tour of Spain (1935); the Race of the Future (France), reserved for amateurs; and the Peace Race (Czechoslovakia, Poland, East Germany), also for amateurs. Also important are the Tour of Egypt and the Tour of Tunisia.

Town to town races include Paris–Roubaix, Milan–San Remo, Bordeaux–Paris, and Paris–Tours.

The Tour de France

The Tour de France went through some difficult first years after its inception in 1903. The racers were hesitant, fearing the long stages 500 to 700 km (300 to 450 mi) and the total distance of 5,200 km (about 3,200 mi). But gradually the formula improved and the event was accepted. In 1906, for the first time, racers went over mountain passes as high as 1,450 m (4,750 ft). Later they had to pedal up slopes 2,600 m (8,500 ft) high in the Pyrenees and the Alps. At that time, around 1910–12, mountain roads were extremely dangerous, and numerous accidents occurred.

In 1930 national team scores were added to the tour. These were replaced in 1962 by trademark team scores, for advertising reasons.

The sport of racing. *Equipment.* The road-racing bicycle is distinguished from a training or touring bike by its lightweight-alloy frame, its lack of mudguards, its double clutch and ten gears, its lighter wheels with narrow rims and thinner, more numerous spokes, and its slimmer, tubular tires. The bike's gearing depends on the racer's height, weight, and power. Usually, the bike travels 4.7 to 7.93 m (15.4 to 26 ft) for each turn of the pedal, depending on the gear used.

The track bicycle is even lighter than the road racer, for it has neither brakes nor gears. Its frame is lower, its crankshaft higher to avoid touching the ground when turning; the tires are extralight. It travels 7 to 8 m (23 to 26 ft) for each turn of the pedal. There is no rear wheel clutch but one fixed sprocket.

Events. Track events include the sprint, or speed, races, 1,000 m (about 1,100 yd), run individually; and, for amateurs only, tandem events, 2,000 m (2,200 yd), run by teams of two racers on tandem bikes. Time trials are run individually, 1,000 m, against the clock from a standing start (Olympic Games) or from a flying start in some events.

Pursuit, or chasing, races

Pursuit, or chasing, races are 5,000 m (5,500 yd) for professionals and 4,000 m (4,400 yd) for amateurs. In pursuit events the individual racers (in team events, the two teams) start on opposite sides of the track and chase each other. The winner is the racer or team that has closed on or overpassed the other within a fixed distance (in the Olympics, 4,000 m).

Other track events include individual races of 5, 10, or 20 km (3, 6, or 12 miles) in which each racer is given points every kilometre, according to his place; eliminations, in which the last racer across the finish line every lap is eliminated; and handicaps, in which the competitors considered to be the best give up some ground to their opponents. The *omnium* is a combination including 10 km, 1 km from a flying start, a 5-km pursuit, and 20 km behind a motorbike. Points are awarded for each event toward an overall score.

In so-called American races, teams of two riders each take their turn whenever they wish. The result depends either on the distance covered or on points gained.

In dragging races, the racer is pulled along by the air current created behind either a motorbike, a motor scooter, or a motorcycle. The bicycles used are fitted with small front wheels and can travel a long distance with each turn of the pedal.

Speed record attempts are made at 5, 10, 50, and 100 km (3, 6, 12, 30, and 60 mi), as well as for distance covered in one hour.

Road events include races of 240 to 280 km (150 to 175 mi) for professionals and 180 to 220 km (110 to 140 mi) for amateurs. The equivalent event in the Olympic Games is the 100-km team race. In addition to the town-to-town and stage events, or tours, discussed above, road races include races against the clock, such as the annual Grand Prix des Nations, 140 km (90 mi), held in the Paris district.

Principles of technique and strategy. The most important point in road-racing technique is the posture of the cyclist on his bicycle. His hands firmly grip the foremost part of the handlebar, ensuring a spinal curve as aerodynamic as possible and preventing tiredness. Training includes exercises for long-distance endurance, body building, and interval training (alternation of fast racing and resting periods so as to improve the cardiac rhythm and resistance to strain). On the road, strategy is mostly a matter of teamwork.

Training for endurance

In sprint races the opponents keep close to each other during the first 700 or 800 m. The oncoming racer usually enjoys advantages over the leaders because he is more or less sucked along by the draft of air his opponent (or opponents) creates. This event can give rise to long periods when no racer wants to precede the other but tries to force less skillful opponents to pass him by.

Usually sprinters and, to a certain extent, chasers have enlarged hearts (more than 1,000 cubic centimetres [60 cubic inches], as compared with the average of 625–650 cubic centimetres [38–40 cubic inches]) and use the interval-training method: repeating short distances covered at full speed, interrupted by short periods of rest.

The drug problem. In cycling the use of drugs has been a most preoccupying problem, not only on the road but also on the track. For a long time, bicycle racers, whether amateurs or professionals, have had set ideas: they think that a cyclist cannot put forth an intense or lasting effort without the aid of drugs, such as antifatigue drugs, which project beyond the usual limits the beginning of exhaustion, or stimulants, which add a push of energy. Professionals are tempted to overuse such drugs —especially in stage races—but many are dangerous (principally the amphetamines) and leave traces in the organism.

As a result of an antidrug campaign within the sport, a wide list of prohibited drugs has been established. The use of banned drugs now can be detected by gaseous-state chromatography. Even amphetamine can be detected, although its by-products are numerous.

In 1966, the French Parliament promulgated a law aiming at repressing the use of stimulating drugs during sport contests. Belgium and Italy have also voted laws to that effect. But, in spite of the severe punishments inflicted—suspension, license cancellation, fines—newly discovered drugs are used that are unknown to experts in charge of controls and are not discoverable by analysis.

Thus drug abuse continues to be a problem and has caused many fatal accidents, such as that of the Briton Tom Simpson, during a stage of the Tour de France in 1967.

Cycling as recreation. In Europe, touring by bicycle and touring clubs have enjoyed wide popularity. Hundreds of thousands of people have joined such clubs, especially in France, Belgium, Italy, and England. Certificates are awarded for covering various distances: 100, 200, 300 km (60, 120, 180 mi). Members ride in rallies, participate in youth camps, and stay in specially booked hotels. France's touring club has an uncommonly active cyclist branch. It constantly organizes excursions of various length in tourist areas and regions with varied terrain.

Activities of touring clubs

Touring is a sport that knows no borders. In Europe, meetings of cyclists from various countries often attract participants who ride international routes (*e.g.*, through France, Belgium, Italy) to attend. Since 1959, a Tour de France for excursionists has been held annually. A national certificate of cyclo-touring has been created: in this type of rally, the rider has to find his way with the help of a map.

The touring bicycle in Europe and North America has been much improved—it is lighter and fitted with gears. Once heavy and fitted with balloon tires, it now looks rather like a racing bike.

Touring by bicycle has also increased in the United States. Famed doctors have asserted that cycling is the healthiest sport for the human body, and 50,000,000 Americans practice cycling sports more or less regularly. With the financial backing of important firms, the Bicycle Institute of America, whose headquarters are in New York, has converted many people of all ages to the pastime.

It is not surprising to find that annual sales of bicycles have risen in the United States from 200,000 in 1932 to more than 5,000,000 in the late 1960s. Hundreds of bicycle touring clubs have been set up, and several states and many municipal and other public park systems have established bicycle paths. Agreements have been reached with youth hostels and with managers of national parks and reserves, where thousands of miles of bicycle paths have been built.

In urban areas both in the United States and Europe, the growth of cycling—as both recreation and transportation—in the early 1970s seemed to offer some promise of eventually relieving the strangling congestion of automobile traffic.

BIBLIOGRAPHY. Further information on the history and other aspects of cycling both as a sport and a recreation may be found in the following references: H. MOORE, *Complete Cyclist*, 5th ed. (1957); A.L. PULLEN, *Cycling Handbook*, 2nd ed. (1953); C.L. FREESTON, *Cycling in the Alps* (1900); G. and L. FRANKEL, *Bike-Ways: 101 Things to Do with a Bike*, rev. ed. (1968); K. BOWDEN, *Cycle Racing* (1958); A.J. PALMER, *Riding High* (1956); E. SHERIDAN, *Wonder Wheels* (1957); H. GRIVELL, *Australian Cycling in the Golden Days* (1953); R. ALD, *Cycling: The Rhythmic Respiratory Way to Physical Fitness* (1968); W. STEPHENSON, "When the World Went Bicycle Crazy," *MacLean's Magazine*, 69:24–25 (1956).

(Ro.P.)

Cyclones and Anticyclones

Cyclones and anticyclones are large rotary wind systems that move across most areas of the earth outside the equatorial belt. Cyclones usually are associated with belts of cloud and rain or snow, whereas the anticyclones, which generally have lighter winds, are typically free from precipitation. The diameter of these wind systems ranges from less than 1,000 kilometres to 3,000 or 4,000 kilometres. In the Northern Hemisphere the cyclonic winds rotate in a counterclockwise direction, and the anticyclonic winds rotate in a clockwise direction. In the Southern Hemisphere the directions of rotation are opposite to those in the Northern Hemisphere (see WINDS AND STORMS).

Cyclones of a somewhat different character occur nearer to the Equator, forming in latitudes 10 to 15 degrees north and south over the oceans. They generally are known as tropical cyclones but, more specifically, as hurricanes in the Atlantic and Caribbean, as typhoons in the western Pacific and China Sea, and as willy-willies off the coasts of Australia. These storms are of smaller diameter than the extratropical cyclones, ranging from 100 to 500 kilometres in diameter, and are accompanied by winds that sometimes reach extreme violence. These storms are more fully described in the article HURRICANES AND TYPHOONS.

The cyclones and anticyclones that form outside the equatorial belt may be regarded as large eddies in the broad air currents that flow in the general direction from west to east around the middle latitudes of both hemispheres. They are an essential part of the mechanism by which the excess heat received from the sun in the equatorial belt is conveyed toward higher latitudes. These higher latitudes radiate more heat to space than they receive from the sun, and heat must reach them by winds from the equatorial belt if their temperature is to be maintained. If there were no cyclones and anticyclones the north–south movements of the air would be much more limited, and there would be little opportunity for heat to be carried poleward by winds of subtropical origin. Under such circumstances the temperature of the equatorial regions would increase and the polar regions would cool; the temperature gradient between them would intensify.

Temperature and pressure

Strong horizontal gradients of temperature are particularly favourable for the formation and development of cyclones. In the absence of cyclones and anticyclones the

temperature difference between pole and Equator would tend to build up until it was sufficiently intense to generate new cyclones. The new cyclones themselves, however, would tend to reduce the temperature difference. Thus, the wind circulation on the earth represents a balance between the effect of solar radiation in building a temperature difference between pole and Equator, and the effect of cyclones, anticyclones, and other wind systems in destroying this temperature contrast.

Cyclones and anticyclones are readily recognized on weather maps because of the centres of low and high pressure that accompany them. The atmospheric pressure is observed by means of barometers at several hundreds of weather stations throughout the world and is corrected to mean sea level. These pressures are plotted on weather maps, and lines of equal pressure (isobars) are drawn on them. As a consequence of the earth's rotation, any large-scale air current that is more or less steady during a period of several hours must flow with higher pressure on the right of the current than on the left (looking downstream) in the Northern Hemisphere. For this reason the rotary motion of the winds in cyclones and anticyclones appears on weather maps as sketched in Figure 1 in relation to the isobars. In the Southern Hemisphere the direction of the circulation of the winds in both systems is reversed.

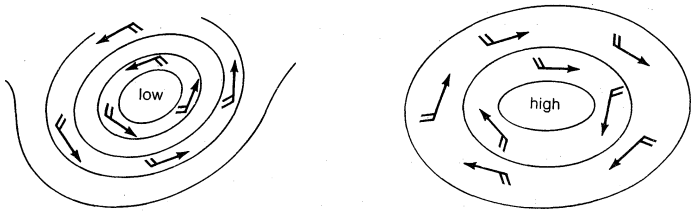


Figure 1: (Left) Cyclone (or depression). (Right) Anticyclone.

Because the cyclone is necessarily an area of lowered atmospheric pressure, with the lowest pressure at the centre of the wind circulation, cyclones are commonly known as depressions.

It has been recognized since weather maps were first drawn at the beginning of the 19th century that depressions are associated with bad weather. The reason has become apparent only recently. In order to understand this association it is necessary to think of the atmosphere (*q.v.*) in depth and to consider what happens in a region where the air flows together (converges) at a level near the ground, then rises as a consequence of the convergence and spreads out (diverges) at some higher level, usually in the upper troposphere; that is, at 8 to 10 kilometres (26,000 to 33,000 feet) above the ground (see Figure 2).

Convergence and divergence of air

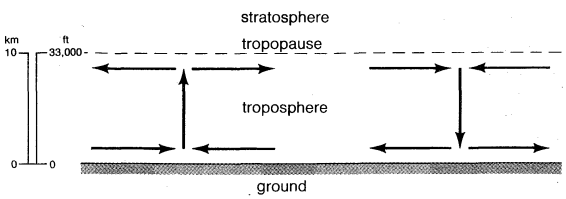


Figure 2: Formation of (left) cyclone and (right) anticyclone.

The air that is converging near the ground is rotating with the earth, as is the rest of the atmosphere; but as it is drawn inward into a smaller area during the convergence, its rate of rotation will increase in a way similar to that in which the rate of rotation of a weight twirled on the end of a string increases as the string is pulled in and the angular momentum of the weight is conserved. The region of convergence thus appears as an area of increased rotation of the winds. This increase in the rotation is the same as the rotation of the earth, counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere, and so the region of convergence appears as a cyclone or depression.

Figure 2 indicates that the convergence of the air in the lower atmosphere necessarily is accompanied by upward motion. As the air rises, it is subject to less pressure from the weight of the air above and it expands. Expanding air cools and is unable to carry as much water vapour as before; part of the water vapour therefore is condensed as cloud and rain. Thus the convergence of the air in the lower layers leads simultaneously to the generation of the cyclonic circulation of the winds and to the bad weather with which cyclones are associated. In order that the proper relation between the wind and the pressure field be maintained, the atmospheric pressure must fall at the centre of the cyclone as it forms. This is accomplished by a slightly greater outflow of air aloft than the inflow near the ground.

The process of formation of an anticyclone is precisely the reverse of that of a cyclone or depression. Anticyclone formation is associated with downward motion of the air and spreading out near the ground. The downward motion compresses, warms, and dries the air, and no thick rain clouds are likely to form. The spreading out slows down the rotation of the air until its rate is less than that of the earth itself. An observer, who of course is rotating with the earth, sees the anticyclone rotating in the opposite sense, namely, clockwise.

FORMATION AND STRUCTURE OF CYCLONES

The middle latitudes of the two hemispheres each contain a broad, generally westerly current of air in which the wind speed increases with height above the ground. This increase of wind with height is associated with the general gradient of temperature from Equator to pole. Cold air is denser than warm air and pressure falls more rapidly with increase in height near the poles than near the Equator. It follows that the pressure difference between the pole and the Equator is greater aloft than near the ground and, because the wind speed is related to the pressure difference across the air current, the wind speed also increases with height in the atmosphere.

The westerly current does not run smoothly, and there are areas where the temperature contrasts are increased. These are areas where the wind increases particularly rapidly with height and in which convergence in the lower layers and consequent upward motion are especially likely. These are the regions that give birth to depressions or cyclones.

Polar-front theory

The association between strong horizontal temperature contrasts and the formation of depressions was first noted by Scandinavian meteorologists working in Bergen about 1920. They developed the polar-front theory of cyclones. They recognized that strong temperature contrasts develop between air masses of different origin, and the division between such air masses is often sharp enough to be indicated on the weather map by a line and be known as a front. At the front the colder air mass lies below the warmer air, forming a wedge. The frontal surface separating the two air masses has a slope of between 1 in 50 and 1 in 200.

Depressions often are seen to form on or in the vicinity of fronts, the first sign being a wavelike distortion of the front. The principal front identified by the meteorologists of the Norwegian school separated air of polar origin from warmer subtropical air and was known as the polar front. It originally was envisaged as extending more or less continuously around the middle latitudes of the earth, but in fact there are usually many substantial breaks in it.

Life cycle of a cyclone

The various stages in the life cycle of a depression on the polar front are illustrated in Figure 3. The front is indicated by the thick line to which triangles or semicircles are attached. The line is placed where the frontal surface separating the two air masses intersects the ground, and the symbols are placed on the side toward which the front is moving. The triangular symbols are used where the cold air is replacing the warm air; the front is advancing as the leading edge of the cold air mass and is known as a cold front. Semicircular symbols are used where the warm air is replacing cold air; here the

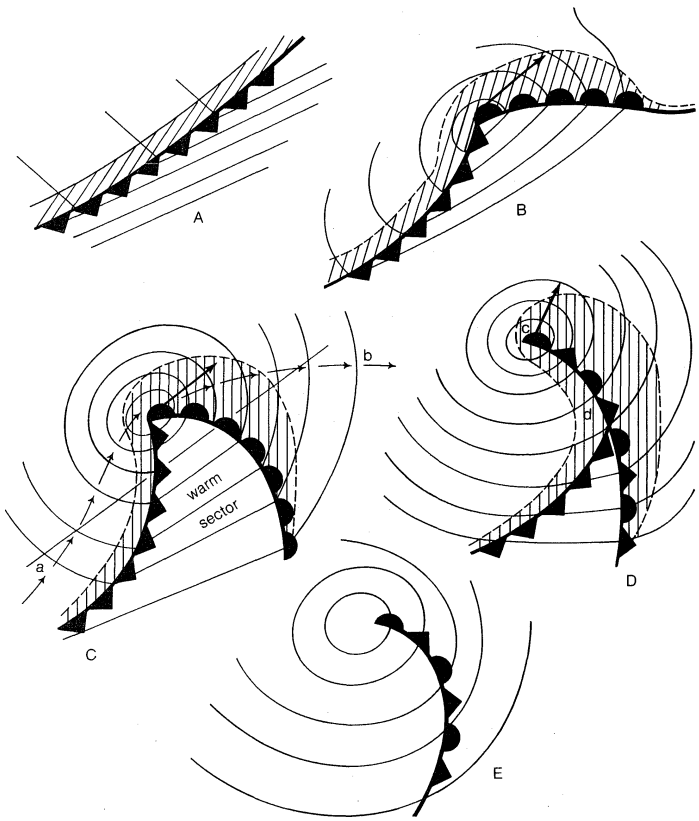


Figure 3: (A) Undistorted front. (B) Wind circulation creates distortion of front. (C) Distortion increases and a warm-sector depression is formed. (D) Development of occluded front. (E) Final stage of a cyclone.

front is the leading edge of the warm air and is known as a warm front.

Figure 3A represents an undistorted front, initially moving slowly southeastward toward the warm air as a cold front. An early stage in the formation of a depression is shown in Figure 3B. The circulation of winds around the depression has created a wavelike distortion of the front. The thin lines denote the isobars, and the lowest pressure, or the centre of the depression, is located at the tip of the wave. At this stage the depression commonly is moving at a speed of from 40 to 90 kilometres per hour in the general direction of the arrow. The area affected by rain or snow is indicated by hatching.

Figure 3C represents a somewhat later stage in the development of a depression. The distortion of the front has increased and the winds have grown stronger. At this stage it is a typical warm sector depression, so called because of the sector between the warm front and cold front that contains warm subtropical air. The warm sector often is a region of much cloudiness, with drizzle on coasts and hills, and sometimes produces rain. Less commonly, the warm sector may contain relatively dry air of continental origin and be a region of high temperatures and sunny skies.

It is helpful to consider the structure of the atmosphere at this stage in terms of a vertical cross section along the line ab. The vertical section cuts through the cold and warm front surfaces, as indicated by the sloping lines. The clouds at the cold front and the warm front lie above the frontal surfaces as indicated, and rain falls through the lower parts of the frontal slope.

The horizontal temperature gradients that accompany a front also are associated with an increase of wind with height as indicated earlier, and a rather narrow band of strong winds is found at a height of about 10 kilometres above the ground. These are the jet streams (*q.v.*). At the stage of depression development illustrated in Figure 3C the jet lies as a somewhat distorted current along the line indicated by the double arrows. The jet stream intersects the cross section ab obliquely above the upper parts of the frontal surfaces.

As indicated earlier, the development of a depression is always accompanied by the convergence of the air horizontally in the lower atmosphere. By the time the mature stage has been reached, the convergence has squeezed the warm air off the ground near the centre of the depression, and the arrangement of the fronts is as shown in Figure 3D. The front from c to d is known as an occlusion. Although the original warm air is lifted off the ground, there may still be some temperature contrast between the cold air mass ahead of the depression and that circulating round the back of it. A front may remain between these two cold air masses and is indicated on the weather map by the symbols for an occlusion, namely, alternating triangles and semicircles.

In the final stage of a cyclone most of the temperature contrasts that generated it have been lost, and its structure is as represented in Figure 3E. There still is a belt of thick cloud and rain along the line of the old occlusion running obliquely from the central region to the periphery of the cyclone. There also are other, narrower spiral bands of cloud that extend outward from the centre; these enable the final stages of such cyclones to be easily recognized on the satellite pictures of cloud distributions. At this stage, the cyclone is usually slow-moving or stationary.

Although the main belts of bad weather in depressions are associated with fronts, the circulation of the depression also includes showers and other local storms in the cold air behind the cold front, where the cold air mass is moving southward and is being rapidly heated from below. During the later stages of the depression these showers tend to spread around the centre in all directions. There may also be some rain in the warm sector in the earlier stages of the cyclone. At all stages the general convergence of air into the cyclone, in addition to the ascent of air, encourages the growth of showers and other precipitation.

Not all depressions and cyclones form on the polar front in the manner indicated. Some are induced by the effect of mountain barriers on winds crossing them, appearing first on the lee side of the mountain range. Tropical cyclones may also form in the absence of any associated fronts.

Apart from tropical cyclones, the initiation of which is still not understood, most cyclones in middle latitudes form in the presence of horizontal temperature contrasts. The wind circulation and convergence lead to the formation of fronts in the vicinity of the depression with a configuration rather similar to that illustrated in Figure 3, even when a front is not present initially. A glance at a series of daily weather maps will show that each cyclone has its individual characteristics, and only broad generalizations in regard to their structure are justified. Nevertheless, recent progress in meteorology is such that calculations can be based on the mathematical equations describing the air motion and the physical processes involved in the heat and water balance. Given a reasonably comprehensive description of the initial state of the atmosphere, such calculations can describe many of the detailed features of the individual cyclones as they develop. With this more specific information, the weather forecaster is becoming much less dependent on idealized descriptions.

ANTICYCLONES AND THEIR FORMATION

For various reasons, the formation of anticyclones is usually a slower process than the formation of depressions; anticyclones tend to be larger in area than cyclones and to persist longer. Whereas the upward motion in depressions is concentrated at the fronts and typically reaches ten centimetres per second, the downward motion in anticyclones is slower, perhaps one to three centimetres per second, and is spread over wider areas. To a great extent, the downward motion in the anticyclones can be regarded as a compensation of the upward currents in the neighbouring depressions.

Types of anticyclones

Most actively developing anticyclones form in the region of cold air behind a depression as it moves away and before the next depression advances; these are known as cold anticyclones. A result of the downward air motion in an anticyclone, however, is compression of the descending air, and as a consequence of this compression the air is warmed. Thus after a few days the air comprising the anticyclone at levels two to five kilometres above the ground tends to become warm for its level, and the anticyclone is transformed into a warm anticyclone. The air near the ground cannot descend and is not much affected by this process.

Warm anticyclones move slowly, and depressions are diverted around their periphery. During their transformation from cold to warm anticyclones they usually move out of the main belt followed by cyclones in middle latitudes, often amalgamating with the quasi-permanent band of relatively high pressure that is found in both hemispheres around latitude 20 to 30 degrees—the so-called subtropical anticyclones. On some occasions the warm anticyclones remain in the belt normally occupied by the mid-latitude westerly winds. The normal tracks of cyclones are then considerably modified; depressions are either blocked in their eastward progress or diverted to the north or south of the anticyclone. Anticyclones that interrupt the normal circulation of the westerly wind in this way are known as blocking anticyclones. They frequently persist for a week or more, and the occurrence of a few such blocking anticyclones may dominate the character of a season. Blocking anticyclones are particularly common over Europe, the eastern Atlantic, and the Alaskan area.

The descent and warming of the air in an anticyclone might be expected to lead to the dissolution of clouds and the absence of rain. This occurs often, particularly in summer. Since the air near the ground cannot take part in the descent, it may become stagnant. In winter the ground cools and the lower layers of the atmosphere also become cold. Fog may be formed as the air is cooled to its dew point. Under other circumstances the air trapped in the first kilometre above the surface may pick up moisture from the sea or moist surfaces, and layers of cloud may form from near the ground to a height of about one kilometre. Such layers of cloud can be persistent in anticyclones (except over the continents in summer), but they rarely grow thick enough to produce rain. If precipitation occurs it is usually drizzle or light snow.

Anticyclones are thus regions of clear skies and warm sunny weather in summer; at other times of the year cloudy and foggy weather may be more typical. Winter anticyclones produce colder-than-normal temperatures at the surface, particularly if the skies remain clear. Anticyclones are responsible for periods of little or no rain, and such periods may be prolonged in association with blocking highs.

DISTRIBUTION OF CYCLONES AND ANTICYCLONES

Cyclones are primarily a feature of the midlatitude belts of the two hemispheres. In the Southern Hemisphere, where most of the earth's surface is covered by oceans, the cyclones are distributed fairly uniformly through the various longitudes. Typically, cyclones form initially in latitudes 30 to 40 degrees south and move in a generally southeastward direction, reaching maturity in latitudes around 60 degrees. Thus the Antarctic continent is usually ringed by a number of mature or decaying cyclones, and the belt of ocean from 40 to 60 degrees south is a region of persistent, strong westerly winds that form part of the circulation to the north of the main cyclone centres. These are the "roaring forties," where the westerly winds are interrupted only at intervals by the passage southeastward of developing cyclones.

Anticyclonic development over the southern ocean occurs only occasionally, and anticyclones are transient features of the weather regime.

In the Northern Hemisphere, the continents extend from the Equator to the Arctic, and mountain belts, such as the Rocky Mountains, interfere with the midlatitude air currents. As a result, there is an important geographical variation in the occurrence of cyclones and anticyclones, and certain tracks are particularly favoured by the wind systems.

The main cyclone tracks lie over the oceans. This is partly because the surface of the oceans is smoother than that of the land and offers less resistance to the stronger winds around depressions. But the water evaporated from the ocean also provides heat when it is condensed into rain within the rain belts of depressions, and the latent heat thus released is an effective energy source for the cyclone. The arrangement of cyclone tracks is such that storms often reach maturity in the vicinity of Iceland or the Aleutians. These regions have a higher frequency of cyclonic weather than others; their climate is notably disturbed and rainy.

Anticyclone tracks are probably less clearly defined than the tracks of cyclones. An important feature of the general circulation of the earth's atmosphere, however, is the existence of a persistent circulation in which air rises in a belt near the Equator, spreads out poleward into both hemispheres in the upper troposphere at about 10 to 15 kilometres (33,000 to 49,000 feet) above the ground, and descends again in latitudes 25 to 30 degrees north and south. As previously explained, such descending air spreading out near the ground leads to the development of anticyclones. The subtropical anticyclones thus formed are found over the oceans throughout the year and are often well developed over the continents during winter months. The high temperatures reached over the continents in summer in these latitudes, however, lead to a different monsoonal type of air circulation, and the anticyclone belt is broken over the continents in summer.

HAZARDS OF CYCLONIC STORMS

Although the strongest winds occur in tropical cyclones, in which speeds of 60 metres per second are sometimes reached, exceptionally strong winds and considerable damage also can be caused by extratropical cyclones. Winds may reach 30 metres per second, with gusts up to 50 metres per second. In certain areas severe damage also can be caused along coastlines by the rise in the water level produced by the storm winds. At the end of January 1953, 307 lives were lost, 156,000 acres of land were flooded, and 25,000 houses were damaged along the east coast of England; again, in February 1962, at least 343 people died in Hamburg as a result of such a tide produced by cyclonic winds.

The weather hazards associated with anticyclones are of a different nature. The stagnant air occurring in anticyclones in winter sometimes leads to persistent fogs, but in industrial areas the greatest nuisance and hazard is probably due to smog, a combination of industrial pollutants and stagnant low-lying air. In December 1952, cold foggy air was stagnant for five days over London and 4,000 more people died than is usual at this time of the year. This increased death rate was largely attributed to the effect of the polluted air on persons already suffering from bronchial illness. The pollution was trapped near the ground beneath an anticyclone that extended westward from continental Europe.

BIBLIOGRAPHY. H.L. CRUTCHER and O.M. DAVIS, *U.S. Navy Marine Climatic Atlas of the World*, vol. 8, *The World* (1969), covers marine areas chiefly, but contains detailed worldwide maps of pressure and winds at sea level for each month, and seasonal maps of wind and topography for the main pressure surfaces up to 200 millibars (about 12 km., 40,000 feet).

General accounts: R.G. BARRY and R.J. CHORLEY, *Atmosphere, Weather and Climate* (1968), an elementary introduction to meteorology and to the atmospheric circulation; F.K. HARE, *The Restless Atmosphere*, 4th ed. (1966), a simple account of the atmosphere's circulation; R.C. SUTCLIFFE, *Weather and Climate* (1966), an authoritative non-technical account of modern meteorology and climatology by a British pioneer of modern dynamic studies.

Theoretical works: E. LORENZ, *The Nature and Theory of the General Circulation of the Atmosphere* (1967), the most

authoritative rigorous treatment of the earth's general circulation, both descriptive and theoretical; S. PETTERSSEN, *Weather Analysis and Forecasting*, 2nd ed., vol. 1, *Motion and Motion Systems*, vol. 2, *Weather and Weather Systems* (1956), a technically advanced treatment of dynamic meteorology, in relation especially to the study of winds, storms and their prediction, by the leading exponent of upper air analysis; B. SALTZMAN (ed.), *Selected Papers on the Theory of Thermal Convection* (1962), reprints of 25 classic papers on the various scales of convective motion in the atmosphere.

(J.S.S.)

Cyclothems

A cyclothem is a complex, but orderly, succession of several lithologically distinctive sedimentary strata. These accumulated during one of several repeated depositional cycles, of a kind that was characteristic of part of the Carboniferous System (280,000,000 to 345,000,000 years ago) in many coalfields, particularly those of the Northern Hemisphere (see also CARBONIFEROUS PERIOD, UPPER; CARBONIFEROUS PERIOD, LOWER).

Historical background

Some features of repeated sedimentary successions in strata of various ages were noted and reported upon during the latter part of the 19th century in both North America and western Europe. Most of these occurrences were repetitions of large dimensions that involved strata of a considerable part of a geologic system. The economic importance of coal beds, however, caused much attention to be directed to Upper Carboniferous or Pennsylvanian stratigraphy, and some features of more restricted sedimentary repetitions were recognized in England and later in Nova Scotia and Ohio. Subsequently, the earliest clear description of what are now termed cyclothems was included in a geological report published in 1912 and devoted to a restricted area in western Illinois adjacent to Peoria. There, a peculiar stratigraphic succession, including coal, was repeated with remarkable fidelity four times.

Little note was taken of this observation until 1926, when a similar repeated succession was distinguished in the Wabash valley of eastern Illinois and western Indiana. At about this time interest in Pennsylvanian stratigraphy revived in the central United States and programs of detailed fieldwork were begun. It was demonstrated clearly that repeated and related sequences of sedimentary strata characterize important parts of the Pennsylvanian System as this is developed from at least western Pennsylvania to northern Texas. As interest in Pennsylvanian stratigraphy expanded, a special term to identify these successions seemed desirable and cyclothem was proposed in 1932. This name gained immediate acceptance and it soon was in general usage in America and abroad.

NATURE OF CYCLOTHEMS

Cyclothems vary greatly in their lithologic development, and some are so different in their makeup that their relationships might not be suspected if intermediate kinds and intergradation did not occur. Cyclic sediments that result from lacustrine processes are termed varved deposits (*q.v.*) and are not considered in this article. An intermediate type generally is considered to be most typical; it is well developed in Illinois and adjacent states. Marine and nonmarine strata are almost equally represented, indicating that in this region the sea probably transgressed repeatedly onto an extensive low-lying coastal plain. Ten members or strata that differ in rock type constitute this kind of cyclothem. Complete development, however, of all the members in any cyclothem at any place is rare, so that the typical or complete cyclothem is, in a sense, an idealized reconstruction. Nevertheless, these ten members and their individual relations to each other are so constant and have been observed at so many places in so many cyclothems that there can be no doubt concerning their order and continuity throughout extensive regions. Some other strata interrupt the sequence in a few areas or zones, but this is not considered to have any broad significance.

Members of a typical cyclothem

The members of a completely developed cyclothem (Figure 1) in the central United States, from base to top, are as follows: (1) *Sandstone.* This generally is fine-grained but massive within local channels cut into underlying strata and thin bedded or even somewhat shaly beyond the channel margins. The older sandstones of the Lower Pennsylvanian are extensive sheet deposits that commonly are conspicuously cross-bedded and consist of relatively pure quartz sand, which may enclose patches or lenses of rounded quartz pebbles. Younger sandstones generally contain much undecomposed feldspar and abundant mica flakes. They are without pebbles except for some that obviously are of local derivation in the lower parts of channels. (2) *Sandy and silty shale.* This member generally is thin and succeeds the sandstone without abrupt transition. (3) *Lower limestone.* Characteristic marine fossils are not present in this member, which is likely to be thin, fine textured, and restricted to small areas except in part of the northern Appalachian region. (4) *Underclay.* Where a lower limestone (member 3) is undeveloped, the sandy shale (member 2) grades upward into this member. The underclay is nonbedded, generally breaks with a blocky fracture, and may contain the compressed and carbonized traces of plant rootlets. Many underclays are vertically zoned after the manner of modern soils (*q.v.*). Sometimes a leached zone that is noncalcareous in its upper part is present. (5) *Coal.* Underclays and coal beds are closely associated with each other. Pennsylvanian coals that do not rest on underclays are rare. Underclays, if not overlain by coals, generally are topped by smutty zones indicating coal horizons.

From Third Conference on the Origin and Constitution of Coal (1961

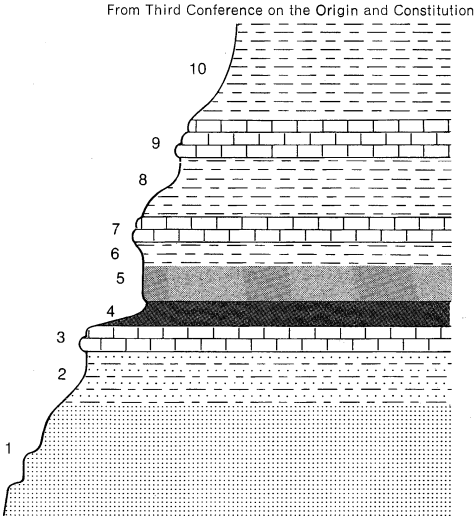


Figure 1: Fully developed "ideal" cyclothem of Illinois showing succession of ten members.

These five members constitute a lower hemicyclothem that seems to consist almost exclusively of nonmarine strata (see also COALS). Much, if not all, of the upper hemicyclothem is marine, however. (6) *Lower shale.* This member is rarely well developed, and it may not be recognizable if a middle limestone (member 7) is not present. It ordinarily is gray, silty, and unfossiliferous, except possibly for carbonized compressions of plant stems and leaves in the lower parts of some members when considerable thickness is attained. Fossils that are indicative of brackish water or marine conditions have been discovered at a few places. (7) *Middle limestone.* Many cyclothems lack this member, and it is unknown in the Appalachian region. In Illinois it is an impure, dark-coloured bed of discontinuous, lenticular character that contains marine fossils. In Kansas it is a compact, dark-coloured bed not more than two or three feet thick of remarkable uniformity and persistence. (8) *Middle shale.* This member is characterized by a bed of hard, black, sheety shale in its lower part, which may contain

fragmentary fish remains in some areas and fine-grained, very hard, black, calcareous concretions of considerable size. The black shale grades upward into soft gray shale that becomes more and more calcareous in its upper part and bears abundant marine invertebrate fossils at many places. (9) *Upper limestone.* This is a light-coloured, fossiliferous member that commonly is the thickest and most prominently exposed limestone of the cyclothem. (10) *Upper shale.* This ordinarily is the thickest member of the cyclothem. Immediately above the upper limestone (member 9) it may be calcareous and fossiliferous in a thin zone, but elsewhere it is dominantly silty; fossils are rarely present. Ironstone concretions occurring singly or at persistent horizons (continuous zones) are common. This member grades imperceptibly into the overlying sandstone (member 1 of the next cyclothem) outside of channels and beyond other areas of obvious erosion.

Complete cyclothems, however they may be developed, probably average less than 50 and rarely exceed 100 feet in thickness. As many as 100 or more cyclothems succeed each other in some regions.

Boundaries. The subdivision of such a repeated succession of distinctive lithologic members may be accomplished in several different ways. As originally proposed, a cyclothem was considered to begin at the base of the lower sandstone (member 1). This seems to provide a logical boundary that is an unconformity at some places. Therefore, it corresponds with at least a local interruption of a thick sedimentary deposit. This is significant from a dynamic, historical, and theoretical standpoint. Other boundaries that have been advocated are the top of the sandstone and the top of the principal limestone, either of which is more likely to be seen and traced in outcrops. The coal bed or coal horizon, which is particularly interesting for economic reasons, also has been suggested in this connection.

Recognition of cyclothems

Variations. The principal variability of cyclothems is related to their position with respect to the more or less regular fluctuations of the ancient sea. Thus in one direction, where deposition was mainly above the most advanced shoreline, marine strata are poorly represented or completely lacking, whereas in the other, where marine conditions prevailed, nonmarine strata may exhibit equally poor development. Thus, the cyclothem becomes more incomplete in both directions and, eventually, may become unrecognizable. The members that are most important in the identification of a cyclothem (Figure 2)

From Third Conference on the Origin and Constitution of Coal (1961)

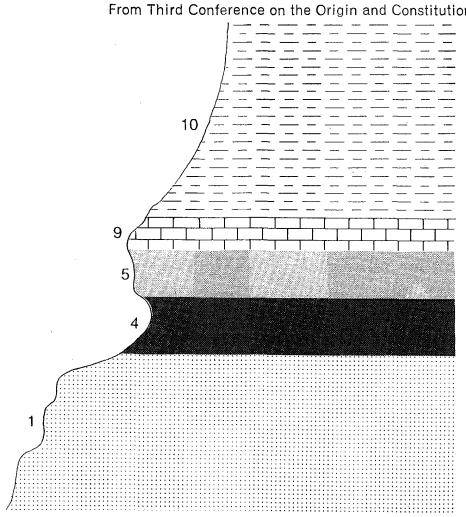


Figure 2: Incomplete but clearly evident cyclothem of five "primary" members.

are the lower sandstone (member 1), the underclay (member 4), the coal (member 5), the upper limestone (member 9), and the upper shale (member 10). The cyclothem can be recognized even if one of these is absent; generally, the cyclothem will be fairly obvious if only three members are present, provided that the underclay (member 4) is one of them. Reduction beyond this limit leaves a succession or alternation of beds that is not uncommon in strata of many ages and may bear no relation to any more complex cyclic deposits.

Distribution. Cyclothems are most completely and persistently developed in an intermediate part of the Pennsylvanian System that reaches a considerable thickness. In both lower and higher zones they are likely to lose some of their identity, either because of incomplete development or because sedimentary deposition was less cyclical. The numbers of cyclothems in equivalent parts of the stratigraphic section in different areas do not correspond exactly, and reduction or expansion of the section seems to be accomplished in several ways, none of which is adequately understood.

Pennsylvanian cyclothems are best known in the central and eastern United States, where gently dipping strata can be traced for long distances both in outcrops and below the surface in the records of wells and drill holes. They occur in comparable parts of the stratigraphic section in some of the principal coalfields of Great Britain, western and eastern Europe, and perhaps in northern China. Somewhat similar sequences also have been recognized in the Permian and Mississippian systems, and less convincingly in the Devonian of several regions, but these are not as clear nor as widely developed.

The term cyclothem has been used for other repeated sequences of strata, particularly those including coal, that seem to reflect types of cycles that differ from those characteristic of the Pennsylvanian. One of the better known is that which occurs in the Upper Cretaceous (about 75,000,000 years ago) of Utah, Colorado, and some adjacent states. This sequence lacks limestone and underclay, and the sandstones below coal beds probably are marine. The question may be raised whether the name cyclothem should be applied to such sequences. Loose usage surely detracts from the meaning and usefulness of the term.

SIGNIFICANCE OF CYCLOTHEMS

Cyclothems are interesting mainly because they reflect some of the physical conditions that prevailed in certain regions in the geologic past and show how these conditions varied contemporaneously in space and time. The distribution, thickness, and types of sediments indicate something about the location and nature of ancient upland areas where the sediments originated, and of depressed areas where they were deposited. The extent, depth, and perhaps salinity of the fluctuating seas are also suggested by the rock units. In a more practical way, cyclothems provide continuous and easily identified stratigraphic units. These are useful for correlating strata, including coals, within large areas more precisely than generally is possible by tracing and matching individual beds. This is economically important. Estimates suggest that some cyclothems represent intervals of time on the order of 75,000 to 100,000 years. Comparisons of the fossilized remains of similar animals contained in several successive cyclothems may show minor morphologic differences that characterize more closely spaced steps in their progressive evolution than ordinarily can be detected.

INTERPRETATION OF THE ORIGIN OF CYCLOTHEMS

Source areas of clastic sediments

Sediments in all of the lower hemicyclothems decrease in coarseness upward from the sandstone (member 1). The overlying sandy shales (member 2) seem to indicate a periodic decrease in the supply of sediments and the efficiency of transportation from their sources. In the United States the principal source regions lay along the eastern and southern continental margins as shown by the general increase in the prominence and coarseness of most sandstones in those directions. These regions contributed enormous amounts of detrital material to the interior of the continent, demonstrating the rapid ero-

sion of prominent uplands of large extent or, more probably, of smaller highlands that were uplifted repeatedly. Mineralogy of the sandstone suggests that during the earlier cycles sediments were derived mainly from the removal of weathered material from a soil-covered ancient land surface. Later sandstones, however, indicate that their sediments probably were produced by the rapid erosion of uncovered fresh rocks of granitic types. Unconformable contacts below the sandstones are most general near the source regions but abruptly entrenched channels occur far from them. Some of the channels were eroded nearly 100 feet into underlying strata and are a mile or more in width. These channels are thought to be segments of normal drainage systems (see also RIVERS AND RIVER SYSTEMS).

The succeeding members of the lower hemicyclothems show that sedimentary accumulation almost ceased. The underclays (member 4) are generally thin and contain no coarse material. The original nature of their sediment seems to have been altered slowly by the growth of plants and other normal soil-making processes. Lower limestones (member 3) probably record the local existence of temporary shallow freshwater ponds or lakes that occupied slight depressions on a vast, very gently sloping alluvial plain. Coals (member 5) record the existence of great lowland forests; plant materials accumulated as peat in extensive swamps. These periodically blanketed the coastal plains at times when appreciable quantities of detrital material were not being contributed by the bordering uplands.

Implications of coals and black shales

Coal swamp conditions were terminated by incursions of the sea. The water was brackish at first, as indicated by fossils in some of the lower shales (member 6). As more normal salinity was established, a restricted fauna mostly composed of small species of invertebrates became established, and their remains contributed to the middle limestones (member 7). The shallow sea seems soon to have been invaded by an abundant growth of kelplike algae. The decomposing remains of these plants furnished much carbonaceous material that produced anaerobic conditions at the bottom and, together with fine mud, accumulated as a rich organic slime. Later, compaction produced the hard, black, sheety strata in the lower part of the middle shale (member 8). As the shoreline advanced upon the land, water deepened, the growth of seaweed was much reduced, and the sea cleared. Black shale gave way to gray and finally to the upper limestone (member 9). The clear-water conditions at this time marked the culmination of a series of events in which, throughout an extended region, continental conditions were gradually supplanted by marine conditions.

The upper shale (member 10) demonstrates a complete reversal of the encroaching sea and deepening water that are recorded by all of the earlier members of the cyclothem. This member was deposited as the shoreline retreated seaward; the water became shallower, and the delivery of abundant detrital sediment was resumed. It is probable that the environment soon became brackish, and perhaps freshwater conditions followed. Finally, the sea withdrew and its bottom was exposed locally to channelling, or to less severe erosion, whereas elsewhere shaly or sandy sediments were deposited on its surface. Thus, the complete cyclothem seems to record a cycle of physical conditions first trending persistently and slowly in one direction and then suddenly and rapidly reversing (for further information see SEDIMENTARY ROCKS; SEDIMENTARY FACIES).

The widespread similarity of cyclothems on both sides of the Atlantic surely records a significant historical pattern of physical changes that followed each other regularly, at least in considerable parts of the Northern Hemisphere, during much of the Pennsylvanian and parts of the Mississippian and Permian periods. This general pattern may not have been restricted to Late Paleozoic time, but a special combination of conditions, including rapid but intermittent erosion of upland regions, sea level critically located with respect to continental platforms, the existence of an abundant and vigorous land flora adapted to marshy conditions, and favourable climates, all combined to produce unique results that were never duplicated exactly either before or since. Other kinds of stratigraphic alternations or successions, at other times and at other places, may record somewhat comparable patterns resulting in widely variable but partly analogous conditions.

Recognition of the existence of cyclothems and their distribution, and identification of the environments that they record, are only remotely related to deductions of the causes responsible for cyclothem development. Evidence pertinent to the historical interpretation is incomplete, scattered, and subject to quite different conclusions. It is not surprising, therefore, that there is much disagreement concerning the basic cause or causes that were chiefly responsible for the production of cyclothems. All present attempts at explanation of cyclothem development are excursions into the realm of theory.

Theories of origin. The theories that have been proposed to account for the development of cyclothems can be classified roughly into three groups, each of which includes several variations. These are (1) diastrophic theories, involving subsiding basins and rising sedimentary source regions that were either continuous, intermittent, or reversing; (2) climatic theories, involving glaciation that led to sea-level oscillations, and rainfall cycles and variable erosion; (3) sedimentation theories, involving differential deposition related to depth of water, strength of currents, distance from a river's mouth, compaction of sediments, and so forth. None of these theories seems adequate to explain completely the origin of cyclothems. When a satisfactory theory is formulated, it probably will include elements derived from several of those that have been proposed.

Diastrophic theories

Subsidence surely prevailed in coal basins where considerable thicknesses of sediment accumulated during extended intervals of times. Similarly, considerable elevation must have prevailed in the uplands where the sediments were produced by prolonged erosion. If basins subsided continuously and evenly, then the development of cyclothems resulted from other factors that controlled the kinds and quantities of sediment provided, and earth movements within the basins were not primarily important in the origin of cyclothems.

The oldest theory accounts for cyclothems and other similar sedimentary successions solely by intermittent subsidence in basins. This simple explanation supposes that deposition lagged behind relatively rapid local subsidence, so that the sea first transgressed low-lying areas and maximum water depth was soon attained. This was followed by a longer period of stability, during which deposition built up the bottom to above sea level and coal swamps developed on the newly emergent land. This theory infers nothing about conditions or the succession of events in the source areas of the sediments. Variations of this theory suggest that sea level fluctuated as the result of earth movements in regions beyond the basins. Such theories seem imperfect because most of the strata of the lower hemicyclothem should have accumulated under slowly shallowing marine conditions, whereas their nature indicates that they probably are nonmarine.

The distinct asymmetry in the succession of cyclothem members has been considered evidence that dominant slow subsidence in basins was interrupted periodically by more rapid and lesser local uplifts. Moreover, the kinds of sediment delivered to the basins can be reasonably explained if movements in the adjacent uplands were similar to and in phase with those of the basins. Thus, as the basins subsided and were progressively submerged, the bordering uplands decreased in height. Consequently they were eroded more slowly and contributed finer grained and lesser amounts of detrital material. Subsequently, as the basins rose, the borderlands were more greatly uplifted, erosion accelerated, and increasing quantities of coarse sediment built up the sea bottom and effected withdrawal of the sea. This theory has been crit-

icized as invoking a repeated series of earth movements so regular and mechanistic as to be unreasonable.

Climatic theories

Climate certainly has played an important part in the production and transportation of most detrital sediment. No matter what other factors were involved, it probably influenced the development of cyclothems. Some theories, however, have relied almost exclusively on climate in attempts at explanation.

Late Paleozoic glaciers spread over considerable parts of Antarctica, South Africa, Australia, India, and perhaps some other regions in the Southern Hemisphere, which may at this time have been combined as a single large continental mass (see CONTINENTAL DRIFT). If these glaciers were as extensive as those of the much later Pleistocene Epoch, they caused worldwide sea level to fall as the glaciers grew and withdrew water from the oceans, and to rise as they subsequently melted and the water was returned. This must have resulted in the alternate flooding and emergence of lowland regions in the Northern Hemisphere where cyclothems occur. The suggestion has been made that these sea level fluctuations can explain the alternation of nonmarine and marine hemicyclothems. One variant of this theory supposes that two glacial advances and retreats were necessary for the production of one cyclothem. Doubts are cast upon these explanations because it is uncertain whether the glaciers were sufficiently extensive to have effected sea level greatly or that glaciation and cyclothemic deposition were exactly synchronous. More seriously, it seems highly unlikely that there were as many glacial advances and retreats as the large number of cyclothems requires (see also CLIMATIC CHANGE; PLEISTOCENE EPOCH).

Sedimentation theories

Alternating humid and arid epochs resulting from glaciation or some other cause have been presented as either the main factor or an important contributory factor in the development of cyclothems. Thus, aridity related perhaps to the low temperature extremes of a glacial epoch would be unfavourable for plant growth in upland regions. A scant vegetative cover would permit maximum erosion of an ill-protected land surface and much detrital sediment would be swept out into a basin. On the other hand, the humid climate of an interglacial epoch probably would favour luxuriant vegetation that would protect uplands from severe erosion. Sea level would be high and submergence extensive at this time, so that conditions would be proper for limestone deposition. Nothing is known, however, about the nature of Late Paleozoic upland floras, so that these responses to climatic cycles are uncertain. In fact, the opposite conclusion has been reached, and it has seemed likely that epochs of aridity might be represented by the limestones and epochs of humidity by detrital strata.

All sedimentation theories are more or less complex. They relate the production, delivery, and deposition of different materials to a variety of factors such as earth movements, fluctuating climate, physiographic development of the land, changing sea level, strength of currents, distance from source, compaction of sediments, and the building and breaking of barriers to the distribution of sediments. Several of these theories include some elements of the foregoing.

Successive uplifts of the land each followed by a period of stability during which erosion reduced its surface to low level might account for a supply of sediment whose coarseness and amount both declined with reduction of the height of land. Under proper conditions, which need not be detailed here, deposition of these sediments in an adjacent basin might produce cyclothems or cyclothem-like sequences. The time estimated as being represented by an individual cyclothem, however, is relatively brief by geologic standards, and it is hardly conceivable that erosion during such an interval could reduce an elevated source region so effectively.

Several theories relate coarseness of sediment and the development of cyclothems to settling in different depths of water and explain vertical changes in the deposits by the advance and retreat of shorelines and the resulting variation of water depth. Finer grained sediments were transported farther out from shore and deposited in deeper water. One theory proposed that because vegetable material floated, it was carried far out to sea, where it finally sank, and, next to limestone, coal might record the deepest water environment. This last, of course, is contradictory to good evidence that peat accumulated where the vegetation grew and later was transformed in place to coal.

Strength of currents in the sea would have much the same effect as depth of water in determining the kind of sediment deposited. Thus as currents failed, the coarser sediments settled out progressively and only the finer material was carried on. All theories of these kinds neglect the implications of alternating nonmarine and marine conditions in the better developed cyclothems.

The distributary channels in a growing delta shifted from side to side, so that detrital sediments were delivered to the sea first in one area and later in another, where sand bodies were built up. There the sedimentary surface rose to or above sea level and coal swamps may have formed. At the same time, normal marine conditions prevailed elsewhere. Lateral shifting of the delta front from time to time and submergence of its older parts resulted in the development of cyclothems. Under conditions of this kind, contemporaneous sediments must have varied greatly from place to place. This does not accord with the observation that many cyclothems and some of their members are remarkably persistent and continue with little change for distances of much more than 100 miles.

The different sedimentary environments existing simultaneously in deltaic areas commonly are separated by physical barriers. These include natural levees, bordering distributary stream channels, and sandbars along coasts. The building or destruction of such barriers is likely to result in sudden and extreme environmental changes that might be recorded in the succession of different lithologic members of a stratigraphic sequence. Thus, erection of natural levees might permit the development of adjacent coal swamps, and the breaching of a coastal barrier might result in their inundation and burial beneath marine deposits. The suggestion also has been made that a dense growth of swamp vegetation could act as a filter and prevent the distribution of detrital sediment far from the channel of a stream. The results of all such actions probably would not be felt beyond the limits of local areas, and they are not likely to have been the principal influences in the development of cyclothems or their members that are persistent throughout large areas.

Shapes of sandstone bodies

The implications of sandstone-filled channels at the base of lower hemicyclothems are disputed. The linear extent of some sandstone bodies, commonly termed shoestring sands, and the abrupt entrenchment of others have been interpreted as evidence that the channels were first eroded and later filled by streams flowing across emergent alluvial plains. The contrary contention that channels were cut below sea level by submarine density currents (*q.v.*) necessitates a different explanation for cyclothem development. In either case, preference in interpretation is likely to be influenced strongly by theoretical considerations concerning the sequence of events and environments in cyclothem development. Marine fossils have been found in a few sandstones, and some sand bodies comparable to those occupying channels have been considered offshore bars. Most of them, however, generally are conceded to be nonmarine.

Differential compaction of sediment has been included as a more or less important part of several theories. Most newly deposited sediments, and especially peat, are highly porous and contain much entrapped water. When this is squeezed out, either by settling or, more effectively, by compression beneath the load of other subsequently deposited sediment, the sediments are compacted and reduced in thickness. Thus, the compaction of a thick bed of peat that had accumulated at or very close above a shoreline might have had the same effect in a basin as subsidence or rising sea level, and this may explain the presence of undoubted marine strata immedi-

ately overlying coal beds. Not explained is why compaction of the peat, resulting from settling alone, was delayed until a considerable thickness had accumulated or the reason for marine beds above a smut streak, indicating that no peat had been preserved.

OTHER CYCLIC UNITS

Megacyclothems. The depositional cycle recognized in the Middle and Upper Pennsylvanian of Kansas includes four or five closely spaced marine limestone members, separated by relatively thin and dominantly shaly intervals. These members are termed the "lower," "middle," "upper," "super," and "fifth" limestones. The interval between the "middle" and "upper" limestones is particularly noteworthy and differs from the others because it generally includes a bed of hard, black, sheety shale. Thin coals are likely to occur close below the "lower" limestones, which are marine. Each sequence of this kind is separated from others higher and lower in the stratigraphic section by thicker shaly intervals that are poorly characterized and may contain thin beds of sandstone, limestone, and coal. Originally, each of the four or five limestones was supposed to be part of an incompletely developed cyclothem of the kind that is well-known in Illinois, and the whole Kansas cycle was termed a megacyclothem.

The relations of these cyclic units of Kansas and Illinois are disputed, and the detailed correlation of most of the Pennsylvanian strata in these two states is quite uncertain. There are fewer megacyclothems in Kansas, however, than there are cyclothems in what seem to be equivalent parts of the section in Illinois. This does not favour the conclusion that megacyclothems are in some way more fully and elaborately developed cyclothems. Also, if such equivalence were so, any explanation of the origin of cyclothems would be much more complicated with respect to rising and falling sea level and other changing depositional conditions.

A much less conspicuous repetition of differently developed cyclothems has been noted in Illinois. All have basal sandstones, many include coals, and the limestones of successive cyclothems are more widely spaced than the limestones in Kansas. One of the cyclothems is particularly characterized by the succession of middle limestone (member 7), middle shale (member 8) with hard, black, sheety strata, and a thin upper limestone (member 9). This at once suggests comparison with the "middle" and "upper" limestones and intervening shale of the Kansas megacyclothem. The next higher Illinois cyclothem is especially noteworthy for its thick, light-coloured upper limestone (member 9). This is suggestive of the "super" limestone of Kansas, which generally is also a thick, light-coloured member. Separating successive pairs of cyclothems of these two kinds in Illinois are one to three less individualistic cyclothems, some of which include strata with brackish water rather than typical marine fossils.

If the foregoing comparisons of Illinois and Kansas limestones are significant, the shale interval between the Kansas "upper" and "super" limestones might be equivalent to the upper shale (member 10) of the first Illinois cyclothem and all of the members of the lower hemicyclothem of the second cyclothem. Close examination of these Kansas shale intervals reveals the intermittent and local occurrence in several of them of thin sandstone or sandy shale, poorly developed underclay, thin coal or smut streak, and reddish shale. All of these are types of strata that normally occur between the upper limestone (member 9) of one Illinois cyclothem and the middle limestone (member 7) of another. Thus it seems possible that at least two cyclothems whose detrital members are very poorly represented constitute parts of a Kansas megacyclothem and include three of the limestones that have been mentioned.

Similar consideration of the lithologic nature of other shale intervals in the Kansas megacyclothem indicate that the "lower" and "fifth" limestone, where this is present, also are probably upper limestones (member 9) of incompletely developed cyclothems. The entire Kansas megacyclothem, therefore, seems to consist of three or more ordinary cyclothems that lack the greater part of their nonmarine and detrital members. Also, the succession of different cyclothems that is repeated in Illinois seems to be a comparable but more fully developed megacyclothem deposited closer to the source of the detrital sediments. Seventeen of these units are clearly evident in Kansas. There are only about seven in Illinois, but the stratigraphic section is truncated above and may not include equivalents of the higher Kansas units. No suggestion of megacyclothem development has been recognized in the Appalachian region or in foreign coalfields. Little thought has been devoted to explaining the origin of megacyclothems.

Hypercyclothems. The Kansas stratigraphic section reveals a still greater Pennsylvanian cycle that has been termed a hypercyclothem. Each consists of four successive megacyclothems and an alternating detrital sequence of more than ordinary thickness and complexity. This is likely to include one or more sandstone members that may occupy erosion channels. Four hypercyclothems increasing in perfection upward are recognizable in Kansas and there is the beginning of a fifth. Nothing suggestive of such large cyclic units is known in other areas, and no explanation for them has been ventured.

BIBLIOGRAPHY. No books have been written on cyclothems as such, and although cyclic sedimentation is treated briefly in textbooks on sedimentology, the best sources of information are within the somewhat technical literature. Articles of greatest relevance include the following: J.R. BEERBOWER, "Origin of Cyclothems of the Dunkard Group (Upper Pennsylvanian–Lower Permian) in Pennsylvania, West Virginia, and Ohio," *Bull. Geol. Soc. Am.*, 72:1029–1050 (1961), a review of various theories explaining cyclothems with a suggestion that climate and discharge rate are the most important factors; D.F. MERRIAM (ed.), "Symposium on Cyclic Sedimentation," *Bull. Kans. Geol. Surv. 169*, 2 vol. (1964), a collection of papers discussing the general and specific aspects of cyclothems and cyclic sedimentation; M.H.P. BOTT and G.A.L. JOHNSON, "The Controlling Mechanism of Carboniferous Cyclic Sedimentation," *Q. Jl. Geol. Soc.* Lond., 122:421–441 (1967), a discussion of the theory that varying rates of crustal subsidence can result in eustatic rises in sea level and are responsible for Carboniferous cyclic sedimentation; B.A. SILVER and R.G. TODD, "Permian Cyclic Strata, Northern Midland and Delaware Basins, West Texas and Southeastern New Mexico," *Bull. Am. Ass. Petrol. Geol.*, 53: 2223–2251 (1969), an illustrated explanation of the cyclic sequences in these basins and a discussion of more general concepts; J.M. WELLER, *Patterns in Pennsylvanian Cyclothems*, Nova Scotia Dept. Mines, Third Conference on the Origin and Constitution of Coal, pp. 129–166 (1961), a review of the development of cyclothemic studies in the U.S. and a comparison of the different types of cyclothems occurring in different regions and at different stratigraphic positions; and "Development of the Concept and Interpretation of Cyclic Sedimentation," *Bull. Kans. Geol. Surv. 169*, vol. 2, pp. 607–621 (1964), a history of cyclical investigations and interpretations, particularly of the Pennsylvanian, presented by brief quotations from nearly 100 articles published between 1830 and 1964.

(J.M.W.)

Cyprian, Saint

Cyprian, bishop of Carthage in the mid-3rd century, led the Christians in North Africa through the crises of the persecutions under the Roman emperors Decius (249–251) and Valerian (253–260). His teaching on the nature of the church and its sacraments and on the office of a bishop molded Christian thought, and in particular the theology of the Donatists, a North African schismatic movement, until the time of St. Augustine in the 5th century. A collection of 81 letters, including some from his contemporaries, a dozen treatises, and his "Life," written by the deacon Pontius, provide a vivid record of Christianity in North Africa at a decisive stage in its history.

Bishop during the Decian persecution. Cyprian (Thascius Caecilius Cyprianus) was born of wealthy pagan parents about AD 200. Educated in law, he practiced as a lawyer for some time in Carthage. He was converted to



St. Cyprian, detail of a polyptych by Girolamo di Giovanni da Camerino, 1473. In the church of Sta. Maria del Pozzo, Monte San Martino, Italy.
By courtesy of the Gabinetto Fotografico Nazionale, Rome

Christianity about 246, finding in Baptism complete release from the sinful and useless life he believed he had led hitherto. Within two years he was elected bishop of Carthage and a few months later, early in 250, was confronted by the Decian persecution. He went into hiding.

Problem of apostasy and the confessors

Bereft of his leadership, thousands of Christians apostatized (rejected their faith) or obtained *libelli pacis* (certificates), by which they declared that they had sacrificed to the pagan gods. When the persecution began to diminish, the confessors—*i.e.*, those who had stood firm for their faith—reconciled the lapsed on easy terms, claiming that as "friends of Christ" they had the right of granting pardon, even more than did priests and bishops. Cyprian returned to Carthage (early 251) and at a council of bishops in May 251 was able to regain his authority. The decision of the council was that, though no one should be totally excluded from penance, those who truly had sacrificed (the *sacrificati*) should be readmitted only on their deathbeds, and those who had merely accepted certificates (the *libellatici*) were to be readmitted after varying periods of penance. Three important principles of church discipline were thus established. First, the right and power to remit deadly sins, even that of apostasy, lay in the hands of the church; secondly, the final authority in disciplinary matters rested with the bishops in council as repositories of the Holy Spirit; and, thirdly, unworthy members among the laity must be accepted in the New Israel of Christianity just as in the Old Israel of Judaism.

In 252 a renewed threat of persecution by the emperor Gallus encouraged a more speedy reintegration of the lapsed, because many now wanted to prove themselves as martyrs. In the same year, the steadfastness of the Christian clergy in face of a plague won for the church further popular support, and Cyprian defeated internal enemies who had set up a rival bishop in Carthage.

Relations with Rome. In the summer of 254 his position was tested again by a dispute with Stephen, bishop of Rome (254–257). Until then relations between the churches of Carthage and Rome had been cordial. In 251 Cyprian had supported Bishop Cornelius against his rival, Novatian, and had written on his behalf the treatise *On the Unity of the Church*, which stressed the centrality of the see of Peter (Rome) as the source of the episcopacy. Though Cyprian may have written two drafts of an important passage concerning the primacy of the chair of Peter, he implied no acceptance of Roman jurisdictional prerogatives. When in 254 two Spanish congregations (Mérida and León) appealed to him against a decision by Stephen to restore bishops who had lapsed during the persecution, he summoned a council to consider the case. The council decided that the congregations not only had a right but a duty to separate themselves from a cleric who had committed a deadly sin such as apostasy. Cyprian wrote (Letter 67) that the Holy Spirit was no longer in such a priest and that his sacraments would lead to perdition and not salvation. The church as the "pure Bride of Christ" might be obliged to absorb a sinful laity, but a sinful priest making offerings on behalf of the people was unthinkable.

Within months there was an even more serious dispute with Rome. For a few years the supporters of Novatian had been active in Africa, asserting against Cyprian that no forgiveness for lapsed Christians was possible. With the recovery of Cyprian's prestige, however, their threat began to fade. Many of those whom they had baptized clamoured to be admitted to the church. Was their Baptism valid or not? In Rome, Stephen, confronted by the same problem, decided that all Baptism in the name of the Trinity was valid. The Africans at first were of two minds. Cyprian held three councils between the autumn of 255 and September 256. The last, at which 87 bishops were present, decided unanimously that there could be no Baptism outside the church, just as there could be neither faith, hope, nor salvation for those outside it. A minister could not dispense what he himself did not possess, namely, the Holy Spirit. Those who had received Baptism from Novatianists must be baptized anew. Behind this clash over rites lay the more fundamental question concerning the nature of the church. Though Rome emphasized the church's universal and inevitably mixed character on earth, the North Africans stressed its integrity under all circumstances. Baptism entailed total renunciation of the world and the reception of the Spirit.

Concern for the validity of schismatic sacraments

Persecution under Valerian. A complete breach between Rome and Carthage was averted by Stephen's death on August 2, 257, and his successor, Sixtus II, was more conciliatory. Meanwhile, persecution had been renewed by the emperor Valerian (253–260). On August 30, 257, Cyprian was summoned before the proconsul, Aspasius Paternus, and assigned an enforced residence at Curubis (Kurba) on the Gulf of Hammamet. Following a more severe edict the next year, he was brought back to Carthage, tried, and condemned to death. He was executed in the evening of September 14, 258, and became the first bishop-martyr of Africa.

During the previous seven years his character had matured. Though not the "man of moderation" eulogized by his biographer, he had shown himself a brave and resourceful leader of the church in Africa. His theology was based on the central idea of the unity and uniqueness of the church: "He no longer has God for his Father, who does not have the Church for his mother" (*On the Unity of the Church*). Unity was expressed through the consensus of bishops, all equally possessing the Holy Spirit and sovereign in their own sees. There was no "bishop of bishops." The church consisted of the people united to their bishop. Schism and rebellion against the priesthood were viewed as the worst of sins. These views —associated with an uncompromising insistence on the integrity and exclusive character of the church, which are believed to have been derived from the North African theologian Tertullian—received divine sanction for most North African Christians through his martyrdom.

Cyprian's theological views

BIBLIOGRAPHY

Texts: Opera omnia, ed. by W. HARTEL, 3 vol. (1868–71); *De Lapsis* and *De Ecclesiae Catholicae Unitate*, text and translation by MAURICE BEVENOT, "Oxford Early Christian Texts," vol. 3 (1971), with good bibliography. English translation by R.E. WALLIS, *The Writings of Cyprian, Bishop of Carthage*, 2 vol., "Ante-Nicene Christian Library" (1868–69). Selection in S.L. GREENSLADE (ed.), *Early Latin Theology*, pp. 113–172 (1956).

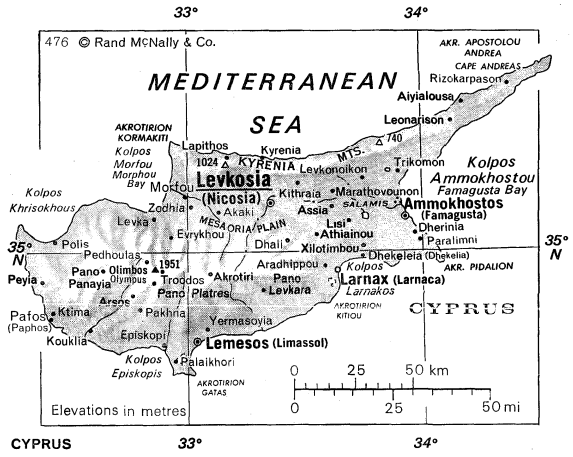
General studies: E.W. BENSON, *Cyprian: His Life, His Times, His Work* (1897), still the best monograph; J.P. BRIS-

SON, *Autonomisme et Christianisme dans l'Afrique romaine de Septime Sévère à l'invasion Vandale* (1958), shows links between Cyprian's theology and that of the Donatists; M. BEVENOT, "St. Cyprian's *De Unitate*, Chapter IV, in the Light of the Manuscripts," *Analecta Gregoriana*, vol. 11 (1937), a significant study of Cyprian's attitude toward the Roman primacy; H. KOCH, *Cyprianische Untersuchungen* (1926), Cyprian's theology of the Church; W.H.C. FREND, *Martyrdom and Persecution in the Early Church*, ch. 12 (1965), deals with Cyprian and the persecutions of Decius and Valerian; L. DUQUENNE, *Chronologie des lettres de Saint Cyprien* (1972).

(W.H.C.F.)

Cyprus

The island of Cyprus (Greek Kipros, Turkish Kıbrıs), a striking setting for often dramatic events, lies in the eastern Mediterranean Basin, about 40 miles (64 kilometres) south of Turkey, 60 miles west of Syria, and 250 miles (400 kilometres) north of the Nile Delta, at about the same latitude as Tokyo and San Francisco. Its maximum length, from Cape Arnauti (Akros Arnoutis) in the west to Cape Andreas (Akros Apostolou Andrea) at the end of the narrow, dagger-like northeastern peninsula, is 140 miles (224 kilometres); maximum north–south extent is 60 miles. With an area of 3,572 square miles (9,251 square kilometres), it is the third largest Mediterranean island.



The island's history prior to the 14th century AD is obscure, though archaeologists date the earliest settlements found in Cyprus from Neolithic times, in about 5800 BC. Since then it has been colonized, occupied, or dominated by virtually every one of the peoples who have, over the centuries, successively exercised power in the region—the Greeks, Phoenicians, Assyrians, Egyptians, Persians, Romans, Byzantines, crusaders of Richard I the Lion-Heart, Lusignans, Venetians, Genoese, Turks, and, more recently, the British.

Independence

The British occupation began in 1878, and it was from Great Britain that the Cypriots finally achieved independence in August 1960, largely by means of an armed rebellion launched in 1955 by the EOKA (Ethnikí Orgánosis Kipriakoú Agónos, or National Organization of Cypriot Struggle).

See further CYPRUS, HISTORY OF.

THE ENVIRONMENT

Topography. The island's unique shape is due, geologically, to the occurrence of two ridges, once part of two much greater arcs running from the Asian mainland toward Crete. These remnants—the Kyrenia Mountains in the north and the imposing Troodos Massif in the south—form the two dominant features of the island. A flat lowland, open to the sea at east and west and known as the Mesaoria Plain, lies between the two ranges. It was here that the ancient city-state of Ledra—now the capital city, Nicosia (Greek Levkosia; Turkish Lefkoşa), encircled by a great wall built in the 15th century during the Venetian occupation—was established.

The Kyrenia range, known as the Pentadaktylos (Five Fingers) from the five-pointed eminence that is one of its main features, is a narrow fold of limestone with occasional deposits of marble; it has a maximum height of just over 3,000 feet (900 metres). The range stretches 80 miles from east to west, leaving along the northern coast a narrow, fertile plain, largely evergreen but also containing olive, carob, and citrus trees.

The Troodos range, fragmented by geologic folding, is mainly igneous rock, impervious to water, but it has a thicker soil and a covering of pine, dwarf oak, cypress, and cedar forest. The highest point is 6,401-foot (1,951-metre) Olympus (Olimbos; formerly Chionistra). The range stretches about 50 miles from near the west coast, culminating in the 2,260-foot (678-metre) eminence of Stravrovouni, 12 miles from the east coast. The general pattern of the 486-mile (778-kilometre) coastline is indented and rocky, with long, sandy beaches.

The Mesaoria Plain, which spans the island from Morphou Bay (Kolpos Morfou) in the west to Famagusta Bay (Kolpos Ammokhostou) in the east, is about 80 miles long and from 15 to 30 miles in breadth. Between autumn and spring it is green and colourful, with an abundance of wild flowers and flowering bushes and shrubs; there are also patches of woodland in which eucalyptus and many types of acacia, cypress, and lowland pine predominate.

Medieval castles

The plain is overlooked by the Kyrenia Heights, on which the ancient medieval castles of St. Hilarion, Buffavento, and Kantara are situated. The plain also furnishes the major supply of grain, principally wheat and barley. About 300,000 acres (120,000 hectares) are cultivated under irrigation, and the remainder is devoted to dryland farming.

Climate. Mean average rainfall is about 19 inches (483 millimetres), the highest precipitation (44 inches) being on the Troodos range and the lowest (12 inches) in the central plain. Temperatures range between an average daily maximum of 97° F (36° C) and an average daily minimum of 69° F (21° C) at Nicosia, though during the high summer in July and August the maximum is often as high as a scorching 105° F (41° C) in the plain, with as much as 30° F variation between maxima and minima. During winter (December to March) Troodos experiences several weeks of below-freezing night temperatures.

Between December and April the weather pattern is variable, influenced by prevailing conditions in the Aegean Sea and on the Turkish mainland and sometimes by complexes affecting Libya, Egypt, and Syria. From mid-June to mid-September, a low-pressure trough becomes established over the eastern Mediterranean and brings high temperatures and steadily increasing humidity with little or no rain.

There are no running rivers, although the island is crisscrossed by dry river valleys that become fast-flowing torrents during periods of heavy rainfall, with consequent severe erosion of soil.

Animal life. Although fossil remains of elephant and hippopotamus have been found in the Kyrenia area, and early writers speak of deer and boar roaming the island in some abundance, the only large wild animal that has survived is the mouflon, a species of wild sheep now believed unique to the island. The mouflon came near to extinction in 1937 and is now under strict protection; its last refuge is the Paphos Forest. Small game abounds but is mercilessly hunted, and by the 1970s it appeared that game laws were inadequate to prevent great annual destruction.

In spring and autumn the island experiences the migration of millions of birds, but experts believe that indiscriminate shooting and poor protection laws are likely to change the pattern of migration. There are many species of birds native to the island, but these also are subjected to the pressure of almost unrestricted hunting.

THE PEOPLE

Ethnic and religious heritages. The people of Cyprus represent two main ethnic groups, Greek and Turkish.

Cyprus, Area and Population

	area		population	
	sq mi	sq km	1960 census*	1973 census
Districts				
Famagusta	764	1,979	114,000	124,000
Kyrenia	247	640	31,000	33,000
Larnaca	436	1,129	59,000	61,000
Limassol	539	1,396	107,000	125,000
Nicosia	1,048	2,714	204,000	233,000
Paphos	538	1,393	58,000	57,000
Total Cyprus	3,572	9,251	574,000†	632,000†

*Includes 1,000 members of the Greek and Turkish military contingents. †Figures do not add to total given because of rounding.
Source: Official government figures.

The former are descended from the earliest inhabitants but have blended substantially with the various waves of invaders, especially with Ionian colonists from Greece. The Turkish Cypriots are the descendants of the Ottoman Turks who conquered the island in 1571 and occupied it until 1878.

Cypriots are a people of rugged individuality, gregarious and friendly and almost embarrassingly hospitable to strangers. Family ties are very close, and the dowry system, under which the woman brings a home to the marriage, is a strong tradition among the Greeks, persisting in spite of efforts by the church to discourage it. By contemporary standards, the Cypriots tend to marry relatively late, average ages being 26 for men and 23 for women, the latter predominating in the population in the ratio of 50.5 to 49.5 percent. Divorce is comparatively rare, the ratio being 0.25 compared with a marriage rate of 8.7 per 1,000.

Eastern Orthodox traditions

The Greek Cypriots are Eastern Orthodox Christians. Since about AD 478 they have enjoyed a degree of autonomy under an ethnarch elected by the bishops of the island, and a 20th-century incumbent, Archbishop Makarios III, became a world-famous political figure. This autonomous authority was bestowed after the remains of the Apostle Barnabas, who was born in Cyprus and met his martyrdom there—at Salamis, near modern Famagusta—had been discovered and restored to the care of the patriarchate. The island was among the earliest converts to Christianity: the Apostles Paul and Barnabas preached there as early as AD 45, achieving the conversion of the Roman proconsul Sergius Paulus.

The Turkish Cypriots are practicing Muslims who follow the secular observances established on the Turkish mainland by Kemal Atatürk. There are also small groups of Maronites, Armenians, and Roman Catholics.

Modern Greek is the language of the Greek majority, and Turkish is spoken by the Muslim minority; but, as a result of the 82 years of British rule, English is widely spoken and understood as a second language. Illiteracy is very limited, thanks to an education system that provides free elementary instruction for all children.

Demographic trends. At the census of 1973 the population was 631,778, a density of almost 160 to the square mile; Greek Cypriots comprised 77 percent and Turkish Cypriots 18 percent of the total. Cyprus has a relatively youthful population, with about 56 percent under the age of 30.

The natural growth rate over the decade 1965–75 declined from 1.5 to 0.9 percent, with an average of 21 births per 1,000 and a death rate of 10 per 1,000. A survey in 1973 disclosed that birth rates had long been substantially overestimated, while death rates, formerly supposed to be among the world's lowest, had been underestimated by a corresponding amount. The death rate, however, is low, which can be attributed in part to a remarkable decrease in the infant mortality rate—from 71.7 per 1,000 births in 1949 to 28.4 in 1973.

Cypriot emigration

Cypriots have long shown an urge to emigrate, largely because of the limited economic opportunity offered in their homeland in the past, and it is estimated that there are almost as many of them living overseas as there are on the island itself.

There was a marked increase in emigration immediately after independence (14,589 in 1960 and 13,489 in 1961); the effect of subsequent economic progress slowed the rate to 1,312 in 1973, but political events of 1974 led to a sharp upsurge of Greek Cypriot emigration in the second half of that year.

The Cypriots are still largely a rural people, despite a marked drift from the countryside. There are more than 600 villages and only six main towns. Nicosia, the capital, lies about 12 miles (19 kilometres) from the north coast; Famagusta, the principal port, is on the east coast 38 miles (61 kilometres) from the capital; Limassol (Lemesos), the port serving the southern area, is 50 miles (80 kilometres) by road from Nicosia; the other towns are Larnaca (Larnax) in the east, Kyrenia on the north coast, and Ktima-Paphos in the southwest. Each of the six towns is a district headquarters.

THE NATIONAL ECONOMY

Policies since independence. Cyprus maintains a free-enterprise economy, and the private sector was allocated a major role in the First (1962–66), Second (1967–71), and Third (1972–76) Five-Year plans. The success that attended this policy is reflected in the growth of trade and financial stability after independence. Exports rose from £20,200,000 in 1960 to £55,200,000, and imports from £39,000,000 to £148,000,000, in 1974. Revenues showed a similar increase, in spite of a levelling down of direct taxation in 1969, which reduced the maximum payable from 75 to 60 percent and also cut company taxes. Government expenditures almost trebled, to about £59,000,000 in 1973, and there was a consistent budget surplus position. Although the visible trade deficit appeared to be of alarming proportions, the overall balance of payments regularly produced a credit until late 1973, thanks to income derived from overseas earnings and remittances from Cypriots living abroad, plus substantial earnings derived from the presence of British and United Nations military forces. Foreign exchange reserves at the beginning of 1975 totalled £103,900,000, having shown in 1973 a net decline for the first time in any year since independence. Customs and excise duties and a direct tax on incomes represent the major part of government revenues. The public debt at the end of 1974 stood at £25,400,000, of which £8,000,000 was external loans.

The economy is predominantly agricultural, and this sector provides full- or part-time employment for nearly 95,000 persons, the main production being wheat and barley, potatoes, carrots, and other vegetables, carobs, citrus fruits, and vine products. During the years following independence, the island achieved a per capita standard of living higher than most of the Middle East, with the exception of Israel, and maintained economic growth till 1973.

UN assistance in the Cypriot economy

This progress was substantially assisted by various agencies of the United Nations, operating through the UN Development Program. Generous financial assistance was given by the World Bank and International Monetary Fund in the form of loans for specific development projects (electricity supply, port development, sewerage, etc.). Experts were made freely available to advise on economic planning and to initiate productive projects, and training for Cypriot specialists was encouraged by scholarships and grants.

Landholdings are very small, highly fragmented, and dispersed under the traditional laws of inheritance. In 1967, on the advice of UN experts, a program of land consolidation was begun; but because of the break with tradition and fear that the program might create problems of rural underemployment, the first phase was expected to take about 25 years to complete. Turkish opposition did not cease, but consolidation gained support from Greek Cypriots. Turkish landowners were encouraged to join.

One-half of the island's manufacturing capacity, as well as the citrus-growing areas and the main tourist centres, fell into Turkish hands in the invasion of 1974. Total losses to trade and industry as a consequence of the invasion were assessed at as much as £350,000,000.

Resources and industry. Raw material resources are very limited, restricting the scope for industrial activity; but there has been useful development of light manufacturing, and the contribution of this sector to the economy more than doubled between 1961 and 1973. Electric power production more than quadrupled between 1960 and the mid-1970s.

Cyprus has for many centuries been a producer of copper (in fact, the metal takes its name from that of the island). As early as 2500 BC the island's mines were being exploited, and traces of Phoenician and Roman workings and surface slags are still to be seen. With the discovery of other sources, the mines remained neglected for many generations until they were reopened by an American prospector in 1925. After World War II they were brought back into production, and since then copper and other minerals—iron pyrites, asbestos, gypsum, chrome ore—have contributed some 40 percent annually to the export trade. Reserves of cupreous ores are fast declining, and it is anticipated that by 1982 the industry will have ceased to be economic. Despite systematic prospecting since 1963, no important new deposits have been located. There are, however, substantial reserves of asbestos, chrome, gypsum, and iron pyrites, which will sustain these branches of the mining industry for many years to come.

There are extensive sandstone quarries, the golden-hued product of which was widely used for building the older style of Cypriot houses before the cement age came to Cyprus. There is now a thriving and expanding cement industry.

Deforestation

Cyprus was once an island of abundant forests, but the demand for timber for shipbuilding, by successive conquerors from the pharaohs onward, and extensive felling for building and for fuel have destroyed the greater part. By the 1970s a policy of conservation and reforestation was being pursued. The forests cover some 670 square miles (1,735 square kilometres), a large proportion of them being found in the mountain areas and in the southwestern district of Paphos.

Transportation. The earliest roads in Cyprus were built by the Romans (58 BC to AD 395), and traces of them still exist. At the time of the British occupation (1878), the only carriage road was between Nicosia and Larnaca; today the island has one of the most comprehensive road networks in the region. All transport is dependent upon the roads, and motor transport has multiplied enormously. In 1907 a narrow-gauge government railway was opened between Famagusta and Nicosia and later extended to Evrykhou in the Troodos foothills, but the installation proved uneconomical and was closed in 1951.

There is no coastal shipping. Some progress has been made toward building up a merchant marine, but the tonnage registered under the Cyprus flag remains small, and most of it is foreign owned. Cypriots, although an island people, have no seafaring tradition.

International air services have expanded greatly since independence; the international airport at Nicosia provides connections to all parts of Europe and the Middle East and some points in Africa, and there is steady expansion of airfreight services.

The great bulk of the island's trade, however, remains seaborne, through the main port of Famagusta on the east coast and Limassol in the south. There are no internal air services.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional framework. The constitution of 1960 established an independent republic with a Greek Cypriot president and a Turkish Cypriot vice president; a Council of Ministers (Cabinet) of seven Greek and three Turkish members; and an elected House of Representatives of 50 members, 35 Greek and 15 Turkish. It also set up Greek and Turkish communal chambers to control community affairs; an army of 2,000, with 60 percent Greek and 40 percent Turkish personnel; and an urban police force and a rural gendarmerie recruited on a 70:30 ratio. It incorporated a Treaty of Guarantee that named Greece, Turkey, and Great Britain as guarantor powers and a Treaty of Alliance between Cyprus, Greece, and Turkey.

Under the Treaty of Establishment of the Republic of Cyprus, Great Britain retained sovereign military enclaves—comprising 99 square miles (256 square kilometres)—at Episkopi and Akrotiri, near Limassol in the south, and also at Dhekelia (Dekeleia) near Larnaca in the east.

There was no acclaim for independence; the anniversary (August 16) is not celebrated, nor is the flag of the republic flown except on government buildings. The EOKA rebellion had been basically a campaign for union of the island with Greece (*énosis*), a union unacceptable to the Turkish Cypriots. Because of this fundamental conflict, the constitution fell into abeyance early in 1964 after political and constitutional disagreement had rendered it unworkable.

The political process. The constitution of 1960 provided for the president and members of the House of Representatives to be elected every five years on a basis of universal suffrage, every national over the age of 21 being entitled to vote. Archbishop Makarios III became president at independence. Before 1970 there was only one political party—the Cypriot Communist Party (AKEL). Other parties emerged in the elections of that year, and three were represented in the legislature: Unified Party (right wing), Progressive Front (right wing opposition), and a socialist party known as the Democratic Centre Union Party (EDEK).

Political parties

Local government and the judiciary. Local government is at district, municipal, rural municipality (improvement board), and village levels. District officers are appointed by the Cabinet, while—according to the constitution—local councils are to be elected; in practice, though, there were no such elections from independence to the end of 1975.

Justice is based on Roman law, dispensed by judges appointed by the president, and the judiciary is entirely independent of the government establishment. The judiciary of the early 1970s, however, was composed only of Greek Cypriot judges. There is a supreme court and an appeals court, district assize courts handling criminal matters, and district courts exercising summary jurisdiction.

The army. The 2,000-member Cyprus Army established by the constitution never actually came into being because of the intercommunal dispute. In 1964 a small but relatively powerful Greek Cypriot National Guard of about 15,000 was organized under an officer corps recruited from mainland Greece, and the Turkish Cypriots established a "freedom fighter" force of about 10,000. There are also Greek and Turkish army contingents—numbering 950 and 650, respectively—stationed on the island under the constitution. Unification of the police force on a 70:30 ratio ceased after 1963, and two separate de facto forces emerged, that of the Greeks—totalling about 2,500 officers and men—and a Turkish unit of about 500.

The number of British servicemen remaining in Cyprus in the sovereign base areas declined by 1975 to "around 5,000." Of these, the greater proportion are Royal Air Force personnel based on RAF Akrotiri in the south. Their presence is officially stated to be in support of NATO in southeastern Europe and of Cento (Central Treaty Organization) in the Middle East.

There are no formal boundaries between the British sovereign area and the territory of the Cyprus republic. Under the 1960 agreements, if Britain at any time relinquishes control of the bases, the area reverts to the Cyprus government's control.

Social development. *Education.* Compulsory elementary education was introduced under the British regime, and some 99 percent of children from nine to 12 were enrolled by the 1970s. Planning envisages the early extension of free education to the age of 15. Secondary education is on a fee-paying basis, and about 80 percent of those completing elementary school continue their education in some manner. Postsecondary facilities include schools for teacher training, technical instruction, hotel

Compulsory elementary education

catering, and nursing and midwifery; there is also a forestry college, generally acknowledged as the best in the Middle East.

There is no Cypriot university, but there is a steady flow of students taking degree courses in Greece, Great Britain, and elsewhere, though there is growing difficulty in absorbing graduates into suitable employment.

Health. Health standards are high, because of a favourable climatic situation and a well-organized public health service, but there is still a relative shortage of medical practitioners in rural areas. The ratio of doctors to populace is one of the best in the Middle East but below what the World Health Organization has laid down as the desirable minimum. Since malaria was eradicated shortly after World War II, the island has been free of major diseases, but there is some anxiety about the rapid rise in the incidence of hydatid disease (echinococcosis), a disease caused by the dog tapeworm that causes a cyst and may be fatal. It has spread in Cyprus to epidemic proportions because of an excessive population of stray dogs. A stringent campaign has been undertaken, and thousands of strays have been destroyed by municipal dog shooters.

Living conditions. Before independence, Cyprus was a low-income country with a low cost of living, but the surge of prosperity and development after 1964 created an inevitable inflationary pressure that was reflected in substantially increased wages and incomes, a steadily advancing cost of living, and an improving standard of living. There is in rural areas a certain degree of poverty among the elderly, but this is susceptible to steady eradication with the advance in per capita income. Average unemployment is very low.

Trade unions

Trade unionism was introduced in 1932, and by the 1970s it was a vigorous movement with a membership representing almost 60 percent of the nominal labour force. There is no restriction on the right to strike, except in essential services, but efficient conciliation machinery under the Labour Department enables labour relations to be maintained on a fairly peaceful level.

In 1961 employers formed a consultative association that became the Cyprus Employers Federation in 1970. It is recognized by both unions and government and takes part in all government committees that deal with labour matters.

Rapid social and economic development has stimulated housing construction, with the growing trend toward urbanization, but an accompanying unprecedented boom in land prices and building costs has led to the predominance of multi-unit apartment housing in place of the traditional Cypriot single-family dwelling. High-cost accommodation has had an adverse influence on a large proportion of the people who wish to own their own homes on an island in which there are virtually no mortgage facilities available. The greatest problem in meeting housing needs has been the lack of a comprehensive town and country planning authority.

CULTURAL LIFE

Government support of the arts

The arts in Cyprus have benefitted from the activities of the British Council, the Goethe Institut (Federal Republic of Germany), and the American Center, and from a growing interest on the part of the government, which has established an art gallery, a public library, a state theatre, and a state orchestra. Frequent visits by internationally known artists, orchestras, and ballet and theatrical companies have stimulated public interest. A useful contribution is also made by Cyprus Television, which makes generous use of imported educational and cultural film material. Some talented local artists have emerged, and their work has appeared in such international salons as the Milan Biennale.

The ancient theatres of Salamis, Curium, and Soli have been restored and are used for the staging of Greek and Shakespearean plays. A Greek theatre has been built in Nicosia.

Broadcasting within the jurisdiction of the republic is a monopoly service operated by the quasi-government Cyprus Broadcasting Corporation, founded in 1953. Broadcasts in Greek, Turkish, Armenian, and English are under strong official influence. A television service was introduced in 1957, and both this and radio are operated on a commercial basis.

There is also a station of the British Forces Broadcasting Service (BFBS), established in 1948 and directed chiefly to British servicemen on the island, and a Middle East relay station of the British Broadcasting Corporation (BBC). Neither of these two stations initiates news broadcasts. The Turkish Cypriots operate a radio station known as Radio Bayrak.

It is official government policy to encourage a free press, and there are daily newspapers in Greek, Turkish, and English and weeklies in Greek and Turkish. There is no formal press censorship.

PROSPECTS

At the end of 1975 the political and economic outlook for Cyprus appeared troublesome. The economy had suffered a severe decline after the Turkish invasion of 1974, and an estimated 70,000 Greek Cypriots were unemployed. Greeks displaced from the north were concentrated in the towns and villages of the south, with some 50,000 living in camps and temporary accommodations supported by the government and by refugee aid funds from international sources.

The situation of the Turkish Cypriot community was difficult, and the community was dependent upon an annual subvention from Turkey of about £14,000,000. Industrial establishments in the area occupied by the Turkish army were almost entirely idle, and there was little agricultural production.

(H.W.G.)

BIBLIOGRAPHY. *General and reference:* There is no current Cypriot national bibliography. The fullest listing of earlier materials on Cyprus is NEOCLES G. KURIAZES, *Kypriake bibliographia* (1936), containing about 5,000 titles. Among the most useful general descriptions of the country are the *Area Handbook for Cyprus* (1971), of the FOREIGN AREA STUDIES DIVISION OF THE AMERICAN UNIVERSITY; *Nagel's Encyclopedia-Guide to Cyprus*, designed for travellers but containing much well-organized background information, revised regularly; and MICHAEL and HANKA LEE'S *Cyprus* volume in the DAVID AND CHARLES "Islands Series." The most complete statistical description is the annual *Statistical Abstract*, published by the MINISTRY OF FINANCE.

The environment: The geological background is summarized in *A Synopsis of the Stratigraphy and Geological History of Cyprus*, by F.R.S HENSON and others (1949); for botanical information, the only comprehensive survey is J. HOLMBOE, *Studies on the Vegetation of Cyprus* (1914), but ESTHER F. CHAPMANN, *Cyprus Trees and Shrubs* (1949), is also useful; detail on wildlife may be found in *Birds of Cyprus*, by D.A. and W.M. BANNERMAN; no comprehensive study of animal life is available.

The people: For the historical background of the development of Cyprus' demographic situation, the best exposition is *The Population of Cyprus (1570–1881)*, by T. PAPADOPOULLOS (1965); for detail on the situation before the invasion of 1974, the reports of the censuses of 1960 and 1973; for materials on the minority community EMEL ESEN, *Aspects of Turkish Civilization in Cyprus* (1965), is recommended.

The national economy: In addition to the *Statistical Abstract* cited above, see also the MINISTRY OF FINANCE'S *Economic Report* (annual), as well as its more specialized serials covering other aspects of Cyprus' economic life. The best general discussions of the economy are to be found in A.J. MEYER, *The Economy of Cyprus* (1962), and D. JENNES, *The Economics of Cyprus* (1962). The materials published in the First, Second, and Third Five-Year plans (1962–66, 1967–71, and 1972–76) detail the programs established by the national government to improve the country's situation domestically and internationally.

Administration and social conditions: For background on Cyprus' situation in the region, JACOB C. HUREWITZ, *Diplomacy in the Near and Middle East*, 2 vol. (1956), provides extensive coverage; problems of constitutional government during the first decade of independence are examined in STANLEY KYRIAKIDES, *Cyprus: Constitutionalism and Crisis Government* (1968), and CRITON G. TORNARITIS, *Constitutional and Legal Problems in the Republic of Cyprus* (1968). Access to data on health and education is primarily through the statistical publications of the government (annually). Living

conditions are examined in the *Household Survey*, 3 vol., of 1971.

Cyprus, History of

The earliest remains so far discovered in Cyprus belong to the Late Neolithic Period. An important settlement not far from the south coast, near the village of Khirokitia, has been dated by carbon-14 tests to the mid-6th millennium BC. Architecture is represented by beehive dwellings, and the utensils in use were of stone. Flint and obsidian implements were current, and stone celts were found in great numbers. This earlier period, which may be called pre-ceramic, was followed by another period marked by the appearance of well-made pottery with painted patterns.

A new era is represented by the introduction of combed pottery, mostly found at the settlement of Sotira about five miles north of the classical town of Curium near the south coast. The Sotira culture has been dated to the second part of the 4th millennium BC. In the early 3rd millennium the combed pottery was gradually replaced by painted pottery, most of which has been found at the settlement of Erimi not far from the south coast. Houses were circular, and the general aspect of the culture shows substantial development. Copper then made its first appearance.

About 2500 BC red and black pottery, probably influenced by the Khirbet Kerak ware of Palestine, began to take the place of the painted wares, and in the 23rd century pottery shapes of western Anatolian types found their way into the island, probably introduced by people in search of copper. This period represents the Early Bronze Age.

The Middle Bronze Age (early 2nd millennium BC) is characterized by the return of painted pottery. In the Late Bronze Age (*c.* 1600–*c.* 1050 BC) Cyprus became the emporium of the East. The most important event during this period was the appearance of Mycenaean traders about 1400 BC, followed by successive waves of Achaeans, who colonized the island and introduced Greek culture and language.

During the Late Bronze Age Cyprus appears in Egyptian records under the name of Asi as a conquest of Thutmose III of Egypt (*c.* 1500 BC). Alashiya (Alasia), mentioned in Egyptian, Hittite, and Mesopotamian records, has also been identified as Cyprus, although doubts of this identification have been expressed.

The culture as reflected mostly in the ceramic production of the first part of the Iron Age (*c.* 1050–*c.* 950 BC) shows a fusion of Cypriot and Achaean characteristics, with some Syro-Anatolian elements. It appears, however, that the Achaean element gained the upper hand. The Cypriot institution of kingship, as it is known from later evidence, was established by the Achaean colonists, although pre-Achaean kingdoms did exist in Cyprus—for example, at Paphos. The character of the kingship, originally Mycenaean, developed later into an autocracy of Oriental type with the king as the head of the state, although the city-state of Idalium, in the south, enjoyed a more democratic form of government. Art, as reflected in pottery and other crafts of the Early Iron Age, is first dominated by the Achaean–Mycenaean features, but later (950–700 BC) Cypriot characteristics are seen to come to the fore.

Early Cypriot kingship

Commercial relations with Syria, Palestine, Egypt, and Anatolia were at first rare, but from the 9th century onward they became closer. Sidon and Tyre on the Syrian coast and Tarsus in Cilicia were centres through which Cypriot wares were transmitted to the interior of the continent. The relations with Greece followed a similar course, Rhodes being the main intermediary between Cyprus and the Greek mainland and islands.

The beginning of Phoenician penetration into Cyprus may be dated to about 800 BC, although a funeral inscription of the 9th century BC shows the appearance of Phoenician elements already in that century.

The city of Citium (Kition), a Tyrian colony, was the main centre of Phoenician activity, and the overwhelming majority of Phoenician inscriptions have been found in that city.

Assyrian domination. The year 709 BC marks the end of political independence with the submission of the Cypriot kings to the Assyrian king Sargon II. The event was commemorated at Sargon's capital, Dur Sharrukin (Khorsabad), as well as in Cyprus, where, at Citium, a stele, now in Berlin, was erected. On the stele Cyprus is referred to as Yatnana; *i.e.*, the isles of Danaoi.

The hold of Assyria over Cyprus continued during the reigns of Sargon's successors. Thus Esarhaddon's prism, written in 673–672 BC, refers to 10 kings and kingdoms of Cyprus, among which is Qarthadast (Citium). Although Cyprus appears as one of the countries paying tribute to Ashurbanipal (667 BC), the Assyrian domination may have ended in 669. From that date Cyprus had almost 100 years of virtual independence—one of the most brilliant periods of Cypriot culture (the first part of the Cypro-Archaic period). The power of the kings became greater, a fact reflected in the monumental tombs built of ashlar masonry. Palaces were built, and shipbuilding and mining attained a high level of production. Ceramic art produced some of the island's finest and most richly ornamented vases, while metalwork, as evidenced by some remarkable silver bowls, reached great perfection under Egyptian and Phoenician influence. The output of Greek epic poetry is also noteworthy; Stasinus, to whom the epic called *Cypria* is attributed, belongs to the 7th century.

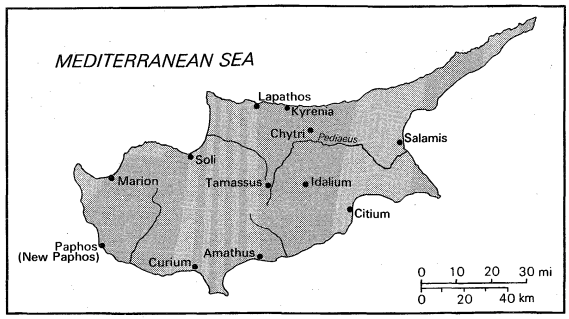
Commercial and cultural relations with the East and West were intensified and Cyprus maintained close intercourse with Rhodes, which continued to be the chief stepping-stone of Cypriot commercial expansion to the West. Mainland and eastern Greek pottery found its way to Cyprus, and Cypriot products were widely distributed in Greece, thus transmitting Oriental characteristics to Greek art.

Egyptian, Greek, and Persian influence. Following the breakup of the Assyrian Empire, Saite Egypt became the great power. Cyprus was necessary to Egypt for minerals and material for shipbuilding, and Ahmose II conquered the island in about 560. This meant a new orientation of Cypriot culture, and the influence of Egypt is reflected mostly in the sculpture and the increased importation of Egyptian goods, chiefly from the Greek colony of Naukratis in the Delta. The Cypriot kings and kingdoms were allowed to exist, but they paid tribute to the Egyptian ruler.

Around 525 BC the island passed to the control of Achaemenian Persia (see also IRAN, HISTORY OF). It formed a satrapy of the Achaemenian Empire, but the Cypriot kings were left undisturbed and struck their own coinage. The Greek cities joined the Ionian revolt about 498 BC, but the Phoenician colonies Citium and Amathus remained loyal to Persia. The rising was soon put down; in 480 Cyprus furnished 150 ships to the fleet of Xerxes and remained subject to Persia during the 5th century. But the Greek cities—the principal of which were then Salamis, Curium, Paphos, Marion, Soli, Kyrenia, and Chytri—retained monarchical governments throughout. Phoenicians held Citium and Amathus on the south coast, Tamassus and Idalium in the interior. At the end of the 5th century a league of Cypriot cities was formed by Evagoras, who became king of Salamis about 410. Aided by the Athenians, he revolted openly from Persia. But the Phoenician states supported Persia as before, the Greeks were divided by feuds, and by 381 the attempt had failed; Evagoras was assassinated in 374, and his son Nicocles died soon after. After the victory of Alexander the Great at Issus in 333, all the states of Cyprus welcomed the conqueror.

Greek cities

These political events are reflected in cultural and artistic development. Architecturally, this period is exemplified by the palace of Vouni, the history of which is interwoven with events of the 5th century. Excavations at Kouklia (Palaipaphos) have shown more evidence of the political history of these times. Sculpture, after first producing remarkable works under Ionian influence encouraged by the annexation of Cyprus to the Persian Empire,



Principal cities of ancient Cyprus.
After J.B. Bury, *A History of Greece* (1934); Macmillan London Ltd.

passed into relative stagnation. Evagoras' accession to the throne of Salamis encouraged Hellenic cultural relations but the effects were meagre. At this time Greek symbols and alphabet were introduced into the local coinage.

After Alexander's death in 323 Cyprus passed, after several rapid changes, to Ptolemy I, king of Egypt. Under Ptolemaic rule Cyprus was usually governed by a viceroy of the royal line, but it gained a brief independence under Ptolemy Lathyrus (107–89 BC) and under a brother of Ptolemy Auletes (80–58 BC). The great sanctuaries of Paphos and Idalium and the public buildings of Salamis, which were wholly remodelled in this period, have produced but few works of art. It is in this period that Jewish settlements are first heard of; these later became well populated. Following the annexation of Cyprus to Ptolemaic Egypt the local coinage ceased to exist and artistic production, mostly illustrated by sculpture, was chiefly dependent on the workshops of the Hellenistic world.

Roman period. In 58 BC Rome, which had made large unsecured loans to Ptolemy Auletes, annexed the island. Under Rome, Cyprus was at first appended to the province of Cilicia; after the defeat of Mark Antony at Actium (31 BC) by Octavian (Augustus), Cyprus became a separate province that remained in the hands of Augustus and was governed by a *legatus Caesaris pro praetore* as long as danger was feared from the East. In 22 BC, however, it was transferred to the Senate, so that Sergius Paulus, who was governor in AD 46, was rightly called proconsul.

Christianization

Christianity reached Cyprus very early. Barnabas, the colleague of the Apostle Paul, was a native of the island, and Paul himself preached there. In 115–117 the Jews of Cyprus, with those of Egypt and Cyrene, revolted, massacred 240,000 persons, and destroyed a large part of Salamis. Hadrian, afterward emperor, suppressed them and expelled all Jews from the island. For the culture of the Roman period there is abundant evidence from Salamis and Paphos, from tombs everywhere, and from the glass vessels, which almost wholly supersede pottery in tombs. Of the sculptural works, the most remarkable are the bronze statue of Lucius Septimius Severus in the Cyprus Museum and two marble sculptured sarcophagi in Cyprus. The most important of the types of coin struck in the island is that with the Aphrodite Temple on the reverse. (P.Di.)

Medieval period. After the division of the Roman Empire (AD 395), Cyprus remained subject to the Eastern emperors, with brief intervals, for more than 700 years (see also BYZANTINE EMPIRE). During the 5th century it formed part of the diocese of the Orient, under the *comes Orientis*, whose seat was at Antioch. This subordination was probably the basis of the claims advanced by the patriarch of Antioch to appoint and consecrate the metropolitan of Cyprus. These claims were successfully resisted by the Cypriot hierarchy. Both the Council of Ephesus in 431 and Council in Trullo in 692 recognized the autocephaly of the Cypriot church and its right to elect and consecrate its own archbishop and bishops.

The island seems to have enjoyed a peaceful existence until the Muslims began their attacks during the 7th century (see further ISLAM, HISTORY OF). The first was launched during the years 647–649, when Muʿa-wiya, amir of Syria, overran Cyprus and laid it under tribute. In 653 it was attacked again, and an Arab garrison was established there. This was withdrawn after 680, and for nearly 300 years Cyprus was to be neither under the full control of Islām nor of Constantinople. Each used Cyprus as a springboard for attacks on the other. The Byzantine fleet prevented permanent Muslim occupation, but their raids were frequent and sometimes prolonged, and it was not until 965 that the emperor Nicephorus II Phocas was able to deliver the island from the constantly renewed attacks. For the next 200 years it remained an integral part of the Byzantine Empire. As an outlying province, it was not infrequently used as the base for revolt against the imperial government. From about 1185 Isaac Comnenus, an ambitious Byzantine governor, held it against two successive emperors, his kinsman Andronicus I and Isaac II Angelus, but in 1191 he showed hostility to the English crusaders led by King Richard I, who then conquered the island. Richard first sold his conquest to the Knights Templars and then, as they could not pay, bestowed it on Guy de Lusignan, the dispossessed king of Jerusalem.

The Latin kingdom

Guy founded a feudal monarchy in Cyprus that survived until the end of the Middle Ages. Much of its recorded history during that period concerns the life of the court and of the territorial aristocracy, which, in contrast to the peasantry, was French in speech and culture. Because of the island's trade, however, Italian merchants, especially those of Genoa and Venice, assumed an ever increasing economic and political control until, in the 15th century, Cyprus became part of the Venetian Empire.

The kings of Cyprus kept alive the crusading idea (see also CRUSADES, THE). Some, such as Amalric II (1194–1205) and Hugh III (1267–84), were also kings of Jerusalem, but they were unable either to impose unity on the warring factions that weakened the Syrian kingdom or to prevent its extinction in 1291 by the Mamlūks of Egypt. Even then, Cyprus remained a base for counterattack against the Muslims. Hugh IV (1324–59) formed an alliance with the papacy, Venice, and the Knights Hospitallers of Rhodes, and in 1344 an expedition fitted out by this league retook Smyrna (Izmir) from the Turks. His son Peter I (1359–69) devoted himself to the organization of a crusade. In 1361 he took Gorigos (Korykos) and Adalia (Antalya) in Anatolia. In 1362 he recruited an army and fleet in western Europe and took Alexandria (Oct. 10, 1365). But he could neither retain the conquest nor persuade the rulers of western Europe to equip another crusade.

After Peter's assassination in 1369 the fortunes of the royal house declined, while those of the Italian republics rose to new heights. The economic activity in Cyprus was a result not only of its own silks, sugar, and salt but also of its importance as an entrepôt for the Eastern trade, especially after the Christians lost the Syrian ports. All the important commercial communities of the Mediterranean therefore established stations and bought trading privileges in Cyprus. In 1373 the Genoese took advantage of the local political situation to acquire complete control of Famagusta, the principal port, and were able to dominate trading and government in the island for 90 years. In 1426 an expedition launched from Egypt overran the island, and for a generation its kings were nominally the tributaries of Cairo.

The situation was redeemed by King James II (1460–73), a bastard of the royal house who in his father's lifetime had become archbishop of Nicosia. With the help of a Muslim force from Egypt he seized power in Cyprus and expelled the Genoese from Famagusta (1464). In 1472 he allied himself to Venice, by marrying a Venetian noblewoman. On his death Cyprus was ruled from Venice in the Queen's name until in 1489 the republic caused her abdication; the island remained a Venetian possession until its conquest by the Ottoman Turks in 1570–71.

Venetian rule

Medieval relics. From the period between the peace of the church and the devastating Arab raids of the mid-7th century there survive numerous remains, including basil-

ican churches, an apse mosaic at Lythrangomi comparable to those of the 6th century at Ravenna, Italy, a bath complex at Curium with fine mosaic pavements, and a treasure from the site of the ancient Lapithos consisting of gold ornaments and silver plate. Wall paintings of Byzantine and later styles exist in churches still in use, some domed in the Byzantine fashion, some wood-roofed and tiled. Outstanding monuments of the Latin kingdom are the Gothic cathedrals of Nicosia and Famagusta, and the abbey of Bellapais. Of military architecture, Byzantine, Frankish, and Venetian, a good deal remains, foremost being the walled city of Famagusta. (Ed.)

Ottoman and British rule. On March 7, 1573, Venice recognized Ottoman sovereignty over Cyprus. The Turkish administration lasted for 300 years. At first comparatively enlightened (serfdom was abolished, the Orthodox archbishopric restored, and the Christian population granted a large measure of autonomy), it became neglectful and often oppressive during the period of Ottoman decline. There were serious risings in 1764, 1804, and 1821. In 1838–39 and 1856 attempts were made to introduce reforms and some self-government with a local *dīwān* (council). The reform movements lagged, however, and there were few internal improvements, since the imperial treasury retained most of the island's income. In the courts, Christian testimony was seldom honoured.

On June 4, 1878, Great Britain took over the administration of Cyprus and occupied it by agreement with the Ottoman sultan, who was promised British help against Russia in return. Cyprus remained under Turkish sovereignty, but from the start elements of the Greek population (80 percent of the total) asked the British to grant *énosis* (union with Greece). For many years this demand was of little more than academic significance.

World War I

Cyprus was annexed to the British crown in November 1914, when war with Turkey broke out. The British government offered the island to Greece 12 months later, if Greece helped Serbia, then invaded by Bulgaria. Greece refused to help Serbia, and the offer lapsed. Turkey formally recognized the annexation under the Treaty of Lausanne on July 24, 1923, and in the following year Cyprus was made a crown colony.

In 1931 serious disturbances arose out of the demand put forward by sections of the Greek population for *énosis*. The Legislative Council was abolished and power to legislate was vested in the governor alone.

Archaeological exploration. This began before the British occupation (T.B. Sandwith, R.H. Lang, L.P. di Cesnola), but it was after the occupation that systematic work was started (under the Cyprus Exploration Fund and by O. Richter). In the early part of the 20th century Sir J.L. Myres and M. Markides carried out excavations for the Cyprus Museum, and later a Swedish expedition devoted several years (1927–31) to exploring many sites. The Cyprus Museum and (from 1935) the department of antiquities investigated many prehistoric and later sites, while expeditions sponsored by the Pennsylvania University Museum, the British School at Athens, and the Louvre, Paris, explored Bronze Age cemeteries and townsites. After 1950 the British excavated sites at Myrtou and at Kouklia (Palaipaphos), and Swedish and French missions undertook excavations at Kalopsida, Ayios Iakovos, and Enkomi. The department of antiquities conducted large-scale excavations at Salamis.

The énosis movement. Apart from a few air raids, Cyprus remained immune from enemy attack throughout World War II. After the war agitation for *énosis* was resumed. Archbishop Makarios III, patriarch of the Orthodox Church in Cyprus, emerged as the leader of the movement, with the support of various political groups. The campaign for *énosis* grew more violent, with persistent attempts to smuggle arms from abroad. This provoked great sympathy in Greece and resentment in Turkey. At the London Conference in September 1955 the British, Greek, and Turkish foreign ministers failed to agree on a solution, and an atmosphere of terrorism, repression, and mistrust prevailed. In this the lead was taken by EOKA (Ethnikí Orgánosis Kipriakoú Agónos, or National Organization of Cypriot Struggle) headed by Col. Georgios Grivas ("Digenes" or "Dighenis"), formerly an officer in the Greek Army. EOKA enjoyed wide support in the Greek Cypriot community and included a guerrilla body that carried out terrorist attacks on British servicemen and establishments. British reinforcements were sent to the island to suppress the guerrillas, and in the autumn of 1955 Field Marshal Sir John Harding was appointed governor. Emergency regulations were introduced. Negotiations opened with Makarios, the main issue being self-determination, which the British government refused to concede at that time although it was ready to grant a large measure of self-government.

Revolt against British rule

In March 1956 Makarios and the Bishop of Kyrenia were deported to the Seychelles. For a year terrorism raged throughout Cyprus, but all attempts to suppress EOKA failed. In the meantime Lord Radcliffe had been asked to advise on a new liberal constitution. His proposals, published in December 1956, were accepted by the British government. On March 14, 1957, EOKA offered to suspend its activities if Makarios was released from exile. A few days later he was freed and permitted to go where he liked except Cyprus. Terrorist activities ceased, and many restrictions were lifted, but progress toward a settlement was slow. Sir John Harding was succeeded in December 1957 by Sir Hugh Foot. By this time the Turkish Cypriot community (about 17.5 percent of the island's population), led by Fazil Kütchük, was thoroughly aroused at the prospect of *énosis* and put forward the demand that the island should be partitioned. Renewed efforts in 1958 by the British government and by NATO to produce a settlement came to nothing in face of the Greek community's desire for union with Greece and the Turkish for partition. Terrorism broke out again and fresh restrictions had to be imposed. But later that year the Archbishop proposed that after a fixed period of self-government Cyprus should become an independent state.

The attainment of independence

This led to negotiations between the Greek and Turkish governments and to a conference in Zürich, Switzerland, in February 1959, at which the British were not represented. Agreement in principle was reached, which was immediately approved by the British government. British sovereignty was to be retained over the two military bases at Dhekhelia and Akrotiri, in all an area of 99 square miles (256.4 square kilometres). The new republic would not participate in a political or economic union with any other state, nor would it be subject to partition. Greece, Turkey, and the United Kingdom guaranteed the independence, integrity, and security of the republic, and Greece and Turkey undertook to respect the integrity of the areas remaining under British sovereignty. Executive power was vested in a Greek Cypriot president and a Turkish Cypriot vice president, to which posts Makarios and Kütchük had previously been elected. They would have a council of ministers comprising seven Greek and three Turkish Cypriots. It was later agreed that the decisions of the council would be binding on the president and vice president, who could, however, impose a veto in matters relating to security, defense, and foreign affairs. There would also be a House of Representatives elected for five years consisting of 50 members, of whom 70 percent were to be elected from the Greek and 30 percent from the Turkish Cypriot communities, and the civil service was to consist of 70 percent Greek and 30 percent Turkish Cypriots. A Turco-Greek military headquarters was to be set up to train the Cypriot Army.

The Republic of Cyprus. The first general election took place on July 31, 1960. Of the 35 seats allotted to the Greek Cypriots 30 were won by supporters of the Archbishop, five by agreement being allotted to AKEL (Anorthotikón Kómma Ergazoménon Laoú), the Cypriot Communist Party. All 15 seats allotted to the Turkish Cypriots were won by supporters of Kütchük. The republic came into being on Aug. 16, 1960, and Cyprus was admitted as a member of the United Nations. It remained a member of the sterling area, and the British

government agreed to make £12,000,000 of financial assistance available, payable over five years. Cyprus was admitted to membership in the Commonwealth of Nations in March 1961. (B.S.-E.)

Foreign influences

The new nation was not to know peace for long. The tensions between the Greek and Turkish Cypriots continued, backed by their mother countries with their own areas of conflict. Members of the NATO alliance, particularly Great Britain and the U.S., were concerned about two fellow members at such constant odds and the vulnerability of their southern flank to Soviet pressure. Outbreaks of violence continued to be frequent, and in March 1964 the UN Security Council agreed to send to Cyprus a mixed force (Unficyp, called the United Nations Peace-Keeping Force in Cyprus); its mandate was extended repeatedly over the next decade.

Makarios as president. Archbishop Makarios had been a leader of *énosis* in the Cypriots' struggle against Great Britain, but after he became president of an independent country his position grew increasingly toward the maintenance of this independence. With Athens' support, primarily through Gen. Georgios Grivas, who led the movement from 1964 to 1967 and again from 1971 until his death in January 1974, the underground terrorist activities of the former EOKA (later known as EOKA B) were a major cause of turmoil. The Turkish Cypriots, with strong support from Ankara, tried repeatedly to assure and improve their minority status and were also responsible for numerous incidents, including a near invasion by Turkey in 1967. Attempts on the Archbishop's life in 1970 and 1973, the assassination of a former government minister (probably involved in the earlier attempt on Makarios), and the kidnapping of an incumbent minister in 1973 (all assumed to be backed by the Greek Cypriots) were only a few of the frequent acts of violence. In contrast, however, the Archbishop was re-elected president for a second five-year term in 1968 by an overwhelming 220,911 votes to 8,577, and in 1973 he won an uncontested third term.

Conflict with churchmen

Makarios also came into conflict with churchmen. In 1972 three Cypriot bishops demanded that he renounce the presidency because it conflicted with his role as spiritual leader. He steadfastly refused to give up the presidency, so in April 1973 the three bishops called a council and declared that the Archbishop was deposed from his spiritual office. This action prompted the convening in July of a synod of the Eastern Orthodox churches under the Patriarch of Alexandria. The synod voided the earlier deposition and stripped all three of the rebel bishops of their jurisdictions and even their priestly offices.

Relations between the president and the Cyprus National Guard, which was officered by Greeks, deteriorated steadily. On July 15, 1974, the inevitable coup occurred. The guard placed Nikos Sampson, a Cypriot newspaper publisher and former guerrilla leader, in the presidency, and Makarios barely escaped with his life. Five days later, with the professed aim of overturning Sampson's government, Turkish forces landed on Cyprus, at Kyrenia. On July 23 the President of Greece dismissed his Cabinet and recalled the self-exiled Konstantinos Karamanlis to form a new government. That same day Sampson was relieved of his post by the Cyprus National Guard, and Glafkos Clerides, president of the House of Representatives, was made interim president. Meetings in Geneva of the three guarantors of the Cyprus constitution, Great Britain, Greece, and Turkey, resulted in several cease-fire agreements, but none was effective until August 16, by which time Turkey controlled the northern third of the island.

Intercommunal talks. Intercommunal talks had begun in 1968 between Clerides and Rauf Denktash, leaders, respectively, of the Greek and Turkish Cypriots, to find a solution to the impasse. The Turks had established a "transitional administration" to administer the affairs of the Turkish Cypriot community until the constitutional provisions of 1960 were fully effective; the Greek community refused to recognize this body. The talks were deadlocked and resumed repeatedly through the years. Early in 1974 the new Turkish prime minister, Bulent Ecevit, demanded that Cyprus form a bi-zonal federation of two separate states with a weak central government. The Greek Cypriot position seemed to evolve into a desire for nationalistic cantons to be held together by a strong central government, a plan that Makarios had earlier opposed. After the Turkish invasion, with Clerides as acting president, the talks centred on the exchange of prisoners and the relocation of refugees. Even after Makarios returned triumphantly to resume the presidency in late 1974, negotiations settled very little. In February 1975 the Turkish Cypriots proclaimed their occupied area to be a separate state. Denktash explained that their intention was not to become independent but, in effect, to accomplish the federation under the 1960 constitution that they had been advocating. The Greek Cypriots objected strongly to the action, as did Greece, but the separate state was now an accomplished fact. Negotiations on the fate of refugees continued. (Ed.)

BIBLIOGRAPHY. The most comprehensive history is that of GEORGE F. HILL, *History of Cyprus*, 4 vol. (1940–52). The later period, covering the acquisition of independence, is dealt with in official publications, notably *Tripartite Conference on the Eastern Mediterranean and Cyprus . . .*, Cmd 9594 (1955), and *Cyprus*, Cmd 1093 (1960), giving text of agreements relating to independence. Since independence, the periodical reports to the United Nations Security Council of the Secretary General requesting renewal of the mandate of UNFICYP (1964–) are indispensable. See also the ROYAL INSTITUTE OF INTERNATIONAL AFFAIRS, LONDON, *Cyprus: The Dispute and the Settlement* (1959); L. DURRELL, *Bitter Lemons* (1957); GEORGE GRIVAS, *The Memoirs of General Grivas* (1964); R.H. STEPHENS, *Cyprus: A Place of Arms* (1966); and P.N. VANEZIS, *Makarios: Faith and Power* (1972). On economics until independence, the annual reports of the Government of Cyprus are informative; since independence, see W.L. THORP, *Cyprus: Suggestions for a Development Programme* (1961); and *The Second Five Year Plan (1967–1971)* (1967). On archaeology, see especially P. DIKAIOS, *A Guide to the Cyprus Museum*, 2nd ed. (1953), and E. GJERSTAD *et al.*, *Swedish Cyprus Expedition*, 4 vol. (1934–37).

(B.S.-E.)

Cyrus II the Great, of Persia

Conqueror and Persian statesman, Cyrus was the founder of the Achaemenian dynastic empire, the first "world state," which included all of the Near East from the Aegean Sea to the Indus River. He is also remembered in the Cyrus legend—first recorded by Xenophon, Greek soldier and author, in his *Cyropaedia*—as a tolerant and ideal monarch who was called father of his people by the ancient Persians, and in the Bible as the liberator of the Jews captive in Babylonia.

Cyrus was born between 590 and 580 BC, either in Media or, more probably, in Persis, the modern Fars province of Iran. The meaning of his name is in dispute, for it is not known whether it was a personal name or a throne name given to him when he became a ruler. It is noteworthy that after the Achaemenid empire the name does not appear again in sources relating to Iran, which may indicate some special sense of the name.

Most scholars agree, however, that Cyrus the Great was at least the second of the name to rule in Persia. One cuneiform text in Akkadian—the language of Mesopotamia (present-day Iraq) in the pre-Christian era—asserts he was the

> son of Cambyses, great king, king of Anshan, grandson of Cyrus, great king, king of Anshan, descendant of Teispes, great king, king of Anshan, of a family [which] always [exercised] kingship.

In any case, it is clear that Cyrus came from a long line of ruling chiefs.

The most important source for his life is the Greek historian Herodotus. The idealized biography by Xenophon is a work for the edification of the Greeks concerning the ideal ruler, rather than a historical treatise. It does, however, indicate the high esteem in which Cyrus was held, not only by his own people, the Persians, but by the Greeks and others. Herodotus says that the Persians called Cyrus their father, while later Achaemenid rulers were not so well regarded. The story of the childhood of

Cyrus, as told by Herodotus with echoes in Xenophon and the Greek historian Ctesias, may be called a Cyrus legend since it obviously follows a pattern of folk beliefs about the almost superhuman qualities of the founder of a dynasty. Similar beliefs also exist about the founders of later dynasties throughout the history of Iran. According to the legend, Astyages, the king of the Medes and overlord of the Persians, gave his daughter in marriage to his vassal in Persis, a prince called Cambyses. From this marriage Cyrus was born. Astyages, having had a dream that the baby would grow up to overthrow him, ordered Cyrus slain. His chief adviser, however, instead gave the baby to a shepherd to raise. When he was 10 years old, Cyrus, because of his outstanding qualities, was discovered by Astyages, who, in spite of the dream, was persuaded to allow the boy to live. Cyrus, when he reached manhood in Persis, revolted against his maternal grandfather and overlord. Astyages marched against the rebel, but his army deserted him and surrendered to Cyrus, about 550 BC. So much for the Cyrus legend.

Cyrus' conquests

After inheriting the empire of the Medes, Cyrus first had to consolidate his power over Iranian tribes on the Iranian plateau before expanding to the west. Croesus, king of Lydia in Asia Minor, had enlarged his domains at the expense of the Medes when he heard of the fall of Astyages, and Cyrus, as successor of the Median king, marched against Lydia. Sardis, the Lydian capital, was captured in 547 or 546, and Croesus was either killed or burned himself to death, though according to other sources he was taken prisoner by Cyrus and well treated. The Ionian Greek cities on the Aegean Sea coast, as vassals of the Lydian king, now became subject to Cyrus, and most of them submitted peacefully. Several revolts of the Greek cities were later suppressed with severity. Next Cyrus turned to Babylonia, where dissatisfaction of the people with the ruler Nabonidus gave him a pretext for invading the lowlands. The conquest was quick, for even the priests of Marduk, the national deity of the great metropolis of Babylon, had become estranged from Nabonidus. In October 539 BC, the greatest city of the ancient world fell to the Persians.

In the Bible (*e.g.*, Ezra 1:1–4), Cyrus is famous for freeing the Jewish captives in Babylonia and allowing them to return to their homeland. Cyrus was also tolerant toward the Babylonians and others. He honoured Marduk and conciliated the local population by supporting local customs and even sacrificing to local deities. The capture of Babylon delivered not only Mesopotamia into the hands of Cyrus but also Syria and Palestine, which had been conquered previously by the Babylonians. The ruler of Cilicia in Asia Minor had become an ally of Cyrus when the latter marched against Croesus, and Cilicia retained a special status in Cyrus' empire. Thus it was by diplomacy as well as force of arms that he established the largest empire known until his time.

Cyrus seems to have had several capitals. One was the city of Ecbatana, modern Hamadan, former capital of the Medes, and another was a new capital of the empire, Pasargadae, in Persis, said to be on the site where Cyrus had won the battle against Astyages. The ruins today, though few, arouse admiration in the visitor. Cyrus also kept Babylon as a winter capital.

No Persian chauvinist, Cyrus was quick to learn from the conquered peoples. He not only conciliated the Medes but joined them with the Persians in a kind of dual monarchy of the Medes and Persians. Cyrus had to borrow the traditions of kingship from the Medes, who had ruled an empire when the Persians were merely their vassals. It is probable that a Mede was traditionally made an adviser to the Achaemenid king, as a sort of chief minister; on later reliefs at Persepolis, a capital of the Achaemenid kings from the time of Darius, a Mede is frequently depicted together with the great king. The Elamites, indigenous inhabitants of Persis, were also the teachers of the Persians in many ways, as can be seen, for example, in the Elamite dress worn by Persians and by Elamite objects carried by them on the stone reliefs at Persepolis. There also seems to have been little innovation in government and rule, but rather a willingness to borrow, combined with an ability to adapt what was borrowed to the new empire. Cyrus was undoubtedly the guiding genius in the creation not only of a great empire but in the formation of Achaemenid culture and civilization.

Little is known of the family life of Cyrus. He had two sons, one of whom, Cambyses, succeeded him; the other, Bardiya (Smerdis of the Greeks), was probably secretly put to death by Cambyses after he became ruler. Cyrus had at least one daughter, Atossa (who married her brother Cambyses), and possibly two others, but they played no role in history.

When Cyrus defeated Astyages he also inherited Median possessions in eastern Iran, but he had to engage in much warfare to consolidate his rule in this region. After his conquest of Babylonia, he again turned to the east, and Herodotus tells of his campaign against nomads living east of the Caspian Sea. According to the Greek historian, Cyrus was at first successful in defeating the ruler of the nomads—called the Massagetai—who was a woman, and capturing her son. On the son's committing suicide in captivity, his mother swore revenge and defeated and killed Cyrus about 529 BC. Herodotus' story may be apocryphal, but Cyrus' conquests in Central Asia were probably genuine, since a city in farthest Sogdiana was called Cyreschata, or Cyropolis, by the Greeks, which seems to prove the extent of his Eastern conquests.

The legacy of Cyrus

It is a testimony to the capability of the founder of the Achaemenid empire that it continued to expand after his death and lasted for more than two centuries. But Cyrus was not only a great conqueror and administrator; he held a place in the minds of the Persian people similar to that of Romulus and Remus in Rome or Moses for the Israelites. His saga follows in many details the stories of heroes and conquerors from elsewhere in the ancient world. The manner in which the baby Cyrus was given to a shepherd to raise is reminiscent of Moses in the bulrushes in Egypt, and the overthrow of his tyrannical grandfather has echoes in other myths and legends. There is no doubt that the Cyrus saga arose early among the Persians and was known to the Greeks. The sentiments of esteem or even awe in which Persians held him were transmitted to the Greeks, and it was no accident that Xenophon chose Cyrus to be the model of a ruler for the lessons he wished to impart to his fellow Greeks.

In short, the figure of Cyrus has survived throughout history as more than a great man who founded an empire. He became the epitome of the great qualities expected of a ruler in antiquity, and he assumed heroic features as a conqueror who was tolerant and magnanimous as well as brave and daring. His personality as seen by the Greeks influenced them and Alexander the Great, and, as the tradition was transmitted by the Romans, may be considered to influence our thinking even now. In the year 1971, Iran celebrated the 2,500th anniversary of the founding of the monarchy by Cyrus.

BIBLIOGRAPHY. R.N. FRYE, *The Heritage of Persia*, pp. 78–87 (1963), gives a general survey with archaeological discoveries at Pasargadae and elsewhere in Iran. HERODOTUS' *History*, XENOPHON'S *Cyropaedia*, and CTESIAS' *Fragments* are not only the principal but practically the only sources on Cyrus; later works only copy these. A.T.E. OLMSTEAD, *History of the Persian Empire*, pp. 34–67 (1948), is a highly readable and scholarly, occasionally imaginative, account of Cyrus.

(R.N.F.)

Czechoslovakia

The position of Czechoslovakia in the heart of Europe has been of great significance, affecting the economic, political, and cultural development of a country in which the most varied of influences and traditions have encountered one another. The elongated shape of the country—its east–west extent is more than 465 miles (748 kilometres), whereas its maximum north–south extent is barely 171 miles—has further differentiated its internal physical and human geography, western European influences being strong in Bohemia while those of eastern Europe predominate in eastern Slovakia. The nation's position on the boundary between Communist and capitalist systems has added a new element to this situation.

An inland country, Czechoslovakia (formally the Czechoslovak Socialist Republic) lies across the great ancient trade routes of Europe, following the Elbe (Labe) River to the North Sea, the Oder (Odra) River to the Baltic, and the Danube (Dunaj) River to the Black Sea. There were also routes to the Mediterranean. Modern Czechoslovakia came into being at the end of World War I, following the collapse of the Austro-Hungarian Empire. By the mid-1970s its area of 49,374 square miles (127,877 square kilometres) held a population of well over 14,850,000. It is subdivided into the Czech Socialist Republic (ČSR), of about 30,450 square miles (78,863 square kilometres), comprising the historic lands of Bohemia and Moravia (often called the Czech Lands) and the Austrian portion of Silesia; and the Slovak Socialist Republic (SSR) of about 18,920 square miles (49,014 square kilometres), which before World War I was a part of Hungary inhabited mainly by Slovaks.

Czechoslovakia is a member of the Council for Mutual Economic Assistance (Comecon), the Communist-bloc trade group, and also of the Warsaw Treaty Organization. Three-quarters of its frontiers are with other Communist countries—the German Democratic Republic in the northwest, Poland in the north, the Soviet Union in the east, and Hungary in the southeast—the remaining borders being with Austria in the southwest and the Federal Republic of Germany in the west. The capital is Prague.

This article is concerned with the contemporary country. For history, see BOHEMIA AND CZECHOSLOVAKIA, HISTORY OF. For information on related subjects, see CARPATHIAN MOUNTAINS; DANUBE RIVER; and PRAGUE.

THE LANDSCAPE

The natural environment. *Topography.* Czechoslovakian geographers divide the country into two large provinces, the Bohemian Massif (České masív) in the west and the Carpathian Mountains in the east; each of these provinces is divided into several smaller regions.

Bohemian Massif

The Bohemian Massif, also called the Bohemian Highlands (Česká vysočina), forms a large elevated basin (the Bohemian Plateau, or Česká tabule) encircled by mountain ranges that occasionally reach heights exceeding 3,000 feet (900 metres) above sea level. These include the Šumava, Český les (Bohemian Forest), Krušné hory (Ore Mountains), Krkonoše, Hrubý-Nizký, Hrubý Jeseník, Nizký Jeseník, and the Bohemian-Moravian Uplands.

The massif may be divided into seven major orographic units. The largest is the Southern Bohemian Highlands (Jihočeská vysočina), which comprise the older, crystalline core of the massif and of which the most important parts are the Bohemian-Moravian Uplands, the Central Bohemian Highlands (Středočeská pahorkatina), the Southern Bohemian Basin (Jihočeské pánve; centred on the lake district to the east of České Budějovice), and, encircling the southwestern end of the country, the Šumava Subsystem (Podsoustava Šumavy; comprising the Český les and Šumava proper).

Northwest of the Southern Bohemian Highlands are the Berounka River Highlands (Poberounská vrchovina). The city of Plzeň lies at the centre of this region. To the northwest of these ranges are the Krušné hory, which form the frontier with the German Democratic Republic, comprising two elongated subsystems paralleling the border. The lowest point in the Bohemian Highlands (384 feet [117 metres]) is found here, where the Elbe (in Czechoslovakia called the Labe) breaches these ranges as it enters East Germany.

To the east of the Krušné hory are the Sudeten (Czech and Polish Sudety) mountains, which form most of the border between these two countries west of Ostrava and in which is found the greatest altitude in the Bohemian Highlands, 5,256 feet (1,602 metres), at Sněžka in the Krkonoše range.

The Brnenska (Brno) Highlands, south of the central Sudeten, are the fifth and last of the uplands which enclose the Bohemian Plateau; here is found the spectacular Moravian Karst.

The plateau itself, also called the Labe plain, or Polabí, after the Labe River, which traverses it from east to west, receives the Vltava on its left bank, about 18 miles (30 kilometres) below Prague, which stands among the hills at the southern edge of the Polabí.

The seventh, and smallest, region is the Oder Lowland (Oderska nížina), a small fringe along the Polish border that may be considered to be a small foreland of the eastern Sudeten.

Carpathian Mountains

The Carpathian Mountains dominate the eastern part of the country; they comprise a system of east–west ranges separated by valleys and intermontane basins. The highest ranges are the Nízke Tatry (Low Tatras), reaching about 6,000 feet, and the Vysoké Tatry (High Tatras), containing the highest point in the republic, Gerlachovský štít (Gerlach Shield), 8,711 feet (2,655 metres).

Four orographic regions are recognized in this province: the Outer Carpathian Depressions (Vnékarpatské sníženiny), the Inner Carpathian Depressions (Vnitrokarpatské sníženiny), the Outer Carpathians, and the Inner Carpathians.

The Outer Carpathian Depressions, from the geographer's viewpoint called the Moravian Corridor, are the narrow lowlands that separate the Bohemian Massif from the easternmost Carpathians; these include the valleys of the upper Oder (Odra) and Morava rivers and the headstreams of the Dyje (although not the lower courses of the last two, as these belong to the Inner Carpathian Depressions). Two other large lowland areas comprise the remainder of the Inner Carpathian Depressions region; these are the Danubian Lowland (Podunajská nížina) and the Eastern Slovakian Lowland (Východoslovenská nížina), which are probably better classified as outliers of the Hungarian Plain, to the south, than as intermontane basins.

The Carpathian Mountains proper consist of two regions, the Outer Carpathians, to the north, composed of flysch, with a well-developed nappe (horizontally overthrust and folded sedimentary beds) structure; to the south, the Inner Carpathians, composed of a crystalline core mantled by Mesozoic sediments and, farther south, by Neogene volcanic series.

According to the definitions of Czechoslovak geographers, mountains and high mountains (lands above 600 metres [1,968 feet]) occupy 23 percent of the country, uplands (200 to 600 metres [656 to 1,968 feet]) occupy 67 percent, and lowlands (below 200 metres [656 feet]) only a scant 9 percent.

Climate. The Czechoslovak climate is mixed, its continental tendencies being masked by large fluctuations in both temperature and precipitation and oceanic influences diminishing from west to east. Relief variations show up in temperature figures: the mean annual temperature drops to 25.4° F (−3.7° C) in the High Tatras, rising to 50.9° F (10.5° C) in the Danube lowlands. Average July temperatures exceed 68° F (20° C) in the Danube lowlands, and average January temperatures can be as low as 23° F (−5° C) in mountain basins. The growing season is about 200 days in the south and less than half that in the mountains. Annual precipitation ranges from 18 inches (450 millimetres) in the central Bohemian basins, to more than 60 inches (1,500 millimetres) on windward mountain slopes in the Krkonoše (Giant Mountains). Maximum precipitation falls in July, minimum in February. There are no recognizable climatic zones but rather a succession of small and varied districts; climate thus follows the topography in contributing to the diversity of the natural environment.

Temperature and precipitation

Soils. Rich, black chernozem soils (10 percent of the total) and good-quality brown soils (another 20 percent) occupy the drier and lower portions of the country. The soils known as podzols, occurring in regions with more than 23 inches (600 millimetres) of rainfall annually, constitute 50 percent of all of the soils in the country but only a third of the agricultural land. Stony mountain soils in the highest regions and rich alluvial soils alongside river courses are characteristic. This soil variety is again further complicated by topographic variations, with loamy-sandy soils commonest in the Bohemian Massif and heavy clay soils predominant in peripheral zones of the Carpathians.

MAP INDEX

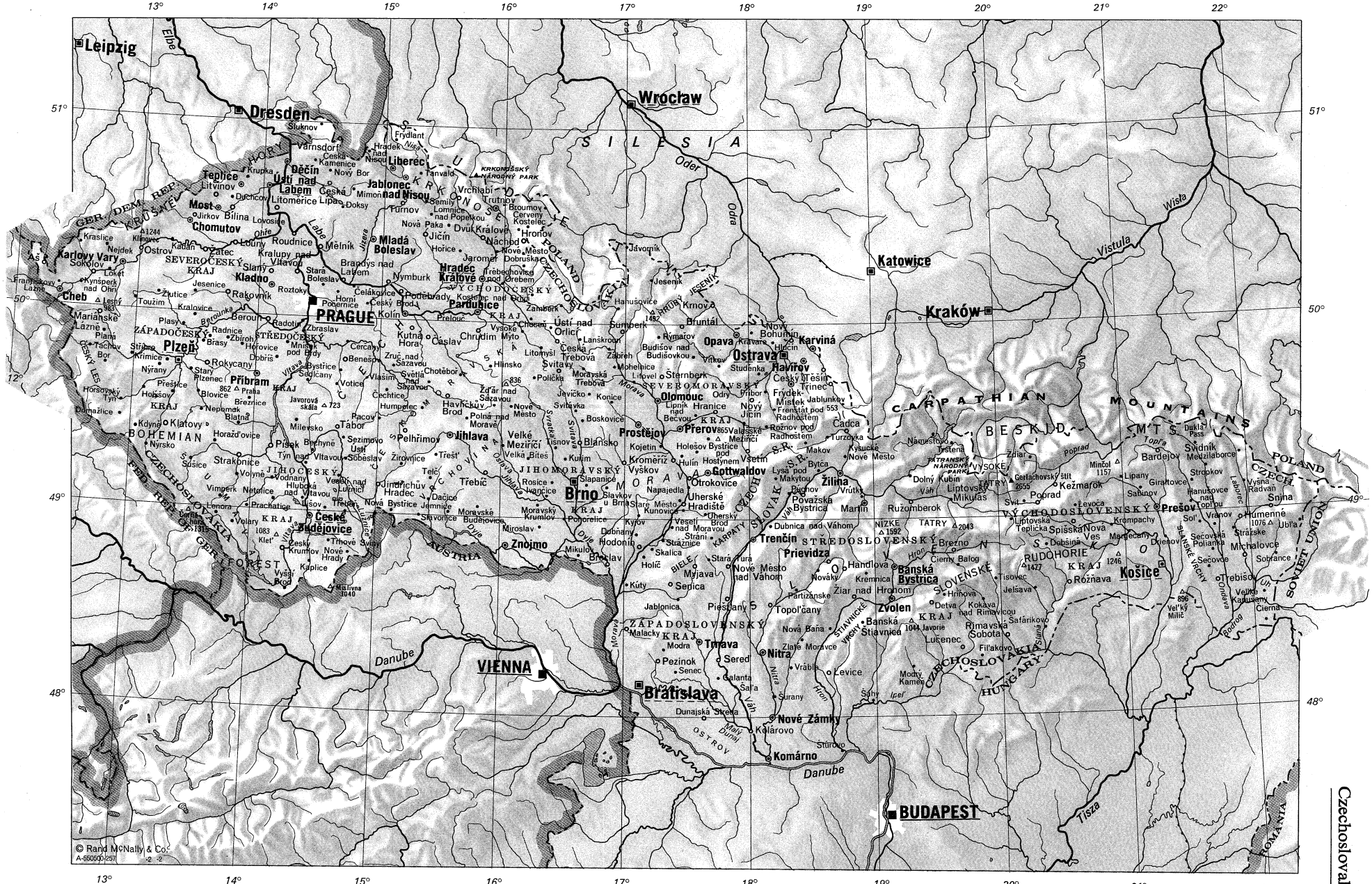
Political subdivisions

Bratislava....48·09n 17·07e
Czech Socialist Republic....49·45n 15·00e
Jihočeský....49·05n 14·30e
Jihomoravský....49·10n 16·40e
Prague....50·05n 14·26e
Severočeský....50·35n 14·15e
Severomoravský....49·45n 17·50e
Slovak Socialist Republic....48·45n 19·30e
Středočeský....49·55n 14·30e
Středoslovenský....48·50n 19·10e
Východočeský....50·10n 16·00e
Východoslovenský....49·00n 21·15e
Západočeský....49·45n 13·00e
Západoslovenský....48·20n 18·00e

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Aš....50·10n 12·10e
Banská Bystrica....48·44n 19·07e
BanskáŠtiavnica....48·28n 18·56e
Bardejov....49·18n 21·16e
Bechyně....49·18n 14·29e
Benešov....49·47n 14·43e
Beroun....49·58n 14·04e
Bílina....50·40n 14·02e
Blansko....49·22n 16·39e
Blatná....49·26n 13·53e
Blovice....49·35n 13·33e
Bor....49·43n 12·47e
Boskovice....49·29n 16·40e
Brandýs nad Labem....50·10n 14·41e
Břasy....49·50n 13·35e
Bratislava....48·09n 17·07e
Břeclav....48·46n 16·53e
Březnice....49·33n 13·57e
Brezno....48·50n 19·39e
Brno....49·12n 16·37e
Broumov....50·35n 16·20e
Bruntál....49·59n 17·28e
Budišov nad Budišovkou....49·47n 17·38e
Bystřice....49·45n 14·41e
Bystřice pod Hostýnem....49·24n 17·40e
Bytča....49·14n 18·36e
Čadca....49·26n 18·48e
Čáslav....49·54n 15·23e
Čechtice....49·37n 15·03e
Čelákovice....50·10n 14·46e
Čerčany....49·51n 14·43e
Červený Kostelec....50·29n 16·06e
Česká Kamenice....50·47n 14·26e
Česká Lípa....50·42n 14·32e
Česká Třebová....49·54n 16·27e
České Budějovice....48·59n 14·28e
Český Brod....50·02n 14·58e
Český Krumlov....48·49n 14·19e
Český Těšín....49·45n 18·37e
Cheb....50·01n 12·25e
Choceň....50·00n 16·13e
Chomutov....50·28n 13·26e
Chotěboř....49·43n 15·40e
Chrudim....49·57n 15·48e
Čierna....48·25n 22·05e
Čierny Balog....48·45n 19·40e
Dačice....49·05n 15·26e
Děčín....50·48n 14·13e
Detva....48·31n 19·28e
Dobříš....49·47n 14·11e
Dobruška....50·17n 16·10e
Dobšiná....48·49n 20·23e
Doksy....50·35n 14·38e
Dolný Kubín....49·12n 19·17e
Domažlice....49·27n 12·56e
Drienov....48·53n 21·17e
Dubňany....48·55n 17·06e
Dubnica nad Váhom....48·58n 18·09e
Duchcov....50·37n 13·45e
Dunajská Streda....48·01n 17·35e
Dvůr Králové [nad Labem]....50·26n 15·48e
Fil'akovo....48·17n 19·51e
Františkovy Lázně....50·07n 12·21e
Frenštát pod Radhoštěm....49·33n 18·14e
Frýdek-Místek....49·41n 18·22e
Frýdlant....50·56n 15·05e
Galanta....48·12n 17·43e
Giraltovce....49·07n 21·31e
Gottwaldov....49·13n 17·41e
Handlová....48·44n 18·46e
Hanušovce nad Topl'ou....49·02n 21·30e
Hanušovice....50·05n 16·55e
Havířov....49·47n 18·27e
Havlíčkův Brod....49·36n 15·35e
Hlinsko....49·45n 15·55e
Hluboká nad Vltavou....49·03n 14·27e
Hlučín....49·54n 18·12e
Hodonín....48·51n 17·08e
Holešov....49·20n 17·35e
Holíč....48·49n 17·10e
Holýšov....49·36n 13·05e
Horažd'ovice....49·20n 13·43e
Hořice....50·22n 15·38e
Horní Počernice....50·06n 14·38e
Hořovice....49·50n 13·54e
Horšovský Týn....49·32n 12·56e
Hradec Králové....50·12n 15·50e
Hrádek nad Nisou....50·48n 14·51e
Hranice....49·33n 17·44e
Hriňová....48·36n 19·31e
Hronov....50·29n 16·12e
Hulín....49·19n 17·28e
Humenné....48·56n 21·55e
Humpolec....49·32n 15·22e
Ivančice....49·06n 16·23e
Jablonec nad Nisou....50·44n 15·10e
Jablonica....48·37n 17·25e
Jablunkov....49·35n 18·47e
Jaroměř....50·21n 15·55e
Javorník....50·23n 17·00e
Jelšava....48·39n 20·14e
Jemnice....49·01n 15·35e
Jesenice....50·04n 13·29e
Jeseník....50·14n 17·13e
Jevíčko....49·38n 16·43e
Jičín....50·26n 15·21e
Jihlava....49·24n 15·36e
Jindřichův Hradec....49·09n 15·00e
Jirkov....50·30n 13·27e
Kadaň....50·20n 13·15e
Karviná....49·50n 18·30e
Kaplice....48·45n 14·30e
Karlovy Vary....50·11n 12·52e
Kdyně....49·23n 13·02e
Kežmarok....49·08n 20·25e
Kladno....50·08n 14·05e
Klatovy....49·24n 13·18e
Kojetín....49·21n 17·18e
Kokava nad Rimavicou....48·34n 19·50e
Kolárovo....47·52n 18·02e
Kolín....50·01n 15·13e
Komárno....47·45n 18·09e
Konice....49·35n 16·53e
Košice....48·43n 21·15e
Kostelec nad Orlicí....50·08n 16·13e
Kralovice....49·59n 13·29e
Kralupy nad Vltavou....50·11n 14·18e
Kraslice....50·18n 12·31e
Kravaře....49·56n 18·01e
Kremnica....48·43n 18·54e
Křimice....49·46n 13·15e
Krnov....50·05n 17·41e
Kroměříž....49·18n 17·24e
Krompachy....48·56n 20·52e
Krupka....50·43n 13·46e
Kunovice....49·03n 17·29e
Kuřim....49·18n 16·32e
Kutná Hora....49·57n 15·16e
Kúty....48·40n 17·03e
Kyjov....49·01n 17·08e
Kynšperk nad Ohří....50·04n 12·32e
Kysucké Nové Mesto....49·19n 18·47e
Lanškroun....49·55n 16·37e
Lenora....48·56n 13·48e
Levice....48·13n 18·37e
Levoča....49·02n 20·36e
Liberec....50·46n 15·03e
Lipany....49·10n 20·58e
Lipník nad Bečvou....49·31n 17·35e
Liptovská Teplička....48·59n 20·06e
Liptovský Mikuláš....49·06n 19·37e
Lišov....49·01n 14·37e
Litoměřice....50·35n 14·09e
Litomyšl....49·52n 16·19e
Litovel....49·42n 17·05e
Litvínov....50·37n 13·36e
Loket....50·09n 12·43e
Lomnice nad Popelkou....50·32n 15·22e
Louny....50·19n 13·46e
Lovosice....50·31n 14·03e
Lučenec....48·20n 19·40e
Lysá pod Makytou....49·12n 18·13e
Makov....49·23n 18·30e
Malacky....48·27n 17·00e
Margecany....48·54n 21·01e
Mariánské Lázně....49·59n 12·43e
Martin....49·05n 18·55e
Medzilaborce....49·16n 21·55e
Mělník....50·20n 14·29e
Michalovce....48·45n 21·55e
Mikulov....48·49n 16·39e
Milevsko....49·27n 14·22e
Mimoň....50·40n 14·30e
Miroslav....48·57n 16·18e
Mladá Boleslav....50·23n 14·59e
Mníšek pod Brdy....49·52n 14·16e
Modra....48·21n 17·17e
Modrý Kameň....48·15n 19·20e
Mohelnice....49·46n 16·55e
Moravská Třebová....49·45n 16·40e
Moravské Budějovice....49·03n 15·49e
Moravský Krumlov....49·03n 16·19e
Most....50·32n 13·39e
Myjava....48·45n 17·34e
Náchod....50·25n 16·10e
Námestovo....49·25n 19·30e
Napajedla....49·10n 17·31e
Nejdek....50·17n 12·42e
Nepomuk....49·29n 13·36e
Netolice....49·03n 14·12e
Nitra....48·20n 18·05e
Nová Baňa....48·26n 18·39e
Nová Bystřice....49·01n 15·06e
Nováky....48·43n 18·34e
Nová Paka....50·29n 15·31e
Nové Hrady....48·47n 14·37e
Nové Město....50·21n 16·09e
Nové Město na Moravě....49·34n 16·04e
Nové Město nad Váhom....48·46n 17·49e
Nové Zámky....47·59n 18·11e
Nový Bohumín....49·56n 18·20e
Nový Bor....50·45n 14·33e
Nový Jičín....49·36n 18·00e
Nymburk....50·11n 15·03e
Nýřany....49·43n 13·12e
Nýrsko....49·18n 13·09e
Odry....49·39n 17·50e
Olomouc....49·36n 17·16e
Opava....49·56n 17·54e
Ostrava....49·50n 18·17e
Ostrov....50·17n 12·57e
Otrokovice....49·13n 17·31e
Pacov....49·28n 15·00e
Pardubice....50·02n 15·47e
Partizánske....48·39n 18·23e
Pelhřimov....49·26n 15·13e
Pezinok....48·18n 17·17e
Piešťany....48·36n 17·50e
Písek....49·19n 14·10e
Planá....49·52n 12·44e
Plasy....49·56n 13·24e
Plzeň....49·45n 13·23e
Poděbrady....50·08n 15·07e
Pohořelice....48·59n 16·32e
Polička....49·43n 16·16e
Polná nad Moravé....49·29n 15·43e
Poprad....49·03n 20·18e
Považská Bystrica....49·08n 18·27e
Prachatice....49·01n 14·00e
Prague....50·05n 14·25e
Přelouč....50·02n 15·34e
Přerov....49·27n 17·27e
Prešov....49·00n 21·15e
Přeštice....49·34n 13·20e
Příbor....49·39n 18·10e
Příbram....49·42n 14·01e
Prievidza....48·47n 18·37e
Prostějov....49·29n 17·07e
Púchov....49·08n 18·20e
Radnice....49·51n 13·37e
Rakovník....50·05n 13·43e
Rimavská Sobota....48·23n 20·02e
Rokycany....49·45n 13·36e
Rosice....49·11n 16·23e
Roudnice [nad Labem]....50·22n 14·16e
Rožňava....48·40n 20·32e
Rožnov pod Radhoštěm....49·28n 18·10e
Roztoky....50·09n 14·22e
Ružomberok....49·06n 19·18e
Rýmařov....49·56n 17·16e
Sabinov....49·06n 21·06e
Šafárikovo....48·27n 20·20e
Šahy....48·05n 18·57e
Šal'a....48·09n 17·52e
Sečovce....48·43n 21·40e
Sečovská Polianka....48·47n 21·42e
Sedlčany....49·40n 14·26e
Semily....50·36n 15·20e
Senec....48·14n 17·24e
Senica....48·41n 17·22e
Sered'....48·28n 17·44e
Sezimovo Ústí....49·23n 14·42e
Skalica....48·51n 17·14e
Slaný....50·11n 14·04e
Slapanice....49·10n 16·44e
Slavkov u Brna....49·09n 16·52e
Slavonice....49·00n 15·21e
Šluknov....51·00n 14·27e
Snina....48·59n 22·07e
Soběslav....49·15n 14·44e
Sobrance....48·45n 22·11e
Sokolov....50·09n 12·40e
Sol'....48·56n 21·36e
Spišská Nová Ves....48·57n 20·34e
Stará Boleslav....50·12n 14·42e
Stará Turá....48·47n 17·42e
Staré Město....49·06n 17·28e
Starý Plzenec....49·42n 13·28e
Šternberk....49·44n 17·18e
Strakonice....49·16n 13·55e
Strání....48·55n 17·42e
Strážnice....48·54n 17·18e
Strážske....48·53n 21·50e
Stříbro....49·46n 13·00e
Stropkov....49·12n 21·40e
Studénka....49·42n 18·05e
Štúrovo....47·48n 18·49e
Šumperk....49·58n 16·58e
Šurany....48·06n 18·14e
Sušice....49·14n 13·32e
Světlá nad Sázavou....49·40n 15·25e
Svidník....49·18n 21·35e
Svit....49·03n 20·12e
Svitávka....49·30n 16·37e
Svitavy....49·45n 16·27e
Tábor....49·25n 14·41e
Tachov....49·48n 12·38e
Tanvald....50·45n 15·19e
Telč....49·11n 15·27e
Teplice....50·39n 13·48e
Tišnov....49·21n 16·25e
Tisovec....48·43n 19·57e
Topol'čany....48·34n 18·10e
Toužim....50·04n 13·00e
Třebechovice pod Orebem....50·12n 16·00e
Třebíč....49·13n 15·53e
Trebišov....48·40n 21·47e
Třeboň....49·00n 14·47e
Trenčín....48·54n 18·04e
Třešť....49·13n 15·30e
Trhové Sviny....48·51n 14·39e
Třinec....49·41n 18·40e
Trnava....48·23n 17·35e
Trstená....49·22n 19·37e
Trutnov....50·34n 15·55e
Turnov....50·35n 15·10e
Turzovka....49·25n 18·39e
Týn nad Vltavou....49·14n 14·26e
Ubl'a....48·55n 22·23e
Uherské Hradiště....49·05n 17·28e
Uherský Brod....49·02n 17·39e
Ústí nad Labem....50·40n 14·02e
Ústí nad Orlicí....49·58n 16·24e
Valašské Meziříčí....49·28n 17·58e
Varnsdorf....50·52n 14·40e
Velká Bíteš....49·17n 16·13e
Vel'ké Kapušany....48·33n 22·04e
Velké Meziříčí....49·21n 16·00e
Veselí nad Lužnicí....49·11n 14·43e
Veselí nad Moravou....48·58n 17·22e
Vimperk....49·03n 13·47e
Vítkov....49·46n 17·45e
Vlašim....49·42n 14·54e
Vodňany....49·09n 14·11e
Volary....48·55n 13·54e
Volyně....49·10n 13·53e
Votice....49·38n 14·39e
Vráble....48·15n 18·19e
Vranov....48·54n 21·41e
Vrchlabí....50·38n 15·37e
Vrútky....49·07n 18·55e
Vsetín....49·21n 17·59e
Vyškov....49·16n 17·00e
Vyšná Radvaň....49·07n 21·56e
Vysoké Mýto....49·57n 16·10e
Vyšší Brod....48·37n 14·19e
Zábřeh....49·53n 16·52e
Žamberk....50·05n 16·28e
Západočeský....49·45n 13·00e
Žatec....50·18n 13·32e
Zbiroh....49·52n 13·47e
Zbraslav....49·59n 14·24e
Žďár nad Sázavou....49·34n 15·57e
Žiar nad Hronom....48·36n 18·52e
Žilina....49·14n 18·46e


Leipzig
Dresden
Wrocław
Katowice
Kraków
PRAGUE
Plzeň
Brno
Ostrava
Olomouc
Pardubice
Hradec Králové
Liberec
Ústí nad Labem
Most
Chomutov
Teplice
Děčín
Jablonec nad Nisou
Mladá Boleslav
Karlovy Vary
Cheb
Kladno
Příbram
České Budějovice
Jihlava
Prostějov
Přerov
Zlín
Gottwaldov
Žilina
Trenčín
Prievidza
Banská Bystrica
Zvolen
Nitra
Trnava
Bratislava
VIENNA
Košice
Prešov
Poprad
Martin
Ružomberok
Dubnica nad Váhom
Nové Zámky
Komárno
BUDAPEST
Znojmo
Havířov
Karviná
Frýdek-Místek
Nový Jičín
Opava
Sumperk
Šumperk
Jeseník
Hradec Králové
Pardubice
Kolín
Kutná Hora
Chrudim
Litomyšl
Ústí nad Orlicí
Česká Třebová
Lanškroun
Svitavy
Moravská Třebová
Šternberk
Olomouc
Prostějov
Přerov
Hranice
Valašské Meziříčí
Vsetín
Rožnov pod Radhoštěm
Frýdek-Místek
Třinec
Český Těšín
Jablunkov
Turzovka
Čadca
Kysucké Nové Mesto
Makov
Bytča
Lysá pod Makytou
Púchov
Považská Bystrica
Vrútky
Liptovský Mikuláš
Poprad
Spišská Nová Ves
Levoča
Kežmarok
Bardejov
Svidník
Humenné
Michalovce
Sobrance
Vranov
Strážske
Trebišov
Veľké Kapušany
Ciema
Rožňava
Jelšava
Hnúšťa
Kokava nad Rimavicou
Rimavská Sobota
Lučenec
Fiľakovo
Veľký Krtíš
Modrý Kameň
Šahy
Ipeľ
Levice
Vráble
Zlaté Moravce
Nitra
Topoľčany
Partizánske
Prievidza
Handlová
Nováky
Kremnica
Žiar nad Hronom
Detva
Banská Štiavnica
Štiavnické vrchy
Nová Baňa
Senec
Pezinok
Sereď
Galanta
Šaľa
Surany
Dunajská Streda
Kálnica
Kúty
Myjava
Senica
Skalica
Holíč
Hodonín
Břeclav
Mikulov
Strážnice
Veselí nad Moravou
Hustopeče
Uherské Hradiště
Napajedla
Otrokovice
Kroměříž
Kojetín
Holešov
Bystřice pod Hostýnem
Kyjov
Slavkov u Brna
Slavonice
Rosice
Telč
Třebíč
Jindřichův Hradec
Dačice
Moravské Budějovice
Moravský Krumlov
Pohořelice
Miroslav
Jemnice
Nová Bystřice
Slavonice
Kunovice
Staré Město
Uherské Hradiště
Vsetín
Valašské Meziříčí
Rožnov pod Radhoštěm
Nový Jičín
Odry
Příbor
Studénka
Bílovec
Fulnek
Hlučín
Bohumín
Orlová
Karviná
Havířov
Frýdek-Místek
Třinec
Jablunkov
Turzovka
Čadca
Kysucké Nové Mesto
Makov
Bytča
Lysá pod Makytou
Púchov
Považská Bystrica
Vrútky
Martin
Ružomberok
Dubnica nad Váhom
Trenčín
Prievidza
Banská Bystrica
Zvolen
Nitra
Trnava
Bratislava
Nové Zámky
Komárno
Šahy
Ipeľ
Modrý Kameň
Fiľakovo
Lučenec
Rimavská Sobota
Veľký Krtíš
Kokava nad Rimavicou
Hnúšťa
Jelšava
Rožňava
Košice
Trebišov
Veľké Kapušany
Ciema
Sobrance
Michalovce
Humenné
Vranov
Strážske
Prešov
Poprad
Svit
Mikuláš
Liptovský Mikuláš
Ružomberok
Martin
Žilina
Kežmarok
Bardejov
Svidník
Stropkov
Giraltovce
Lipany
Sabinov
Snina
Vyšná Radvaň
Humenné
Michalovce
Sobrance

MAP INDEX (continued)

Žirovnice.....49·15n 15·11e
Zlaté Moravce...48·25n 18·24e
Žlutice.....50·03n 13·10e
Znojmo.....48·52n 16·02e
Zruč nad Sázavou.....49·45n 15·07e
Zvolen.....48·35n 19·08e

Physical features and points of interest

Berounka, *river*...50·00n 14·24e
Beskid Mountains.....49·40n 20·00e
Biele Karpaty, *mountains*.....49·00n 18·00e
Bodrog, *river*.....48·25n 21·40e
Bohemian Forest (Šumava), *mountains*.....49·00n 13·30e
Carpathian Mountains.....49·30n 22·00e
Čechy, *historic region*.....49·50n 14·00e
Černá hora, *mountain*.....48·58n 13·48e
Českomoravská vrchovina, *mountains*.....49·20n 15·30e
Český les, *mountains*.....49·40n 12·40e
Danube, *river*....47·48n 18·50e
Dukla Pass.....49·25n 21·42e
Dyje, *river*.....48·37n 16·56e
Gerlachovský štít, *peak*.....49·12n 20·08e
Hron, *river*.....47·49n 18·46e
Hrubý Jeseník, *mountains*.....50·00n 17·20e
Ipel', *river*.....47·49n 18·52e
Javorie, *mountain*.....48·27n 19·18e
Javorová skála, *mountain*.....49·31n 14·30e
Jihlava, *river*.....48·55n 16·37e
Jizera, *river*.....50·10n 14·43e
Kleť *mountain*...48·52n 14·17e
Klínovec, *mountain*.....50·24n 12·58e
Krkonoše, *mountains*.....50·46n 15·35e
Krkonošský Národný Park...50·45n 15·35e
Krušné hory, *mountains*.....50·30n 13·15e
Labe, *river*.....50·50n 14·12e
Laborec, *river*....48·31n 21·54e
Lesný, *mountain*..50·02n 12·37e
Lužnice, *river*....49·15n 14·22e
Malý Dunaj, *river*.....48·08n 17·09e
Minčol, *mountain*.....49·15n 20·59e
Mislivna, *peak*....48·40n 14·44e
Morava, *historic region*.....49·20n 17·00e
Morava, *river*....48·10n 16·59e
Nisa, *river*.....51·02n 15·00e
Nitra, *river*.....47·46n 18·10e
Nízke Tatry, *mountains*.....48·54n 19·40e
Ohře, *river*.....50·32n 14·08e
Ondava, *river*....48·27n 21·48e
Oslava, *river*.....49·05n 16·22e
Ostrov, *physical region*.....47·55n 17·35e
Poprad, *river*.....49·25n 20·45e
Praha, *mountain* .49·40n 13·49e
Slaná, *river*.....48·20n 20·20e
Slanské Vrchy, *mountains*.....48·50n 21·30e
Slovenské Rudohorie, *mountains*.....48·45n 20·00e
Slovensko, *historic region*...48·50n 20·00e
Štiavnické Vrchy, *mountains*.....48·40n 18·45e
Sudety, *mountains*.....50·30n 16·00e
Sumava, see Bohemian Forest
Svitava, *river*.....49·09n 16·37e
Svratka, *river*.....49·11n 16·38e
Tatranský Národný Park...49·15n 20·00e
Topl'a, *river*.....48·45n 21·45e
Uh, *river*.....48·35n 22·10e
Váh, *river*.....47·55n 18·00e
Vel'ký Milič, *mountain*.....48·35n 21·28e
Vltava, *river*.....50·21n 14·30e
Vysoké Tatry, *mountains*.....49·12n 20·05e

Drainage. Czechoslovakia lies in the headwater area of the main European watershed, with the Elbe draining 39 percent of the country, the Oder and Vistula 7 percent, and the Danube 54 percent. There are many small rivers of little economic importance, and only five rivers have a mean annual discharge of more than 3,500 cubic feet (100 cubic metres) per second. Only the Danube, which barely touches Czechoslovak territory, stands comparison with the great rivers of the world, with a discharge of 78,000 cubic feet (2,200 cubic metres) per second at Bratislava. The greatest flows generally occur in spring, when the rivers are swollen by melting snow, and the lowest in summer. Weather variations can seriously affect water supply, as the ratio of minimum to maximum flow can approach 1:250. The country is rich in mineral springs, and good use is made of the groundwater reserves.

Danube River

Plants and animals. Czechoslovakia's vegetation may be divided into three major groups: Hercynian (central and western European), Pannonian (steppe), and Carpathian (eastern Europe and Siberian). Each of these groups is modified by altitudinal and other climatic influences. Although large areas of original forest have been cleared for cultivation, forests still cover more than a third of the country. The major forest types include the oak (*Quercus robur*), steppe oak (*Q. pubescentis*), and oak-grove assemblages of the lowlands and uplands adjoining the Danubian Plain; the beech forests of the Carpathians (*Fagus sylvatica*); and the spruce (*Picea abies*) forests of the Carpathians and adjoining areas of Bohemia and Moravia. In the very highest elevations can be found taiga and tundra assemblages characteristic of more northerly or more elevated regions elsewhere in Europe. Since cultivation is mostly restricted to areas of level terrain, the proportion of original forest cover is highest in mountainous areas.

The timberline runs at about 4,500 feet above sea level in the west, rising to about 5,000 feet in the east. At these higher elevations, particularly in the Tatras, the tree cover below the timberline consists of little more than dwarf pine (*Pinus mugo*). The alpine zone contains grass and bushes (*Vaccinium, Calluna, Salix retusa,* and others); above 7,500 feet, lichens predominate.

Czechoslovakia's wildlife is extensive and varied; the High Tatra National Park probably shelters the most exceptional collection of wild animals, including bear, wolf, lynx, wild cat (*Felis silvestris*), marmot, otter, marten, and mink. Most of these may be hunted, although the parks themselves are fully protected. Certain other animals, however, are protected everywhere; the chamois, for example, very near extinction for many years, now appears to be increasing in numbers. Birdlife is everywhere; game birds are common in forests and marshes, especially pheasant, partridge, wild goose, and duck. All may be hunted, but larger species such as eagle, vulture, osprey, stork, eagle owl, bustard, and capercaillie are much rarer and often protected.

Wildlife

Although the history of wildlife protection in Czechoslovakia is not long, and although tourism is one of the major purposes of these areas, much attention is devoted to the preservation of the natural heritage. Game reserves exist for the breeding of rare or endangered species such as the mouflon. Nature reserves have been created to preserve especially important landscapes: the Šumava Forest, Moravian Karst, and Jizerské hory (mountains and peat bogs) are among the most significant.

The two largest and most important areas, however, are the High Tatra and Krkonoše (Giant Mountains) national parks. The former, established in 1948 and occupying 190 square miles (500 square kilometres), is an international area, being contiguous with, and jointly administered with, the Polish Tatrzański Park Narodowy. It preserves most of the species of wildlife mentioned above, together with a glacial landscape, alpine flora and fauna, and relict species from the Pleistocene glaciations. Krkonoše National Park, established in 1963 and occupying some 147 square miles (380 square kilometres), protects glacial and alpine landscapes and vegetation and some relict boreo-Arctic species, such as the alpine shrew (*Sorex alpinus*); it is most extensively developed, however, as a ski resort.

Traditional regions and the human imprint. Industrialization and urbanization have wiped out many of the former traditional regions of Czechoslovakia, although, reflecting differing national and cultural heritages, Bohemia, Slovakia, and to a lesser extent Moravia are still recognizable entities. Local traditions are preserved in the mountainous parts of Slovakia in the wearing of folk costumes on holidays and in the preparation of local dishes, the making of cheese, and the preserving of fruits and vegetables. Similar traditions have been preserved in southern Bohemia and in southeastern Moravia.

The country has a high density of settlement, with communities lying, on average, only a few miles apart. The total number of separate settlements exceeds 20,000, but there are fewer than 11,000 communes. Rural settlements are characteristically compact, but in the mountainous regions, colonized during the 13th and 14th centuries, straggling row villages are common. Dispersed rural settlements tend to be exceptional, occurring mostly on the Moravian–Slovakian boundary and in central Slovakia and reflecting the later colonization of the 17th and 18th centuries. Rural settlements with fewer than 500 inhabitants tend to prevail except in south Moravia and southwest Slovakia. The collectivization of farmland that took place in the decades following World War II has resulted in a pattern of large, regularly shaped fields, replacing the centuries-old division of land into small, irregular, privately owned plots.

Villages

Urbanization in Czechoslovakia is not particularly high for an industrialized country, only 47 percent of the population residing in communes with more than 5,000 inhabitants (1970) and 15 percent living in communes with 2,000–5,000. Even the smallest urban centres, however, usually contain some manufacturing industry. There are six cities with more than 100,000 inhabitants (accounting

Czechoslovakia, Area and Population

	area		population	
	sq mi	sq km	1970 census	1976 est.
Republics (*republiký*)				
Czech Socialist Republic				
Capital city (*hlavní město*)				
Prague	191	496	1,078,000	1,170,000*
Regions (*kraje*)				
Jihočeský (South Bohemia)	4,381	11,348	653,000	670,000
Jihomoravský (South Moravia)	5,802	15,027	1,938,000	1,992,000
Severočeský (North Bohemia)	3,015	7,810	1,103,000	1,140,000
Severomoravský (North Moravia)	4,273	11,067	1,800,000	1,883,000
Středočeský (Central Bohemia)	4,248	11,003	1,192,000	1,136,000†
Východočeský (East Bohemia)	4,340	11,240	1,202,000	1,227,000
Západočeský (West Bohemia)	4,198	10,872	849,000	875,000
Slovak Socialist Republic				
Capital city				
Bratislava	142	368	284,000	341,000*
Regions				
Středoslovenský (Central Slovakia)	6,941	17,976	1,403,000	1,462,000
Východoslovenský (East Slovakia)	6,247	16,179	1,256,000	1,324,000
Západoslovenský (West Slovakia)	5,595	14,491	1,599,000	1,636,000†
Total Czechoslovakia	49,374‡	127,877	14,358,000‡	14,857,000‡

*Change partly attributable to increase in area.
†Change partly attributable to decrease in area.
‡Figures do not add to totals because of rounding.
Source: Official government figures.

for 16 percent of the national population, and nearly half of this is attributable to Prague), but there are 64 in the 20,000–100,000 range (15 percent). Urbanization is greater in Bohemia than in Slovakia, although postwar industrialization in the latter has reduced the disparity. Prague, with a population of 1,170,000 (1976), is the national capital, historically occupying a predominant role. Bratislava, capital of Slovakia, has a population of 341,000, and Brno, chief city of Moravia, 360,000. The other large cities are Ostrava (301,000), the leading Czech coal–mining and steel centre; Košice (174,000), with its large steel complex; and Plzeň (156,000), with old established engineering and brewing industries.

THE PEOPLE

Composition. Archaeological findings confirm that what is now Czechoslovakia contains regions with some of the oldest settlements in Europe; in particular, a continuous agricultural settlement of the forest–steppe areas, up to 900 feet above sea level, can be dated back to the 4th millennium BC. An ensuing series of invading waves of population was brought to an end by the West Slavs, who settled the whole territory—with the exception of the mountainous regions—in the 5th and 6th centuries AD. The later German and Hungarian minorities were the result not only of immigration but also of some assimilation of the native population.

The Czechs and the Slovaks

Differing economic and political developments resulted in a differentiation of the originally uniform Slav population into Czech and Slovak components, the latter people forming one of the youngest such entities in Europe. The very similar Czech and Slovak languages belong to that group of Slavic languages that uses the Roman rather than the Cyrillic alphabet. The literary Czech language dates from the beginning of the 15th century, Slovak from the 19th.

Anthropologically, the population of Czechoslovakia is very mixed, about half the people exhibiting the physical traits of the Alpine local race and thus reminiscent of western Europe, while east European Baltic elements predominate farther east, and Dinaric elements manifest themselves in Slovakia. Cultural exchanges resulting from dependence on the Austro-Hungarian monarchy until 1918 resulted in an influx of German and west European influences, whereas the linguistic affinity, together with political developments since World War II, has cemented ties with eastern Europe.

Czechs and Slovaks form the great majority (more than 94 percent) of the population, in contrast to the situation before World War II when minorities made up 39 percent. The postwar expulsion of Germans, an exchange of populations with Hungary, and the cession of Ruthenia (with a Ukrainian population) to the Soviet Union brought about the present homogeneous state. By the mid-1970s Czechs numbered 9,500,000 and Slovaks 4,400,000. There were about 580,000 Hungarians in southern Slovakia, small groups of Ukrainians near the Soviet frontier, and a few Poles in northeastern Moravia. Fewer than 80,000 Germans remain, living in northwestern Bohemia. Gypsies—who have no legal status as an ethnic group—are estimated to number between 200,000 and 300,000; according to some sources, they constitute the largest Gypsy population in any one country.

Religion. No official statistics on religion are kept, though the activities of the churches have been financed by the government since the nationalization of all church property in 1949. Roman Catholics are preponderant, but there is an Evangelical element in Slovakia, and a significant number of people have no formal religion. Atheism is the official government philosophy, and the churches' role is largely restricted to religious rites. Clergy are civil servants and required to take an oath of loyalty to the government.

Demographic trends. Although Slovakia shows a natural increase that is higher than that in the ČSR, the country as a whole has a moderate to low rate of natural increase. The annual rate decreased from nearly nine per thousand in 1955–59 to only five in 1965–69, essentially because of a falling birth rate. The rate of natural increase had recovered significantly by the mid-1970s, however, to about eight per thousand, owing to a rising birth rate (19.5 per thousand in 1975), which is attributable in large part to pro-natalist government policy. The mortality rate, about 11.5 per thousand, conceals an important difference between the ČSR (12.4) and Slovakia (9.5), this being largely attributable to the more youthful age structure of the latter. Infant mortality, at about 21 per thousand, is comparable to that of the more advanced countries.

A serious demographic problem is created by the fact that there are few families with more than two or three children, and the number of children per family is decreasing. Although the marriage rate is rather high, the divorce rate is climbing; by the 1970s (as in other European countries) it exceeded 20 percent of new marriages. This factor, together with the changes in way of life associated with urbanization and employment of women in the labour force, helps to explain the rather low rate of population increase.

Population movement

Internal movements of population are of increasing significance with the further modernization of the economy. Migration is directed from the country to the towns and from the smaller towns to the larger cities. In fact, there is a net decline in the number of local inhabitants in most Czechoslovakian communes. Although the lower natural growth rate in the west has been supplemented by immigration from Slovakia, this movement has declined with the growth of urbanization in Slovakia itself.

In the past, population growth was slowed by emigration to the urban centres of the Austro-Hungarian Empire and farther afield, especially to the United States. The main emigrant regions were eastern and central Slovakia, and by the 1970s it was estimated that there were about 1,000,000 Czech and 1,500,000 Slovak emigrants and their descendants living in the United States, although only fragile ties were retained with the homeland after the second generation. Emigration now is a mere trickle.

The most complicated demographic problems in contemporary Czechoslovakia stem from the fact that, with only moderate birth rates and a prolongation of average life-span (approaching 67 for men, 74 for women), the population is aging. In 1950 less than 12 percent of the people were over 60; in 1975 the figure exceeded 17

percent. The highest percentage of old people occurs in central and eastern Bohemia; the Slovak population is generally younger because of the higher birth rate.

Female employment

Half of the total population is in the labour force, a total of 7,435,000 people. The female employment rate is among the highest in the world, almost half of all the employed people being women. Industry and construction take up about 48 percent of the total labour force, with agriculture, transport and trade, and services taking almost 52 percent of the remainder. There are considerable regional variations to this pattern, with industry predominating in urban areas and the number of persons engaged in agriculture increasing eastward, although only a third of the rural population works, on the average, in agriculture. Independent small producers and cooperative farmers represent only 2 percent and 11 percent, respectively, of the labour force.

Geographic conditions make possible the settlement of almost the whole of Czechoslovakia, but the population tends to be concentrated at lower altitudes, more than half the population living below about 1,300 feet above sea level. There is an abrupt decline in population density with increasing altitude. A notable feature is the low density in some frontier areas, partly reflecting the induced emigration of minorities in the post-World War II years; the ČSR was one of the few European political entities to have fewer people (9,800,000) in 1970 than in 1910 (10,080,000), a circumstance explained by war losses and the expulsion of 2,500,000 Germans.

The demographic outlook for Czechoslovakia is not dynamic; the total population will grow only slowly, and the aging of the population will continue unless the recent revival of the birth rate can be maintained. The situation is most serious in agricultural areas and the interior of the ČSR, and there are prospects that the Slovak share of the national population will increase, shifting the demographic centre of gravity eastward. (M.Bl./R.H.O.)

THE NATIONAL ECONOMY

Although Czechoslovakia is not rich in natural resources, its economy is one of the most advanced in eastern Europe. It uses more chemical fertilizers and machinery per acre and boasts higher agricultural yields than most other east European countries. Its industry is highly developed, providing a substantial share of total east European production. It is a very important supplier of a wide range of machinery for fellow Comecon members and is eastern Europe's largest manufacturer of a number of engineering products. National income has grown relatively slowly since World War II, if compared with the performance of other European countries in the Communist bloc, but in the 1970s it was estimated that, after the German Democratic Republic, Czechoslovakia was the second most prosperous country in the entire Communist world. It is also eastern Europe's largest foreign trader in per capita terms.

Natural resources. *Mineral resources.* Black and brown coals are produced in significant quantities, output in relation to population being high by world standards, but deposits of petroleum and natural gas are small.

Coals

Most of the black coal is derived from the Ostrava–Karviná coalfield, though it is also mined near Kladno, in the Plzeň basin, near Trutnov, and near Brno. A high proportion of the black coal is of coking quality. Production of brown coal (including small quantities of lignite) increased rapidly up to the mid-1960s but remained static thereafter; open-pit methods are used. The main mining areas are around Chomutov, Most, Teplice, and Sokolov. The brown coal is used extensively for thermal power stations and for domestic fuel, and large quantities are also utilized as raw material in the chemical industry. New investment should lead to increased brown coal output by the late 1970s. Petroleum and natural gas are produced near Hodonín, but only in small quantities. There is a great reliance on pipelines importing Soviet oil and natural gas, the latter supplementing existing coal gas supplies.

Czechoslovakia has a limited endowment of metallic ores. Iron ore is mined chiefly in the Slovak Ore Mountains (Slovenské Rudohorie), while areas between Prague and Plzeň are historically important. Because of the gradual exhaustion of reserves, annual production fell after 1960 and remained static through the late 1960s and mid-1970s at about 1,700,000 tons. There is a great reliance on imported iron ore, especially from the Ukrainian S.S.R. Copper and manganese ores are also mined in the Slovak Ore Mountains. Lead and zinc ores are mined near Kutná Hora, Příbram, and Banská Štiavnica and in the Jeseník Mountains. Uranium is also mined near Příbram. The Ore Mountains (Krušné hory) of Bohemia yield small quantities of tin. Imported bauxite and nickel ore are refined at Žiar nad Hronom and Sered', respectively. Other mineral resources include salt in east Slovakia, graphite near České Budějovice, and kaolin near Plzeň and Karlovy Vary.

Biological and hydroelectric resources. Just under 55 percent of the country's total land area consists of agricultural land, and the bulk of this is arable land, with few meadows and pastures. On the lowlands the soil is generally fertile, but in the mountainous regions it is considerably less productive. More than one-third of the country's surface is wooded (see also above *Soils; Plants and animals*). Hydroelectricity plays a small part in the country's total energy supplies. Power stations are located on the Váh and Vltava rivers, but their output represents less than a tenth of total electricity production.

Sources of national income. *Agriculture.* At the beginning of the 1960s, agriculture accounted for 16 percent of total net material product, but by the end of the decade its share had dropped by a quarter. Growth in this sector —which employed only about 14 percent of the economically active population by the mid-1970s—has been adversely affected by the government planners' emphasis on industry and, in the opinion of Western observers, by the disincentives associated with collectivization. Commentators also avow that another weakness of the authorities' farm policy has been the low prices paid to farmers and that, although these have been revised upward, there is still a considerable difference between rural and urban incomes.

Declining share of agriculture

Czechoslovak agriculture is, nevertheless, unquestionably one of the most advanced in the Communist bloc, and yields tend to be better than the average for eastern Europe as a whole. The program of rural collectivization was completed by 1961, although much consolidation of cooperative farms took place afterward (between 1971 and 1976 their total number fell from 6,200 to 2,200, while land held by them rose by 7 percent). Production is organized on the basis of these 2,200 cooperatives and a small number of state farms. Cooperatives in 1975 owned 63 percent of the total agricultural land (including 2.4 percent that is allotted to personal use of individual members) and accounted for 61 percent of total farm output (in addition, private production, 5 percent). State farms worked 30 percent of the land and provided 27 percent of food production. For private farms and plots, the percentages were 6.7 and 7, respectively.

On the fertile lowlands, wheat, barley, sugar beets, maize (corn), and fodder crops are the most important, but on the relatively poor soils of the mountains the principal crops are rye, oats, and potatoes. Cereals lead in total production, with wheat, barley, oats, and rye, in that order, as the most important crops. Other important agricultural items include potatoes. The country also raises a sizable amount of livestock.

Industry. By the mid-1970s industry accounted for nearly two-thirds of the country's net material product and employed nearly 40 percent of the economically active population. The period from 1955 to 1970 was characterized by a very rapid growth in industry, the index of industrial production advancing more than 160 percent. Growth tended to be fastest in heavy industry, but it fell behind—in terms of both quality and quantity —in light industry and in consumer goods. Engineering is the largest branch of industry, constituting more than a quarter of total industrial production. In close second place is the food industry, followed by ferrous metallurgy (about 10 percent of the total), fuel mining and process-

ing, and the chemical, rubber, and asbestos industries. Czechoslovakia's iron and steel industries are the second largest in eastern Europe, after the Soviet Union, based largely on imported ores. Steel production is centred on the plants of the Ostrava area, with lesser amounts produced at Kladno, Plzeň, and Chomutov (all in Bohemia); the major centre in Slovakia is Košice.

Czechoslovakia is eastern Europe's largest producer of a number of engineering products, including electric and diesel locomotives and passenger cars.

Energy. Czechoslovakia is dependent on solid fuels for 70 percent of its energy supplies, which are not adequate to meet domestic requirements. The importance of liquid fuels and natural gas is expected to grow, while it is estimated that the share of hydroelectricity will remain unchanged at 1.5 percent. Electricity output has risen substantially, but shortages still pose a serious problem. The bulk of this output is derived from brown coal. There is a nuclear power station near Trnava. Oil and natural gas imports from the Soviet Union are necessary.

Banks

The financial sector and foreign trade. Banking is nationalized, with the State Bank of Czechoslovakia (Státní banka československá) as the most important of the country's financial institutions; it is the sole bank of issue, provides credit to enterprises, manages the foreign-exchange reserves, and administers the government's overall financial policy. Some loans are also provided by the Investment Bank, while the Commercial Bank of Czechoslovakia (Československá obchodní banka) specializes in trade and related transactions. There are also the Czech and Slovak State Savings Banks (Česká i Slovenská státní spořitelna) and the Trade Bank (Živnostenská banka), which provides banking service for foreigners.

The country is highly dependent on foreign trade; in per capita terms it is the largest foreign trader in the Communist bloc. The foreign-trade account has generally been in surplus, although trade with the advanced capitalist world usually yields a deficit. The rising prices of raw material imports had become an adverse factor by the mid-1970s. The lion's share of foreign trade—about 70 percent—is accounted for by Communist countries, including the Soviet Union (about one-third), followed by the German Democratic Republic and Poland. Outside the Communist bloc, its largest trading partner is the Federal Republic of Germany, which accounts for more than 6 percent of total imports and exports.

About 47 percent of all of the imports are raw materials, of which fuels, minerals, and metals are the most important. The second largest group consists of machinery and tools, followed by food and consumer goods. On the export side, about one-half is made up of sales of machinery—including motorcars and machine tools, electric-power machinery, switchgear, and textile and leather machinery—much of which is shipped to fellow Comecon countries. This is followed by sales of raw materials (*i.e.*, wood, metalliferous ores and scrap, and some chemicals), which accounted for about 30 percent of the total, and consumer goods (about 18 percent).

The role of the state in the economy

The management of the economy. The state plays an extremely important role in the economy. Approximately 90 percent of all of the agricultural land is in the hands of the state farms and the cooperatives, and industry, banks, and most trading and service enterprises are owned and closely controlled by the government. Private enterprise is tolerated on a very small scale only and mainly in certain specified areas such as handicrafts, some service trades, and some farming activities. All in all, the private sector accounts for less than 5 percent of the net material product and employs no more than 4 percent of the country's labour force. Economic planning is based on detailed five-year plans, which lay down a large number of targets in all the areas of the economy but, as elsewhere in the Communist bloc, are less centralized than they were in earlier years.

Taxation. As the state derives substantial revenue from the profits of the state enterprises, taxes play a relatively small role in the budget—less than 15 percent of total budgeted revenue. The principal tax is the turnover tax on consumer goods that is levied on producer, wholesaler, retailer, and ultimate buyer and that accounts for 63 percent of all tax receipts. This is followed by the payroll tax on enterprises, with a 28 percent share; the remainder is made up of a large number of different levies, the yield of which is insignificant.

Trade unions. Most workers belong to the trade-union movement, which is organized by industry and, as in other Communist countries, plays a central role in implementing national economic policy. The major function of the unions is to mobilize the working force to support the regime.

Economic policies, problems, and prospects. Up to 1960, economic progress was fairly rapid, but the subsequent years brought a slowdown in growth, and the authorities drew up an extensive program for reforming the system of economic management. In broad terms, the avowed aim was to reduce reliance on rigid central regulation of the economy—which had proved to be unable to match supply and demand or to ensure maximum efficiency—and to replace it with greater emphasis on a balancing of economic forces. A program embodying these goals came into effect from 1966; its principal features were a revision of the pricing system, which freed some prices from central control, and the granting of greater independence to individual enterprises.

By 1968, however, it became apparent to both Czechoslovak economists and Western commentators that the reform program had run into serious trouble. It led to a much faster rise in prices than anticipated, and it expanded consumer and investment demand to levels that could not be satisfied. After the invasion of the country by Warsaw Pact forces in August 1968, the newly installed government introduced firmer control over prices, investment, and the whole economy; according to the new plans, trade with the Soviet Union would grow faster than that with the West. As a result of these and other changes, the system of economic management is, in the opinion of Western observers, somewhat more rigid than that of some other east European countries, although it is generally accepted that it is still more flexible than it was before the ill-fated reform program.

Continuing problems of the economy

It is also generally accepted that the Czechoslovak economy, like the economies of many other countries, both East and West, is faced with a number of difficulties in the 1970s. The reforms outlined by the Dubček regime of January–August 1968, along with earlier defects in the economic system, have been decisively rejected. Yet a new, comprehensive strategy was, not unexpectedly, slow to evolve. The difficulties of the 1960s left a heritage of lack of purpose and confidence, a problem made even more serious by widespread industrial apathy and poor labour discipline. Among the more pressing specific problems acknowledged by both Czechoslovak economists and outside observers in the 1970s were the low level of productivity, the lack of advanced equipment in many industries, the shortage and poor quality of consumer goods, and the neglected state of the country's infrastructure.

In spite of such difficulties, the Five-Year Plan for 1971–75 was a notable success, national income rising by about one-third. The plan for 1976–80 envisaged further economic expansion, aided by selective industrial re-equipment, a continued export drive, and greater collaboration within the Comecon bloc. (E.I.U./R.H.O.)

TRANSPORTATION

Czechoslovakia has a typical inland transport system, the modernization of which has taken place only in recent decades. Transit functions, whereby the country was merely a link in trans-European trade, lost much of their importance in the new political and economic conditions that prevailed on the Continent after 1945.

The railways. The most important element of the transport system is the railways, which are especially important in freight transport, notably of coal (about 40 percent of the total), ores, metals, and building materials. Density of the railway network is fairly high at 26.8 miles per 100 square miles (10.3 kilometres per 100 square kilometres) but is not always effective. The rail-

ways developed in the second half of the 19th century and mostly ran from north to south, to Vienna and to Budapest, whereas the main trend of present-day transportation runs east–west. Communications with the main trading partner of Czechoslovakia, the Soviet Union, have been facilitated by an extensive program of reconstruction and electrification; more than 20 percent of all lines had been electrified by 1975. Most freight moves along main-line routes, but short journeys in the vicinity of the larger towns are of decisive importance in passenger traffic.

Road, water, and air transport. The density of roads and the number of scheduled bus routes are also high, and the transport of passengers and freight by state-owned vehicles predominates. Again, the urban centres are the focus of the most intensive road networks. Although Czechoslovak roads are of good quality, there are few of the throughways and expressways familiar in the West, and the Western visitor is surprised by the small number of private cars seen on the Prague–Brno road. Only in Prague and its surroundings and on weekends in other large cities does traffic flow approach the west European level; the emphasis on cheap mass transportation is an attempt to forestall in Czechoslovakia the traffic paralysis that plagues many countries.

Waterways

Czechoslovakia's position in the headwater area of the main European watershed means that the potential of water transport is small; of the rivers flowing through the country, including the Danube, only small portions actually in the country are navigable. Freight is transported on the Elbe to Hamburg, which functions as a port for mixed cargo, but Danube river transport, with Komárno the largest Czechoslovak port, now predominates, carrying raw materials imported from the Soviet Union and its neighbours. Although natural conditions would facilitate the construction of canals linking the Danube, Oder, and Elbe, such projects would be very expensive, and their construction—which would make Czechoslovakia the centre of European waterway routes—would also require international cooperation.

Air connections between Prague, Bratislava, and other regional centres are of considerable importance, and Prague itself is a major international air centre. Finally, major changes in freight transportation have been introduced by the Druzhba (Friendship) oil pipeline running into the country from the Volga River region of the Soviet Union, and also by a gas line from Siberia, which runs beyond Czechoslovakia to western Europe.

ADMINISTRATION AND SOCIAL CONDITIONS

The constitutional framework. According to the constitution of 1960, as modified by the Constitutional Law of 1968 and by subsequent amendments in 1970 and 1975, Czechoslovakia is a Socialist federative republic, with all power derived from the people. The head of the federation is the president, elected by the two-chamber Federal Assembly, which consists of the Chamber of the People (Sněmovna lidu) and the Chamber of the Nations (Sněmovna ńarodù). The latter consists of equal delegations from the national councils; *i.e.*, the legislative bodies of the Czech and Slovak Socialist republics.

The government of the federation (*i.e.*, the president, with the premier, deputy premiers, and ministers), appointed by the president, is theoretically responsible to the Federal Assembly. In practice, however, the power of the government is greater than that of the assembly, and the power of the Communist Party is greater than that of either. Individuals often occupy seats of rank in both government and party.

The Communist Party

Political life is organized by the National Front, a union of political parties and mass organizations, including the trade unions and women's and youth groups. The Communist Party of Czechoslovakia (Komunistická Strana Československa, or KSČ), with more than a million members, is the dominant element in the National Front, and its leading position is embodied in the constitution. It thus has an effective monopoly of political power. Though there are several other political parties, they function as auxiliaries to the KSČ rather than as an opposition. Organizations that are not members of the National Front are not permitted to function.

The president of the Central Committee of the National Front is also the general secretary of the Communist Party, and the front controls elections at all levels, presenting single lists of candidates, one candidate for each seat. All of the component organizations of the front have constituent branches representing the local government units, as well as individual plants and economic enterprises.

Local government. Local administration is controlled by "national" committees that operate on three levels. Czechoslovakia is divided into seven Czech and three Slovak *kraje* (regions), the capitals, Prague and Bratislava, each occupying independent positions; the regions are subdivided into *okresy* (districts); and these, in turn, are subdivided into numerous *obce* (local communities, or communes). The last named came into being in the 18th century as basic groups of settlements and are being merged. The administrative division underwent changes in 1948, 1960, and 1968.

Justice. In the civil division, the lawcourts are most concerned with family affairs, including divorce cases; the criminal courts deal with crime and traffic accidents. Conflicts in the sphere of economic law between the various components of the Communist system are dealt with by independent, government-organized arbitration.

The armed forces. The four military forces are integrated as the National People's Army, including ground forces, air and air defense forces, frontier guards, and interior guards (the two last under the Ministry of the Interior till 1965). Their numbers were estimated by Western experts at 180,000 in 1976. Two years of service in the military forces are required from men over age 18 who are physically and otherwise fit and who are called. Discharged conscripts usually remain in the reserves till age 50.

Services. The management of houses and apartments is one of the more critical areas of social policy in Czechoslovakia. Local "national" committees control the housing policy, including the 60 percent of apartments that are in private homes as well as the state-owned urban housing projects. Cooperative housing projects are increasingly popular, and private ownership remains important. The housing problem continues to be one of the most severe afflicting the country; by the early 1970s, the average age of housing units was about 60 years, and 10 percent or more of the apartments were more than a century old, although almost a quarter had been built since 1945. About eight new apartments per 1,000 inhabitants are built each year. Almost all housing units are now supplied with electricity, two-thirds with water, and about half with bathrooms and gas connections. There is little regional variation in this regard.

The housing problem

A centrally organized police force, the National Security Corps, under the Ministry of the Interior, takes care of public safety and is supplemented by the voluntary organization of the People's Militia in factories and similar enterprises. The police also combat unauthorized political activity.

Consumer goods and services are made available through national and cooperative retail stores; more than half of the 70,000 or so shops are food outlets. Large department stores and shopping centres are found in the cities and in the new housing developments, and remote settlements are served by travelling stores. Specialized stores and service centres tend to be concentrated in larger urban areas.

Education. The system of schools in Czechoslovakia has been built up in the tradition of the great 17th-century European educational reformer John Amos Comenius, who was born in Moravia. Compulsory education, which has existed since the middle of the 18th century, lasts nine years. Preschool education is an important consideration with the children of employed mothers, and there are some day nurseries in operation. Study in secondary schools lasts three or four years. The secondary schools are so located as to be accessible from all the districts. Specialized schools include technical, agricultural,

and commercial schools and schools of social hygiene (which hold nursing classes). The larger towns, as well as traditional cultural centres, are the areas served best.

Colleges and universities

Eight undergraduates per 1,000 inhabitants attend the nearly 40 universities and colleges. In the mid-1970s, most students were studying teaching, medicine, engineering, social sciences, and agriculture. The leading educational institutions, providing four to five years of intensive study, have very ancient traditions. Charles University (Universita Karlova, founded 1348) and the Czech Technical University in Prague (České vysoké učení technickév Praze, founded 1707) are the oldest in central Europe. After World War II, however, higher educational institutions increased in number, especially in Slovakia, where by 1975 about two dozen universities and colleges existed, in comparison with one university before the war. The largest student populations are at Prague (about 51,000) and Bratislava (36,000).

In small nations, instruction in languages is very important. Language teaching starts in Czechoslovakia with Russian in the third year of school, followed later by another major language, usually German or English.

The emphasis on adult education is not so strong as it was in the early years of the Communist regime, but the opportunities are still good.

Research work in Czechoslovakia is organized not only in the universities but also in special research institutions. Basic research is carried out by the Czechoslovak Academy of Sciences, which has institutes in Prague, Brno, and Bratislava.

Health, recreation, and tourism. Czechoslovakia is traditionally among the very best equipped countries medically, although Slovakia is somewhat less well served than the Czech Lands. One-half of all of the doctors work in specialized medical establishments and in district and regional hospitals. Czechoslovakia has more than 50 sanatoriums, most of them in the Czech Lands, where more than 360,000 patients, including children and pensioners, as well as more than 19,000 visitors from abroad, are treated each year. For most patients, treatment is free. Half of the spas are in northwestern Bohemia, in such internationally known resorts as Karlovy Vary, Mariánské Lázně, and Františkovy Lázně.

Football (soccer), ice hockey, track-and-field sports, and skiing are among the most popular sports, and playgrounds and gymnasiums can be found in all of the larger communities. The larger towns have winter stadiums and sports grounds. Camping is also popular. The rest centres run by the trade unions are visited by 400,000 persons annually and their summer camps for children by more than 250,000. The main tourist regions are the Krkonoše, the Tatras, and other scenic mountain regions, although many citizens continue to be attracted by the beauty of Prague and other ancient centres.

Half of the families of Prague, Ostrava, and Bratislava have a cottage or cabin in the country near the city. Excursions into the country and the picking of wild berries and mushrooms are favourite forms of recreation, as is gardening: not surprising, as much of the urban population of Czechoslovakia originally came from the country.

Foreign tourists numbered some 8,500,000 annually in the mid-1970s, not a great number considering the attractions of the country. There were about 13,000,000 foreign visitors every year (including those coming for reasons other than holiday), two-thirds of them from other Communist countries, notably the German Democratic Republic, Hungary, and Poland. A considerable number of tourists also come from the Federal Republic of Germany. Czechoslovak citizens themselves mostly visit Hungary, Yugoslavia, the German Democratic Republic, Poland, and Bulgaria.

Living standards. Czechoslovakia has a generally good standard of living as compared to the other Communist-bloc countries, about 52 percent of the real national income going to personal consumption. The mean monthly wage shows some variation, industrial workers as a whole getting above-average wages, though lower wages can be found, for example, in the clothing industry and in some services; coal miners receive maximum wages. Farmers' incomes are about average. Unearned income is forbidden. All workers participate in pension insurance, and the retirement age is 60 years for men and 55 for women. Old-age pensions, however, remain low, approaching only 50 percent of mean wages. Regional variations in income may be up to 40 percent above or below the average, as, for example, between mining districts and cities, on the one hand, and predominantly agricultural districts, with families with many children, on the other; some peasants in remote areas of Slovakia live at subsistence level.

Wages and pensions

In the average Czechoslovak household of the 1970s, the greater part of the family income is provided by the husband and his working wife. Sick benefits, old-age and disability pensions, and child allowances represent most of the remainder. In farmers' families, the income structure is different, about 15 percent of the income coming from crops they grow and sell privately. In addition, medical care, medicines, and education are free, and there are reduced fares for transportation to and from work. Vacations range between two and four weeks per year, and there is generous paid maternity leave.

About 30 percent of average family expenditure goes to relatively costly food and drink; about 30 percent to manufactured goods; and 30 percent to services, rent, and taxes. As much as 10 percent of the income is reserved for savings. The national beer consumption per capita is the highest in the world. The high rate of calorie intake is largely accounted for by consumption of bread and wheat products, including the ubiquitous national dish, dumplings, of which no fewer than 345 pounds (156 kilograms) per person are eaten each year.

CULTURAL LIFE AND INSTITUTIONS

The setting. The rich cultural milieu in contemporary Czechoslovakia is rooted in history and the geographic position of the Czechs and Slovaks in Europe. The earliest preserved writings in Czech are hymns dating from the 13th century. The most varied artistic schools have penetrated the country, and Prague and other historic cultural centres exhibit a mixture of architectural styles. The threat of doom to the Czechs and Slovaks as peoples in the 18th and 19th centuries gave national cultural life a strong patriotic element.

Literature and the theatre. Literature has occupied an important position in Czechoslovak cultural life since the late 18th century, when the development of an urban middle class created a public for it. Romanticism in the early 19th century was best exemplified by Karel Hynek Mácha, whom some regard as the greatest Czech poet of all time. Janko Kral' was a major Slovak poet of this period. The reaction to Romanticism came in the 1840s with the Realist writings of the Czech novelist Božena Němcová and the political journalist Karel Havlíček. National traditions were reasserted by the Czech poets Jan Neruda and Svatopluk Čech and by the Slovak poet Pavel Országh (whose pen name was Hviezdoslav). After independence in 1918 came the plays and novels of Karel Čapek, well known in the West for *R.U.R.* (*Rossum's Universal Robots*), and the plays of František Langer. Jaroslav Hašek's novel *The Good Soldier Schweik* is still highly popular. After the establishment of the Communist regime in 1948, Socialist Realism became dominant stylistically, and interest in literature declined. The period of the "thaw" brought a number of writings critical of the events of Stalin's era, the first success among which was that of *The Taste of Power*, by Ladislav Mňačko. From 1969, literature again declined.

Translations

Editorial activities are concentrated in numerous specialized government publishing houses, and books are among the cheapest goods in Czechoslovakia. Translations of foreign literature predominate, and there is interest in both Western and Soviet literature. Asian and Latin-American books and publications from other parts of the developing world are also available.

Music and films. The Czechs and Slovaks are traditionally musical, and operas and symphonies have a high place in cultural life. Such composers as Bedřich Smetana, Antonín Dvořák, Leoš Janáček, and Bohuslav Mar-

tinů have international reputations, and their works are often played in the annual spring music festival in Prague. In addition, jazz and related musical forms, and their performers, are also very popular, as is folk music.

Czechoslovak motion pictures are noted for their delicate and sympathetic treatments of basic human situations. In fact, modern Czechoslovak literary, theatrical, and film creations are all characterized by an interest in everyday life combined with the application of new forms of expression. A notable example of this has been the continued success of the combination of film, ballet, and theatre known as Laterna magica, which first attracted international attention at the world expositions in Brussels (1958) and Montreal (1967).

Fine, applied, and popular arts. In painting and in sculpture, abstract schools have made themselves felt, but Realism generally prevails. One of the best known painters before independence was Josef Mánes, and after his time Alphonse (Czech Alfons) Mucha gained world renown, as did the half-Czech Oskar Kokoschka. The glass designer and sculptor René Roubíček is also world renowned in his restricted field. In the applied arts, the manufacture of glass ornaments, the traditional north Bohemian costume jewelry, and toys is probably best known. Popular art has been preserved above all in useful objects in ceramics and wood; embroideries and traditional costumes are now of less importance.

State Library

Libraries and museums. The largest library is the State Library of the Czech Socialist Republic in Prague, created in 1958 by merger of several older libraries (one of which, the University Library, was founded in 1348); its collections include 4,600,000 volumes. The National Museum Library, also in Prague and founded in 1818, has about 2,400,000 volumes. Other major collections are in the Slovak National Library in Martin (2,040,000 volumes), the Slovak Technical Library of Bratislava (1,100,000), and the university libraries in Bratislava and Brno (1,330,000 and 2,390,000, respectively). Public lending libraries are found in every community.

Among the many museums, in both Prague and the provinces, may be mentioned three in Prague: the National Museum (founded 1818), the National Gallery (1796), and the Museum of Decorative Arts (1885), the last housing one of the world's largest glass collections.

The media of mass communication. Among the nearly 1,200 varied periodicals published in Czechoslovakia, there are only about 30 newspapers, the leader of which is the KSČ organ *Rudé Právo*. The national television network reaches the whole country. Half of the programs are devoted to music and a quarter to news and reportage.

PROSPECTS

Czechoslovakia, a rather small country situated in the centre of Europe, is characterized by a great internal variation, a fact that Czechoslovaks feel justifies the words of the national anthem, which claim that the country is "a paradise to look at." Complicated natural conditions, limited resources of raw materials, and a landlocked position are reflected in the national character, demography, and economy. Czechs and Slovaks entered the modern period as peoples without a large noble class and higher social order, and subsequent developments have made them cosmopolitan, with social consciousness and a strong democratic spirit. Their prospects centre on the intensification and modernization of the national economy, but also important are the improvement of surviving backward areas and the smoothing out of regional differences.

BIBLIOGRAPHY. The literature on Czechoslovakia is predominantly written in Czech. Foreign sources are often out of date, but competent works with sections on Czechoslovakia include P. GEORGE and H. SMOTKINE, *Les Républiques socialistes d'Europe centrale* (1967); A.F.A. MUTTON, *Central Europe: A Regional and Human Geography*, 2nd ed. (1968); R.H. OSBORNE, *East Central Europe: A Geographical Introduction to Seven Socialist States* (1967), N.J.G. POUNDS, *Eastern Europe* (1969), and R.E.H. MELLOR, *Eastern Europe: A Geography of the Comecon Countries* (1975), all detailed works. Regional geographies include M. BLAZEK, *Ökonomische Geographie der Tschechoslowakischen Republik*, in German (1959; also pub. in Russian, 1960); J. DEMEK *et al.*, *Geography of Czechoslovakia* (1971); V. HAUFLER, *Changes in the Geographical Distribution of Population in Czechoslovakia* (1966), and V. HAUFLER *et al.*, *Czechoslovakia: Land and People* (1968); population data are in WORLD POPULATION YEAR, *La Population de la Tchécoslovaquie* (1974), and detailed analysis in the official *Vývoj společnosti ČSSR*, based on the 1970 census figures; detailed information may also be found in English in the *Atlas Československé socialistické republiky* (1966); M. BLAZEK, *Ekonomická geografia ČSSR* (1964); *Príručni slovník naučný*, vol. 1, pp. 389–462 (1962), a concise encyclopaedia; and in the *Statistical Yearbook of the Czechoslovak Socialist Republic* (annual).

(M.Bl./R.H.O./Ed.)

Dahomey

The Republic of Dahomey—which in 1975 was renamed People's Republic of Benin—is one of the smaller independent states of West Africa. With an area of 43,500 square miles (112,600 square kilometres), it consists of a long wedge of territory extending for about 420 miles (675 kilometres) from the Niger River, which forms part of its northern frontier, to the Atlantic Ocean in the south, on which it has a 78-mile (125-kilometre) seaboard. Its population, estimated to be 3,100,000 in 1975, is composed of a variety of ethnic groups, the largest of which is composed of Fon, or Dahomean, people, numbering about 850,000 (1969 estimate). Dahomey is bounded to the west by Togo, to the northwest by Upper Volta, to the northeast by Niger, to the east by Nigeria, and to the south by the Atlantic Ocean. The capital is Porto-Novo (population, according to 1972 estimate, 100,000). Cotonou is the de facto capital.

A grouping of ancient kingdoms

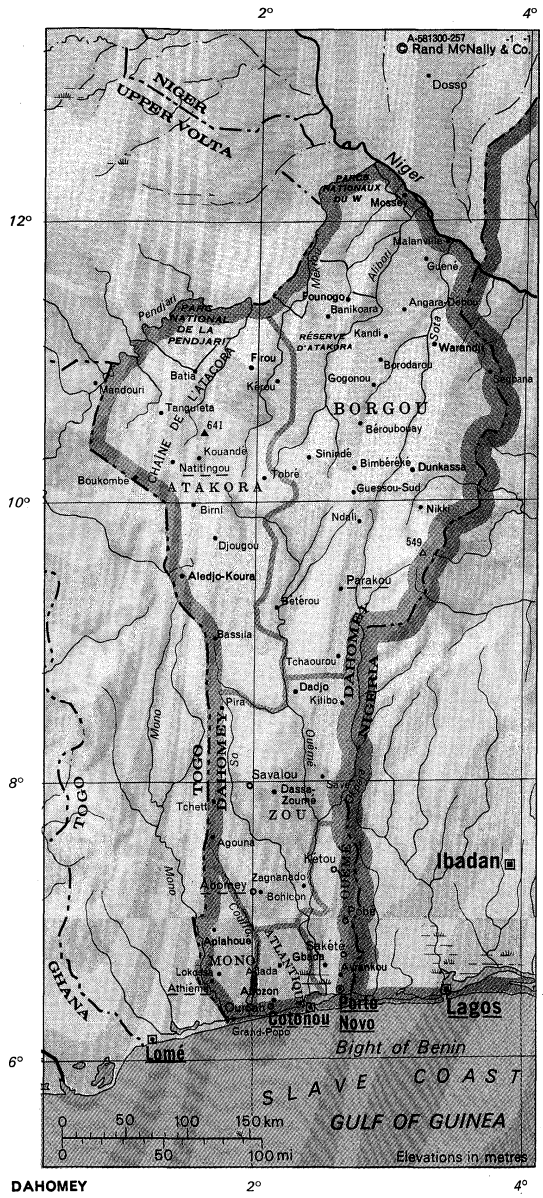
A former French colony, which gained its independence in 1960, Dahomey has diverse and distinctive cultural heritages. Today it is composed of a somewhat uneasy grouping of ancient kingdoms or chiefly states that were formerly at war with one another; one of these states, Dan-Homé ("On the Stomach of Dan," Dã, or Dan, being a conquered 17th-century kinglet), gave its name to the entire country. It was this kingdom, which had its capital at Abomey, that was renowned for its Amazon warriors —an elite corps of women soldiers who fought in the armies of the king of Abomey. The name Benin was that of a major West African state of earlier times.

Modern Dahomey has been characterized by political instability; there were eight changes of government and four military coups in the first decade of independence alone. The economy is dependent upon palm oil—which, along with palm kernels and products, accounts for two-fifths of the value of all exports—and upon aid, primarily from France. The discovery of offshore oil has raised hopes for a new source of national income. The economy was strengthened by the opening of an artificial deep-water harbour at Cotonou in 1965, a port that is now the commercial capital and largest city in the country (population estimated to be 175,000 in 1972). (For coverage of an associated physical feature, see NIGER RIVER; for historical aspects, see WEST AFRICA, HISTORY OF.)

The landscape. *Relief.* Dahomey consists of five natural regions—a coastal region, the *barre* country, the Dahomeyan plateaus, the Atacora Massif, and the Niger plains.

The coastal region is low, flat, and sandy, backed by lagoons. It forms, in effect, a long sandbar on which grow clumps of coconut palms; the lagoons are narrower in the western part of the country and wider in the east, and some are interconnected. In the west the Grand-Popo Lagoon extends into neighbouring Togo, while in the east the Porto-Novo Lagoon provides a natural waterway to the port of Lagos, Nigeria, although its use is discouraged by the political boundary. Only at Grand-Popo and at Cotonou do the lagoons have outlets to the sea. The outlet at Grand-Popo is known as the Bouche du Roi ("King's Mouth"), a French corruption of the Portuguese name, Bocco de Rio ("River Mouth").

Behind the coastal region occurs the *barre* country—the word being a French adaptation of the Portuguese word *barro* ("clay"). A fertile clay plateau, the *barre* region contains the Lama Depression, a vast swampy



DAHOMEY

MAP INDEX

Cities and towns

Abomey	7·11n	1·59e
Adjohon	6·42n	2·28e
Agouna	7·34n	1·42e
Ahozon	6·23n	2·11e
Aledjo-Koura	9·21n	1·27e
Allada	6·39n	2·09e
Angara-Débou	11·19n	3·03e
Aplahoué	6·56n	1·41e
Athiémé	6·35n	1·40e
Avrankou	6·33n	2·40e
Banikoara	11·18n	2·26e
Bassila	9·01n	1·40e
Batia	10·54n	1·29e
Béroubouay	10·32n	2·44e
Bétérou	9·12n	2·16e
Bimbéréké	10·13n	2·40e
Birni	10·00n	1·31e
Bohicon	7·12n	2·04e
Borodarou	10·59n	2·53e
Boukombé	10·11n	1·06e
Cotonou	6·21n	2·26e
Dadjo	8·34n	2·14e
Dassa-Zoumé	7·45n	2·11e
Djougou	9·42n	1·40e
Dunkassa	10·22n	3·08e
Firou	10·55n	1·56e
Founogo	11·28n	2·32e
Gbada	6·37n	2·31e
Gogonou	10·50n	2·50e
Grand-Popo	6·17n	1·50e
Guené	11·44n	3·13e
Guessou-Sud	10·03n	2·38e
Kandi	11·08n	2·56e
Kérou	10·50n	2·06e
Kétou	7·22n	2·36e
Kilibo	8·34n	2·36e
Kouandé	10·20n	1·42e
Lokossa	6·38n	1·43e
Malanville	11·52n	3·23e
Mossey	12·17n	2·54e
Natitingou	10·19n	1·22e
Ndali	9·51n	2·43e
Nikki	9·56n	3·12e
Ouidah	6·22n	2·05e
Parakou	9·21n	2·37e
Pira	8·30n	1·44e
Pobé	6·58n	2·41e
Porto Novo	6·29n	2·37e
Sakété	6·43n	2·40e
Savalou	7·56n	1·58e
Savé	8·02n	2·29e
Segbana	10·56n	3·42e
Sinindé	10·21n	2·23e
Tanguiéta	10·37n	1·16e
Tchaourou	8·53n	2·36e
Tchetti	7·50n	1·40e
Tobré	10·12n	2·08e
Warandji	10·57n	3·27e
Zagnanado	7·16n	2·21e

Physical features and points of interest

Alibori, *river*	11·56n	3·17e
Atacora, Chaîne de l', *mountains*	10·45n	1·30e
Atakora, Réserve d', *wildlife refuge*	11.30n	2·10e
Benin, Bight of	6·00n	3·00e
Couffo, *river*	6·35n	1·59e
Guinea, Gulf of	5·30n	3·00e
Mékrou, *river*	12·24n	2·49e
Mono, *river*	6·17n	1·51e
Niger, *river*	11·37n	3·37e
Okpara, *river*	7·40n	2·35e
Ouémé, *river*	6·29n	2·32e
Pendjari, *river*	10·54n	0·51e
Pendjari, Parc National de la, *national park*	11·20n	1·15e
Slave Coast	6·25n	3·00e
So, *river*	6·28n	2·25e
Sota, *river*	11·52n	3·24e

region stretching from Abomey to Allada. The landscape is generally flat, although occasional hills occur, rising to about 1,300 feet.

The Dahomeyan plateaus, four in number, occur in the Abomey, Kétou, Aplahoué, and Zagnanado districts. The plateaus consist of clays on a crystalline base. The Abomey, Aplahoué (or Parahoué), and Zagnanado plateaus are from 300 to 750 feet high, and the Kéton plateau is up to 500 feet in height.

The Atacora Massif, in the northwest of the country, forms a continuation of the Togo mountain chain, running southwest to northeast and reaching an altitude of 2,146 feet at its highest point. It consists of a quartzite plateau.

The Niger plains, in the northeast of Dahomey, slope down to the Niger River Valley. They consist of clayey sandstones.

Drainage. Apart from the Niger River, which, with its tributaries the Mékrou, Alibori, and Sota, drains the northeast part of the country, the three principal rivers in Dahomey are the Mono, the Couffo, and the Ouémé. The Mono, which rises in Togo, forms the frontier between Togo and Dahomey near the coast. The Couffo, near which stands Abomey, flows down from the Dahomeyan plateaus to drain into the coastal lagoons at Ahémé. The Ouémé rises in the Atacora Massif and flows southward for 310 miles; near its mouth it divides into two branches, one draining into Porto-Novo Lagoon and the other into Nokoué Lake. The Atacora Massif forms a watershed between the Volta and Niger basins.

Climate. Two climatic zones may be distinguished—a southern and a northern. The southern zone has an equatorial type of climate with four seasons—two wet and two dry. The principal rainy season occurs between mid-March and mid-July; followed by the shorter dry season, which lasts to mid-September; followed by the shorter rainy season, which lasts to mid-November; after which the principal dry season occurs, lasting until the rains begin again in March. The amount of rain increases toward the east. Grand-Popo receives only about 32 inches a year, whereas Cotonou and Porto-Novo both receive approximately 50 inches. Temperatures are relatively constant, varying between 72° and 93° F (22° to 34° C).

Two climatic zones

In the northern climatic zone only two seasons, one dry and one rainy, occur. The rainy season lasts from May to September, with most of the rainfall occurring in August. Rainfall amounts to about 53 inches a year in the Atacora Massif and in central Dahomey; further north it diminishes to about 38 inches. In the dry season the *harmattan*, a hot, dry wind, blows from the northeast from December to March. Temperatures average about 80° F (27° C), but the temperature range varies considerably from day to night. In January, the hottest month, day temperatures may rise to 110° F (43° C).

Vegetation. The original rain forest, which covered most of the southern part of the country, has now largely been cleared, except near the rivers. In its place, many oil palms and rônier palms have been planted and food crops are cultivated. North of Abomey the vegetation is an intermixture of forest and savanna (grassy parkland), giving way further north to savanna. Apart from the oil and rônier palms, trees include coconut palms, kapok, mahogany, and ebony.

Animal life. The animal life is varied and includes elephants, panthers, lions, antelope, monkeys, wild pigs, crocodiles, and buffalo. In the extreme north is a game reserve, the Parc National du "W," which extends into Upper Volta and Niger. There are many species of snakes, including pythons and puff adders. Birds include guinea fowl, wild duck, and partridge, as well as many tropical species.

The landscape under human settlement. The southern third of the country is by far the most densely peopled; it contains about two-thirds of the population. On the coastal sandbars, coconut plantations, often European-owned, have been established. In the lagoon region, fish are caught in wicker traps; fishing villages often consist of houses built upon stilts. The Ouémé Delta region is heavily populated. The cultivation of subsistence crops, such as maize (corn), cassava, and sweet potatoes, is intensive on the outskirts of the towns. The *barre* region and the Dahomeyan plateaus are planted with oil palms, which form the cash crop, as well as with subsistence crops. To the north, the aspect of the countryside changes as savanna vegetation increases and the population diminishes; some areas are uninhabited, except by Fulani nomads. Villages, instead of being encountered frequently as in the south, become scattered.

Although some of the towns are beginning to assume a modern aspect, vestiges of former centuries are still to be seen, such as the old Danish and English forts at Ouidah, dating from the time of the slave trade, for which Dahomey was a centre. Cotonou is primarily a modern European-built city, whereas Porto-Novo, an old African-founded town, combines both African and European features. Abomey, the ancient capital, now somewhat in decline, remains an important market and is also a centre for crafts. Parakou is an important market town.

The people. As mentioned, most of the people live in the southern third of the country. Many of these are clustered about the new port of Cotonou, which is the focus of the commercial and political life of the country.

Ethnic and linguistic groups. While French is the official language, the most widely spoken languages are Fon, Mina, Yoruba, and Dendi. The Fon people are predominant in the south; together with related groups they number about 700,000. The Yoruba, related to the Yoruba of Nigeria, number about 220,000. The Adja, Ge (Mina), and Gun peoples, living in the southeast and southwest, number about 200,000. In upper Dahomey the principal group is the Bargu (Bariba), a Voltaic people numbering about 175,000; the Somba, who live between Atacora and Togo, number another 90,000; and the Dendi, who are associated with the Niger Valley, number about 30,000. There are also a number of smaller African ethnic groups. About 5,000 Europeans, mostly French and Portuguese, also live in the country.

Dahomey, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 estimate
Départements				
Atlantique	1,200	3,200	309,000	461,000
Atacora	12,000	31,200	313,000	372,000
Borgou	19,700	51,000	305,000	385,000
Mono	1,500	3,800	290,000	390,000
Quémé	1,800	4,700	463,000	598,000
Zou	7,200	18,700	425,000	586,000
Total Dahomey	43,500*	112,600	2,106,000*	2,792,000

*Figures do not add to total given because of rounding.
Source: Official government figures.

Religious groups. Dahomey has always been a centre for African religion, and the great majority of the population remain animist. In the south, fetish practices (centring on an object regarded with awe as the embodiment of a powerful spirit) are maintained, and the cult of vodun (voodoo), for which Dahomey is renowned, retains its traditional strength. The town of Ouidah is a religious centre. Islām has adherents in the north and southeast, about 15 percent of the total population being Muslims. About 17 percent of the total population is Christian, of which about 85 percent are Catholic.

Demography. The population numbered 2,800,000 in 1971. The annual rate of natural increase is about 2.6 percent. About 90 percent of the population are rural, about 10 percent urban.

The national economy. *Resources.* The principal agricultural resource of Dahomey consists of its palm-oil plantations. Only a few stretches of forest remain, mostly in the south and in central Dahomey; these contain mahogany, iroko, samba, and other tropical hardwoods. The rivers and lagoons are rich in fish. Known mineral deposits include iron-ore deposits in the Atacora and Kandi region; limestone deposits suitable for the manufacture of cement at Onighlo; chrome ore and a little gold in the northwest near Natitingou; about 5,000,000 tons of marble at Dadjo; an important deposit of pottery clay at Sakété; and ilmenite (a mineral source of titanium) near the coast. In addition, oil has been located about eight miles offshore, about 18 miles from Cotonou, at a depth of about 100 feet below the seabed.

Sources of national income. About 90 percent of the population lives solely by agriculture. As mentioned, palm oil constitutes the main cash crop; together with palm kernels, this represents about two-fifths of the value of all exports, amounting to about 380,000 tons a year. Other exports are shelled peanuts (groundnuts), copra, coffee, and cotton. About half of the 2,470,000 acres under cultivation are devoted to palm-oil production. Subsistence crops include manioc, sweet potatoes, maize (corn), and sorghum. Mining is as yet of little importance since the distance of most mineral deposits from the coast renders them uneconomic to exploit. Some industrial diamonds, however, are exported. About 17,000 tons of fish are caught annually in the lagoons and rivers, while coastal fishing produces a further 5,000 tons. Most of the fish is exported to Nigeria or Togo. Shrimp fishing is beginning, using modern vessels.

Manufacturing plants and secondary industries include several palm-oil-processing plants, in Ahozon, Avrankou, Bohicon, Cotonou, Gbada, and Pobé; several modern bakeries; a soft-drink plant; a brewery; and two shrimp-processing plants. Other plants specialize in cotton milling, processing kapok, and the manufacture of jute.

Energy. Electricity is generated by diesel engines. There are three power-generating plants, located at Cotonou, Porto-Novo, and Parakou, with a combined capacity of 6,500 kilowatts. It is anticipated that Dahomey will purchase electricity from Ghana's Volta River Project. An agreement has also been concluded between Dahomey and Togo for the construction of four dams on the Mono River as part of a hydroelectric project that would have a capacity of 115,000 kilowatts.

The principal imports are food, drink, tobacco, cement, petroleum products, machinery, automobiles, and paper; the principal exports are palm oil and palm kernels, industrial diamonds, cotton, and peanuts. France is the principal trading partner for both exports and imports, followed by The Netherlands and the United States for exports and Italy and West Germany for imports.

Management of the economy. During the colonial period, the economy of Dahomey was integrated with that of former French West Africa; since independence, Dahomey has remained partially dependent upon France for financial aid, so that the economy—which relies primarily upon the production of a single commodity, palm oil—still retains some of the characteristics that it possessed during the colonial era. Palm oil is purchased from individual producers by the *Société Nationale des Huileries du Dahomey* (Snahda)—a quasi-governmental agency, which markets the oil abroad.

The political-basis economy remains mixed, combining both governmental controls with private enterprise. An investment code seeks to attract capital from abroad by providing tax benefits and other incentives. Large enterprises are established for periods of 25 years or more, with the cooperation of the state. Economic policy from 1966 to 1970 was established in a five-year plan that sought to increase production and to stimulate economic growth, giving priority to rural and industrial development and to investing in transport and other necessary services. The present government envisages that, henceforth, economic priorities will be established by a series of two-year plans. About half the government's regular budget is obtained from indirect taxes, and about a

quarter from direct taxes; the remainder is obtained from various sources.

There are four trade unions—the Union Générale des Travailleurs Dahoméens (UGTD), the Union Générale des Syndicats du Dahomey (UGSD), the Confédération Dahoméenne des Travailleurs Croyants (CDTC), and the Confédération Générale des Travailleurs Dahoméens (CGTD).

The standard of living

The principal difficulty faced by successive Dahomeyan governments has remained the low standard of living that prevails. Difficulties have been accentuated by the rate of growth of the population, the uncertain rate of economic growth, and an insufficient agricultural production; these and other factors have combined to produce a financial imbalance that remains quasi-permanent. Various governments have sought to overcome these problems by seeking to stimulate further agricultural production and to promote industrial expansion.

Transportation. There are 4,100 miles of roads, of which about 470 miles are surfaced. The principal roads are two: one running along the coast from the Togo border in the west via Cotonou and Porto-Novo to the Nigerian border in the east, and one running north from Cotonou to Parakou and on to the Niger frontier. The coastal route is currently being improved, and a second bridge is to be added where the road crosses the lagoon at Cotonou. Road transport is usually by truck; a central truck park at Cotonou is capable of accomodating 15,000 vehicles.

There are about 400 miles of railroad track consisting of three lines. One line runs for about 30 miles from Cotonou to the Togo border, from where it continues to the Togolese port-capital of Lomé. A second line runs for about 66 miles from Cotonou via Porto-Novo to Pobé. The third line runs north from Cotonou to Parakou, a distance of about 270 miles. Following the discovery of important uranium deposits in the Republic of Niger, an extension of the Parakou line to Dosso in Niger was planned.

The lagoons, running parallel to the coast, are used as waterways by small craft. The Ouémé River is navigable for about 125 miles, the Couffo for about 80 miles, and the Mono for about 60 miles. About 600 ships enter the port of Cotonou each year. The airport at Cotonou links Dahomey with other West African countries and with France. The government maintains a small domestic air service.

Administration and social conditions. *The structure of government.* A national charter, adopted by the government in 1970, is intended to be Dahomey's fundamental constitutional instrument until elections are held in 1976.

The presidential council

Executive and legislative powers are held by a presidential council of three, which is headed by the president, and a group of 11 ministers. The president has responsibility for the Ministers of Defense and of the Interior. Despite the regular changes that have marked Dahomey's political life and that may be said to amount to a state of institutionalized instability, some continuity is preserved by the various ministries.

For purposes of regional government Dahomey is divided into six *départements* (provinces)—Atlantique, Atacora, Borgou, Mono, Ouémé, and Zou—each of which is administered by a prefect. Each *département* is divided into subprefectures administered by a subprefect, of which there are 35 throughout the country; each subprefecture is divided into an *arrondissement* (district), headed by a district leader, of which there are 64. In addition, there are five townships.

At the head of the judicial system is the supreme court. There is also a council of the magistracy, a court of appeal at Cotonou, a tribunal of the first instance, and local legal tribunals.

Defense is the responsibility of the Dahomeyan Army and the gendarmerie, which together number about 3,000. The army itself consists of three battalions with a strength of 1,700 men. There is a small air force. The president is the supreme commander of the Dahomeyan forces and is also minister of defense; there is no army chief of staff, the highest ranking officer being secretary general of defense. Internal security is the responsibility of the Ministry of the Interior, which supervises the police services in each *département*.

Administration. About 30 percent of the school-age population is receiving education. There are about 150,000 students in primary schools, of which about 85,000 are attending public schools, and 50,000 private (missionary) schools. About 13,000 students are receiving secondary education, and another 1,000 are engaged in technical studies. A Dahomeyan university was opened in 1970. It is estimated that about 700 Dahomeyans are receiving higher education abroad, mostly in France or at universities elsewhere in Africa.

Health services are limited by budgetary restrictions. There are about 80 doctors in the country—one for every 31,000 inhabitants. There are four hospitals, located at Cotonou, Porto-Novo, Abomey, and Parakou. In addition there are 30 clinics, about 150 dispensaries, about 50 maternity centres, and six leprosariums. Campaigns are waged periodically against smallpox, meningitis, leprosy, malaria, and sleeping sickness.

Unemployment constitutes a serious problem. In 1961 it was estimated that wage earners numbered about 30,000, of which about 13,000 were employed by the public sector and 17,000 by the private sector (primarily industry, construction, transport, and commerce). By 1964, however, it was estimated that the total number of wage earners had fallen to 18,000. Since then, austerity measures have further reduced the number of wage earners. In 1967 a labour code was adopted; different wage scales are applied to industrial and to agricultural employment. The great majority of Dahomeyans, however, continue to live by subsistence agriculture.

Cultural life and institutions. As already mentioned, Dahomey has inherited a strong and distinctive form of African culture, which is manifest in music, dance, legend, and craftsmanship. Radio Dahomey broadcasts from Cotonou but also has a transmitter at Parakou. There is no television. There is one daily newspaper, *Daho-Express*, and two biweeklies, *L'Action Populaire* and *La Dépêche Dahoméenne*, the last intended for rural readers. The *Journal Officiel* appears bi-weekly.

Prospects for the future. The instability that has characterized Dahomey's political life and that stems from its difficult economic situation appears likely to continue, particularly if popular dissatisfaction with taxes and separatist tendencies in the north continue to exert an unsettling influence. The adherence of Dahomey to the franc zone, however, assures some continuing viability of the national economy. Dahomey's adherence to the Conseil de l'Entente, of which the other members are the Ivory Coast, Niger, Togo, and Upper Volta, may also prove beneficial. The extension of the rail system from Parakou to the Republic of Niger is expected to stimulate further economic development. The fragile economy will also be strengthened should the recently discovered offshore oil deposits prove exploitable.

BIBLIOGRAPHY. H. HUBERT, *Mission scientifique au Dahomey* (1908), a monumental work; P. HAZOUME, *Le Pacte de sang au Dahomey* (1937), and M. QUENUM, *Au Pays des Fons* (1938), for aspects of the peoples of Dahomey, their civilizations and traditional religions; SIR RICHARD BURTON, *A Mission to Gelele, King of Dahomé*, 2 vol. (1864), the classic account by an eyewitness of the customs; M.J. HERSKOVITS, *Dahomey: An Ancient West African Kingdom*, 2 vol. (1938); B. DAVIDSON, "Dahomey in the 17th and 18th Centuries," in *A History of West Africa, 1000–1800*, new ed. (1968), a succinct presentation; R. CORNEVIN, *Histoire du Dahomey* (1962), a general history of Dahomey from prehistoric times to the present, with an extensive bibliography; B. SCHNAPPER, *La Politique et le commerce français dans le golfe de Guinée, de 1838 à 1871* (1961), the most important work on 19th-century Dahomey. Twentieth-century accounts are: J.F. RESTE, *Le Dahomey: réalisation et perspectives d'avenir* (1934); H. DESANTI, *Du Danhomé au Benin-Niger* (1945); A. AKINDELE and C. AGUESSY, *Le Dahomey* (1955); G.M. CARTER (ed.), *Five African States* (1963). For social evolution since World War II, see C. TARDITS, *Porto-Novo: les nouvelles générations africaines entre leurs traditions et l'occident* (1958); and P. BRASSEUR, *Porto-Novo et sa palmeraie* (1963).

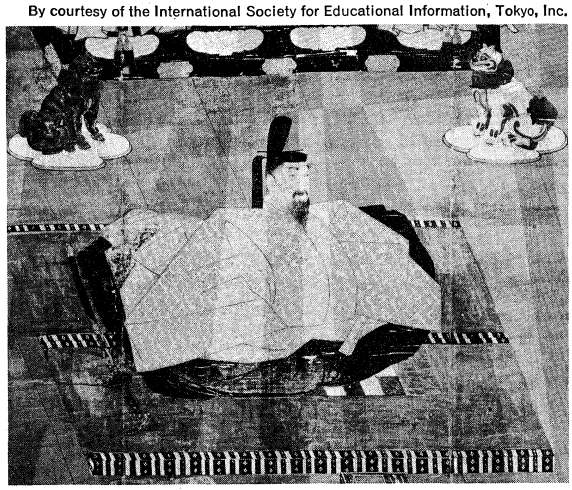
(S.S.A.)

Daigo II

Few emperors in Japan's history occupy as controversial a place as Daigo II (also known as Go-Daigo), who reigned in the 14th century. Ranked by some as one of his nation's greatest sovereigns, he is regarded by others as a ruler of limited abilities. To some he personifies courage and integrity for daring to lead a revolt against the military government in whose shadow emperors and their courts had existed for a century and a half. Others see in him a proud and inept ruler, whose clumsy efforts to restore the monarchy plunged the nation into civil war and divided the Imperial family into two rival dynasties.

Born in 1287, Daigo ascended the throne in 1318, when the nation was in one of the more turbulent periods of its history. Politically, authority was uncertainly divided between two governments—the de jure government of the Emperor and his court in Kyōto, and the de facto government of the Shogun (the military overlord) and his court in Kamakura in eastern Japan. Neither government was stable and united, and neither Emperor nor Shogun actually wielded authority in his respective government, each having become the puppet of powerful families.

By courtesy of the International Society for Educational Information, Tokyo, Inc.



Daigo II, colour on silk by an unknown artist. In the collection of Daitoku-ji, Kyōto.

Rule by Emperor and Shogun

In Kyōto, political authority was still further diffused by the introduction in the 11th century of a curious practice known as *insei* ("cloistered rule"). Emperors, in their desire to recover their prerogatives, abdicated and entered a monastery, where they organized a new government and proceeded to rule from retirement. A minor would be placed on the vacated throne and would await the day he, too, could retire so that he could begin to rule. Of Daigo II's seven immediate predecessors, six were minors, one of whom acceded at the age of three and another at seven.

Adding to the political confusion in Kyōto was the practice of alternating the throne between the senior and junior branches of the Imperial family, which had been feuding over the question of succession for years. It was in accordance with the agreement forced upon the quarrelling factions by the shogunate that Daigo of the junior branch ascended the throne in 1318.

No less anomalous was the situation in Kamakura, where control of the shogunate had passed from the Minamoto to the Hōjō family. Not being eligible for the office of shogun, the Hōjō were content to rule as regents for the puppet shoguns they appointed, at first from among the younger scions of the Fujiwara family and later from the Imperial family. But by the 14th century Hōjō influence itself had declined considerably, and the regency had become the instrument of yet another family, the Nagasaki. Its ascendancy was facilitated by the youth of the regent, Hōjō Takatoki, who, at investiture, was only eight years old. As he grew to manhood, his questionable intelligence and dissolute ways—spending much time, for example, watching dogfights—led to a widespread loss of confidence in the shogunate. Further alienation of many traditional supporters of the shogunate was caused by the favouritism to friends and relatives shown by Nagasaki Takasuke, the man who controlled the regent.

In view of these and other signs of growing discontent, Daigo, even before his enthronement, began to plot the overthrow of the shogunate and the restoration to power of the Imperial court. He continued these secret efforts after his installation and throughout the decade of the 1320s with encouraging results, but in 1331 the plot was exposed. Captured by Kamakura forces while attempting to flee from Kyōto, Daigo was sent into exile to the Oki Islands in the Sea of Japan.

But Daigo had triggered the revolt he had plotted so long. During his first year of confinement there was sporadic fighting between his supporters and shogunate forces, and in 1333 decisive events, marked by much treachery and violence, took place. The commander of some of the shogunate's forces turned and attacked Kamakura instead of the shogunate's enemies. Caught by surprise, Hōjō Takatoki and his supporters chose to take their own lives, thus bringing to an end the 150-year regime of Japan's first military government. Equally dramatic was the defection of Ashikaga Takauji, who was in command of the main armies of the shogunate. Instead of seizing Kyōto for the shogunate, he attacked the shogunate garrison there and turned the city over to Daigo, who, in the meantime, had contrived to escape from Oki.

The shogunate overthrown

If Daigo was grateful to the former vassals of the Hōjō for having made possible the "Kemmu restoration," as this series of events is known, he failed to show it. He neither rewarded them nor brought them into his new government, except for Takauji, who was given some lands and the comparatively low rank of counsellor. Takauji expected to be designated shogun, but Daigo foolishly named his own son, Prince Morinaga, to the post. By such callous acts the sovereign alienated Takauji as well as a large segment of the warrior class at a time when their continued support would have assured the success of the restoration. In fact, historians are generally agreed that Daigo's greatest failing and the cause of many of his troubles was his inability to see that rule by a strictly civilian aristocracy was no longer feasible.

By 1335 open warfare broke out between Takauji, who proclaimed himself shogun, and Daigo, who pronounced him a rebel. Although the loyalists distinguished themselves in battle, in the end they were overpowered by the numerically superior forces of Takauji, who had gone as far as Kyushu in the southwest to raise a vast army. In 1336 Daigo, who had fled from Kyōto a number of times as the tide of battle flowed in and around the city, left the ancient capital for the last time. Takauji re-entered Kyōto and promptly elevated Kōgon of the senior Imperial line to the throne. Daigo established his own court in the Yoshino Mountains to the south of Kyōto, where he died in 1339. Thus, from 1336 until 1392, when the rival factions of the Imperial family were to be reunited, Japan witnessed the spectacle of two contending Imperial courts—the southern court of Daigo II and his descendants, whose sphere of influence was restricted to the immediate vicinity of the Yoshino Mountains, and the northern court of Kōgon and his descendants, which was under the domination of the Ashikaga family.

BIBLIOGRAPHY. The *Taiheiki* (Chronicle of the Grand Pacification) of unknown authorship is the most detailed primary source on Daigo II, of which there is an excellent translation by HELEN CRAIG MCCULLOUGH, *The Taiheiki: A Chronicle of Medieval Japan* (1959). For biographies of and relevant data on all 124 emperors of Japan, see R.A.B. PONSONBY-FANE, *The Imperial House of Japan* (1959). The most detailed and authoritative general history in English on the subject is GEORGE SANSOM, *A History of Japan*, vol. 1 and 2 (1958–61). For the views of Japanese thinkers of the Tokugawa period on the subject, see DAVID MAGAREY EARL, *Emperor and Nation in Japan: Political Thinkers of the Tokugawa Period* (1964).

(M.Sh.)

Dairying and Dairy Products

Dairying is the production and marketing of milk, usually cows' milk, and its products; it includes the care of cows, their breeding, feeding, management, and milking. The milk must be collected, processed into dairy products, and marketed. All of these operations have been studied and improved by physiological, genetic, nutritional, chemical, microbiological, technological, economic, and marketing research and development.

Historical development. Cattle, goats, and sheep have been kept by man for the production of milk throughout the time of recorded history. Milk and especially soured milks, butterlike products, and cheese were probably common foods of the peoples roaming the grasslands of Asia with their sheep and cattle thousands of years ago.

Biblical references

In the biblical narratives, Abel, son of Adam, was a "keeper of sheep," and Abraham served milk to the three divine beings who appeared to him at Hebron. Canaan was idealized as "a land flowing with milk and honey," and cheese is mentioned in the Book of Job. The Hindu Vedas, written before 1200 BC, mention the use of butter as food. In all these instances the mention of milk or its products implies much earlier use.

The ancient Europeans did not use butter as food but as a pharmaceutical, for skin injuries and sore eyes. It was widely used as a hair oil and as fuel for lamps. Whey, the watery part left after separation of the solid curd from coagulated milk, was employed widely in Europe as a medicine in the Middle Ages. Milk sugar (lactose) seems to have displaced whey as a panacea subsequent to its isolation from whey in 1628.

The Mongols in the Middle Ages prepared concentrated milks in paste, and probably in dry form, and used them as rations on the march. Commercial processes for making concentrated and dried milk did not appear until the 19th century. A patent for "the concentration of milk" was granted to Gail Borden of the United States in 1856, and the utility of canned concentrated milk was demonstrated by its use in army rations during the American Civil War. F.S. Grimwade's British patent for producing dried milk was issued in 1855, but large-scale production of dried milks did not begin until 50 years later.

The domestication and keeping of milk-producing herd animals spread from southwest Asia to other parts of the world. At first the same animals were used for work, meat, and the production of milk. The milk was consumed at the point of production as milk or as domestically made dairy products. As urban centres developed, cows and goats were kept in larger numbers to produce milk in quantity, and transportation by wagon, rail, and finally tank truck became necessary.

Early man found that goats and sheep produced enough milk for his family. As milk needs increased, the cow became established as a volume producer. The Middle Ages saw many advances in development of milk-producing breeds, and by the 18th century the practice of selective breeding was well-established. Factories to process and pasteurize the milk became accepted as a safeguard for the milk supply. The largest herds today are found within a few hundred miles of heavily populated areas and usually where abundant feed is available. Rapid transportation and advanced technology are encouraging the rise of large, specialized dairy manufacturing plants in potentially high milk producing sections distant from population centres.

In the 20th century the dairy industry of the world was well established, but it faced a challenge from the growing number of economical substitute foods, notably vegetable-base margarine, but including imitation creams, cheese dips, evaporated milks, ice creams, and even milks. Many of the simulated products employ skim milk or sodium caseinate as a protein source since there is yet no vegetable protein having the mild, pleasant flavour of milk protein.

Milk production

Milk for man is produced in largest volume in dairy countries by the cow and the water buffalo. The goat also is an important milk producer in China, India, Egypt, and in many other Asian countries. Goat's milk is also produced in Europe and North America but, compared to cow's milk, goat's milk is relatively unimportant. Buffalo's milk is produced in commercial quantities in some countries, particularly India. Where it is produced, buffalo's milk moves in the same channels as cow's milk and in some areas the community milk supply consists of a mixture of both.

DAIRY HERDS

Dairy cows are divided into five major breeds: Ayrshire, Brown Swiss, Guernsey, Holstein-Friesian, and Jersey. There are many minor breeds, among them the Red Dane, the Dutch Belted, and the Devon. There are also dual-purpose breeds used to produce milk and meat, notably the Milking Shorthorn and the Red Polled.

Breed characteristics

The Ayrshire breed originated in Scotland. Animals of this breed are red and white or brown and white in colour, and they are strong, vigorous, and good foragers. Ayrshire milk contains about 4.1 percent butterfat. Switzerland is the native home of the Brown Swiss. These cows are silver to dark brown in colour with a black nose and tongue. Brown Swiss are strong and vigorous. The average fat test of the milk is 4.1 percent. The Guernsey breed originated on Guernsey Island off the coast of France. The Guernsey is fawn-coloured with clear white markings. The milk averages about 4.8 percent fat and carries a deep yellow colour. The Holstein-Friesian originated in The Netherlands. It is black and white in colour and large in size. Holsteins give more milk than any other breed; average butterfat is 3.7 percent. The Jersey breed originated on the isle of Jersey, Great Britain. Jersey cows are fawn in colour with or without white markings. They are the smallest of the major breeds but their milk is the richest, containing on the average 5.2 percent butterfat. The protein content of milk is highest for Guernsey (3.91 percent) and Jersey (3.92 percent) and lowest for Holstein (3.23 percent).

Breeding and herd improvement. The breeds of dairy cattle have been established by years of careful selection and mating of animals to attain desired types. Increased milk and butterfat production has been the chief objective, although recently the objective has tended to shift to increased milk and protein production. Production per cow varies with many environmental factors, but the genetic background of the cow is extremely important. Production per cow for cows in the 12 highest milk-producing countries is shown in Table 1.

The principles of breeding to improve production are being applied in lesser developed countries and show promise of increasing milk production. Progress is being made in India with cows and water buffalo.

Artificial breeding has developed into a worldwide practice. Bulls with the genetic capacity to transmit high milk-producing ability to their daughters are kept in studs. Dairy-farmer cooperatives usually operate the studs, with artificial insemination generally used. Semen may be frozen for shipment to any part of the world.

Feeding dairy cattle. The dairy cow is an efficient producer of human food from roughage not digestible by humans, thanks to a unique digestive system consisting of what is really a four-compartment stomach capable of handling roughages not digested by monogastric (one stomached) animals.

Pasture is the natural feed for dairy cattle, and an abundance of good pasture provides most of the requirements of a good dairy ration. An outstanding example of grassland dairying is New Zealand, where cows are on pasture all year and milk production costs are at a minimum. The farmer does not need to prepare and store feed for a long winter period. Feeding a balanced ration, however, rather than grass alone, increases milk production. In 1970 the average annual production per cow in New Zealand was 5,511 pounds (2,500 kilograms) of milk, while in the U.S., where supplemental feeding is common, it was 9,388 pounds or 4,258 kilograms (see Table 1). Pastures of poor quality must be supplemented with other feed such as green crops, summer silage, or hay.

During seasons when pastures are inadequate, cows need hay, silage, and grain in sufficient amounts and balance to supply nutrient needs, and to guarantee a nutritional reserve to keep milk volume and composition from declining.

Disease prevention. Disease is one of the greatest problems of the dairy farm. It is a constant threat and may make removal of valuable animals from the herd necessary. One study of removal of cows from a typical dairy herd showed that an average of 22 percent of cows were removed yearly and about a third of these were lost.

Herd management

Good health herd management includes cleanliness, isolation of sick or injured animals, keeping premises free of hazards that might cause injury, and continuous protection against poisonous plants and other material. Certain diseases, such as tuberculosis, require injections. Others, such as mastitis, require constant treatment. For some diseases there is no known cure; slaughter of the animal is the only way to stop spread of the infection. Foot and mouth disease is the most notorious of these; severe measures have been taken by most governments to exclude or control it.

MILKING AND BULK HANDLING ON THE FARM

Milk is produced by the cow from her blood, and a large amount of food is necessary for maintenance of a high producing cow. The products of digestion and absorption enter the blood and are carried to the udder. There the raw materials are collected and changed into milk components. Each time the blood passes through the udder a small fraction of the components is removed to make the milk. Some 400 pounds (50 gallons, or about 200 litres) of blood must pass through the udder to make one pound (about 450 grams) of milk. A daily flow through the udder of ten tons (20,000 pounds) of blood is required for a cow producing 50 pounds (22.5 kilograms) of milk per day. The energy required to produce milk components and to circulate the blood indicates the great importance of proper and abundant feed.

Use of milking machines

Today, most milking is done with machines by a carefully trained operator, usually twice a day, in stanchion barns or milking parlors. An experienced milker handles one to three machine units. The cows are first cleaned, and the teat cups put on. Pulsating vacuum draws the milk into a receiver or through piping into the farm milk tank (see Figure 1).

Milk is an extremely perishable commodity that must be cooled to 50° F (10° C) or less within two hours and must be maintained at that temperature until it is delivered to the consumer.

Milk is transported from farm to plant in a variety of ways, depending on the part of the world. In the Gujarat region of India, the milk is carried to a receiving station in jars on the heads of women who do the milking. The receiving station transports the milk in large cans to the plant by truck.

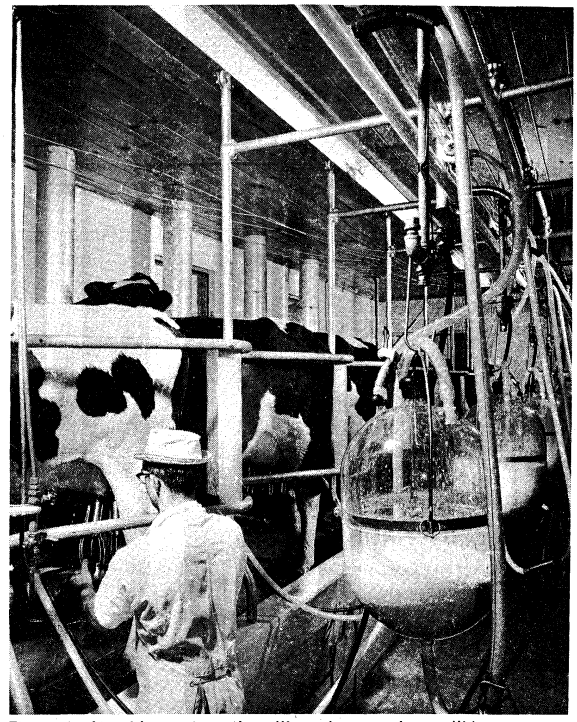


Figure 1: Attaching automatic milkers in a modern milking parlor (United States).
Grant Heilman

In the major milk-producing countries the milk is held cold in the farm tank or in cans until picked up once or twice daily by tanker or truck. Tankers pump the milk in at the farm and out into plant tanks on delivery. The tanker driver measures and samples each farmer's milk; fat and bacteria tests are run at the plant.

Pipelines have been introduced on a small scale in Europe for delivery of milk from farm to factory. Whether the method will be practical for long-range or large-scale operations remains problematical.

PRODUCTION AND CONSUMPTION STATISTICS FOR MILK

The production and utilization of milk by the largest milk-producing countries where statistics are available is shown in Table 1. The heaviest producing areas are the U.S., Europe, Australia, and New Zealand. In the U.S., almost half the milk produced is utilized in fluid form but in the dairy product exporting countries of central Europe and in Australia and New Zealand, the population requires a smaller percentage of production, leaving an exportable surplus of butter, cheese, casein, and milk sugar.

Table 1: Milk Production and Utilization in Specified Countries
(1970) Preliminary

country	production			utilization (percentage of production)				
	milk production (000,000,000 lbs)	milk cows (000,000 head)	production per cow (000 lbs)	fluid milk	butter	cheese	other uses*	feed
U.S.	117.4	12.5	9.4	46.2	20.5	16.7	15.1	1.5
France	65.5	9.6	6.9	19.7	33.1	23.4	4.6	19.1
West Germany	48.2	5.6	8.6	26.3	52.1	7.9	7.0	6.7
U.K.	27.3	4.5	6.1	68.2	12.2	11.3	4.8	3.6
Italy	20.6	3.5	5.9	36.3	14.3	32.1	2.3	15.0
Canada	18.3	2.5	7.4	32.8	42.3	13.0	8.2	3.7
The Netherlands	18.2	1.9	9.5	24.6	14.3	33.2	24.5	3.4
Australia	17.0	2.7	6.4	22.8	59.0	10.1	7.6	0.5
New Zealand	13.0	2.4	5.5	8.3	69.5	16.9	2.8	2.5
Denmark	10.2	1.2	8.9	18.5	55.7	14.8	6.7	4.3
Belgium	8.4	1.0	8.1	23.2	61.7	5.9	3.4	5.8
Ireland	8.0	1.7	4.8	18.7	47.5	7.8	10.6	15.4

*Includes milk used for ice cream, dried whole milk, canned milk, minor products, waste and balance.

Large quantities of dairy products move in international trade. Japan, the countries of Indochina, India, Central America, and other regions bordering the Equator are in general poorly suited for dairying. Where economic conditions permit, their dairy product deficit is overcome by imports. The basic requirement for a viable dairy industry is an abundance of feed and a temperate climate for cattle. Optimum conditions exist in Australia, New Zealand, the U.S., and Europe, hence the large milk production of these areas. Per capita consumption of dairy products tends to be largest in the milk-producing countries.

Prices for milk and dairy products have not increased as rapidly during the last half century as have prices of many other products. But greater economies of milk production and product processing, together with some government price-support activity, has kept profit margins attractive in most of the dairy countries.

Dairy products

Milk has been used by man to make many products both to increase variety and appeal and as a means of preservation. Almost half the milk produced in many countries is consumed as a fresh whole and skim-milk beverage type product. The remainder is manufactured into the more stable dairy products of commerce, principally butter and cheese, but including condensed and evaporated milks, dried milks, ice cream, and dairy by-products.

The following discussion of dairy products is based on cow's milk as the raw material because the processes so far developed have overwhelmingly used cow's milk. Similar processes are doubtless applicable to the buffalo's milk of India, China, Egypt, and the Philippines, the goat's milk of the Mediterranean region and elsewhere, to the milk of the reindeer, in northern Europe, and to the sheep's milk used in southern Europe. Special processing problems may be encountered with the milk of each of these species because of differences in composition and in the nature of the proteins. Simple pasteurization changes the flavour of goat's milk, and producers of this product often prefer to sell it unpasteurized. The fat of goat's milk is naturally homogenized and this is thought to improve its digestibility. Buffalo and goat milk when concentrated do not easily withstand sterilization for evaporated milk manufacture and so are rarely used for this purpose. A Cheddar-type cheese is made from buffalo milk in India, but its protein-calcium complex appears different from that of cow's milk and a way has not yet been found to cure the cheese as well as cow's milk cheese. Buffalo milk is successfully dried in India to produce powder for both infant and adult use. With research, the behaviour of the milk of various species can no doubt be adapted to the manufacture of all dairy products. The basic processes have been developed around cow's milk because of its supply and the need to preserve it for future use in distant markets.

Table 2: Average Composition of Milk of Different Species of Mammals
(percentage)

species	water	protein	fat	milk sugar	ash
Cow	87.2	3.5	3.7	4.9	0.7
Goat	87.0	3.5	4.3	4.3	0.9
Sheep	82.0	5.8	6.5	4.8	0.9
Indian buffalo	82.7	3.6	7.4	5.5	0.8
Chinese buffalo	76.8	6.0	12.6	3.7	0.9
Egyptian buffalo	82.1	4.2	8.0	4.9	0.8
Philippine carabao	78.5	5.9	10.4	4.3	0.9
Reindeer	63.3	10.3	22.5	2.5	1.4

FLUID AND CONCENTRATED MILKS

Fluid or market milk is generally considered to refer to bottled, fresh milk and its associated products, while concentrated milk refers to plain and sweetened condensed and evaporated milk.

Milk as it is secreted by the cow contains 12 to 13 percent solids. It is used in this form or it may be fermented, flavoured, or concentrated for improved utilization. It is the basic material from which all dairy products are produced. The fluid milk industry is composed of city milk plants that receive milk and process it into a variety of products, including pasteurized milk, cream, cultured buttermilk, chocolate milk, fortified skim milk, or yogurt. Specialized factories that make concentrated and dry milks, butter, cheese, and even ice cream are often located in rural heavy milk producing areas where hauling distances for the raw milk are minimal.

Composition. The same constituents are present in the milks of all mammals, but the proportions differ both among species and within a species (Table 2).

Table 3: Composition of Milk Products*

product	water (%)	fat (%)	protein (%)	lactose (%)	ash (%)	calcium (mg/100 g)	other (%)
Milk	87.0	3.9	3.5	4.9	0.7	118	
Half-and-half	80.2	11.5	3.1	4.5	0.7	108	
18% cream	74.5	18.0	2.8	4.1	0.6	102	
30% cream	63.3	30.0	2.5	3.6	0.6	85	
36% cream	58.0	36.0	2.2	3.3	0.5	75	
Plastic cream	18.2	80.0	0.7	1.0	0.1	—	
Dry cream	0.7	65.0	13.4	17.9	3.0	—	
Butter	16.0	80.6	0.6	0.4	2.4	20	
Butter oil	0.2	99.5	0.3	—	—	—	
Sweet-cream buttermilk	91.0	0.4	3.4	4.5	0.7	—	
Skim milk	90.5	0.1	3.6	5.1	0.7	121	
Cultured buttermilk	90.5	0.1	3.6	4.3	0.7	121	lactic acid 0.8
Yogurt	89.0	1.7	3.4	5.2	0.7	120	
Plain condensed skim milk	66.0	0.4	12.7	18.4	2.9	—	
Sweetened condensed skim milk	28.0	0.3	11.2	16.3	2.5	—	sucrose 42.0
Sweetened condensed milk	26.5	8.1	8.1	11.4	1.6	262	sucrose 44.3
Evaporated milk	73.8	7.9	7.0	9.7	1.6	252	
Ice cream	63.2	10.6	4.5	6.6	0.9	146	sugar 14.2
Ice milk	66.7	5.1	4.8	7.0	1.0	156	sugar 15.4
Sherbet, orange	67.0	1.2	0.9	1.4	0.1	16	sugar + trace lactic acid 29.4
Dry whole milk	2.0	27.5	26.4	38.2	5.9	909	
Dry skim milk	3.0	0.8	35.9	52.3	8.0	1308	
Dry malted milk	2.6	8.3	14.7	20.0	3.6	288	maltose and dextrin 50.5, fibre 0.3
Cheese, Cheddar	37.0	32.2	25.0	2.1	3.7	750	
Whey (Cheddar)	93.0	0.3	0.9	4.9	0.6	51	lactic acid 0.2

*These values represent average composition of consumer products, not legal minimum standards.
Source: B.H. Webb and A.H. Johnson, *Fundamentals of Dairy Chemistry*, 1965, and *Composition of Foods*, USDA Handbook No. 8, 1963.

Table 4: Composition of Some Common Cheeses

	moisture (percent)	fat in solids (percent)	fat (percent)	protein (percent)	calcium (mg/100 g)	ash (percent)
Parmesan						
Typical	30.0	37.1	26.0	36.0	1,140	5.1
Federal standard*	<32.0†	>32.0‡				
Cheddar						
Typical	37.0	51.1	32.2	25.0	750	3.7
Federal standard	<39.0	>50.0				
Pasteurized process American						
Typical	40.0	50.0	30.0	23.2	697	4.9
Federal standard	<39.0	>50.0				
U.S. Swiss						
Typical	39.0	45.9	28.0	27.5	925	3.8
Federal standard	<41.0	>43.0				
Roquefort§						
Typical	40.0	50.8	30.5	21.5	315	6.0
Federal standard	<45.0	>50.0				
Limburger						
Typical	45.0	50.9	28.0	21.2	590	3.6
Federal standard	<50.0	>50.0				
Camembert						
Typical	52.2	51.7	24.7	17.5	105	3.8
Federal standard		>50.0				
Cream						
Typical	51.0	76.9	37.7	8.0	62	1.2
Federal standard	<55.0		33.0			
Cottage, creamed						
Typical	78.3	19.3	4.2	13.6	94	1.0
Federal standard	<80.0		4.0			

*Listed in approximately the order of their decreasing hardness. †< Signifies "not more than." ‡> Signifies "not less than." §Blue cheese is identical in composition.
Source: R.W. Bell and E.O. Whittier in B.H. Webb and A.H. Johnson, *Fundamentals of Dairy Chemistry*, 1965.

Many factors influence the composition of milk, including breed, the genetic constitution of the individual cow, and the interval between milkings. Since the last milk to be drawn at each milking is richer in fat than the rest, the completeness of milking influences the composition of the sample. The age of the cow, the stage of lactation, and certain disease conditions are among other factors affecting composition. In general, the kind of feed has slight effect on the composition of milk, but feed of poor quality and insufficient quantity causes both a low yield and a low percentage of nonfat solids.

The compositions of milk, creams, and various milk products are shown in Table 3. Those that contain high moisture and that are not artificially preserved by addition of sugar or by sterilization require refrigeration during storage and distribution. Most of these are products of the fluid milk industry, as distinguished from the concentrated and dry milks, butter, and ice cream. (The composition of cheeses is discussed later and shown in Table 4).

Milk is a good source of many of the vitamins, as shown in Table 5; however, only one milligram of vitamin C or ascorbic acid is present in each kilogram of milk and this is easily destroyed by heating. Vitamin D is formed in milk fat by ultraviolet irradiation, and beverage milk is now commonly fortified by additional vitamin D. Both vitamins A and D are fat soluble and are often added to skim milks to improve nutritive value.

Properties of milk. The properties of milk are important in controlling its behaviour during manufacture into various products.

Colour and flavour

The white colour of milk is caused by the fine dispersion of calcium caseinate, which is hydrated and permanently suspended in the milk. The carotenoids, largely as alpha and beta carotene (related to vitamin A) impart a natural yellow colour to the milk fat, in which they are soluble. A greenish-yellow colour in milk, noticeable particularly in whey, is caused by the presence of vitamin B_2 or riboflavin. The quantities of these pigments in milk are given in Table 5.

The flavour of milk is mild and bland unless it has been affected by the cow's consumption of strong flavour-producing feed such as wild garlic. Pasteurization changes flavour only slightly, but sterilization, as in the making of evaporated milk, imparts a definite cooked or heated flavour. This is caused by the liberation of volatile sulfides during heat treatment. Newer methods of sterilization using rapid ultra-high temperatures up to 300° F (149° C), attained in two or three seconds and held for one or two seconds, generate fewer sulfhydryls and less cooked flavour than older sterilization processes, with temperatures of 242° F (117° C) held for 15 minutes.

Other undesirable flavours can be developed in or absorbed by milk. Growth of bacteria and protein-splitting organisms can produce obnoxious flavours. Chemical changes take place when milk fat is oxidized or when a

Table 5: Vitamin Content of Some Milk Products

product	A carotene (IU/100 g)	B_1 thiamine (mg/kg)	B_2 riboflavin (mg/kg)	nicotinic acid (mg/kg)	B_6 pyridoxine (mg/kg)	pantothenic (mg/kg)	biotin (mg/kg)	B_{12} cyanocobalamin (mg/kg)
Milk	156	0.44	1.75	0.94	0.64	3.46	.031	.0043
Table cream	880	0.3	1.4	0.4	0.40	—	—	—
Butter	3,108	0.03	0.16	0.5	0.40	2.3	—	—
Skim milk	9	0.4	1.7	0.86	0.45	3.6	.016	.0038
Evaporated milk	369	0.56	3.8	2.0	0.74	7.0	.056	.0014
Ice cream	523	0.48	2.3	1.1	—	—	—	—
Cottage cheese	291	0.26	3.3	0.92	0.54	2.2	.020	.0085
Cheddar cheese	1,169	0.30	5.0	0.49	0.75	2.7	.022	.013
Whey	11	0.4	1.2	0.85	0.42	3.4	.014	.0020

Source: A.M. Hartman and L.P. Dryden in B.H. Webb and A.H. Johnson, *Fundamentals of Dairy Chemistry*, 1965.

product becomes stale. These changes are permanent and cannot be removed by further processing, but their development can be retarded by refrigerated storage, and oxidation of milk fat can be inhibited by absence of oxygen. Certain flavours imparted to milk by the cow's feed can be removed by heating the milk and passing it through a vacuum chamber.

Producing desirable flavours

Desirable flavours such as the clean lactic flavour of cultured buttermilk or of yogurt are developed by controlled fermentation of pure cultures. The characteristic flavours of various cheeses are produced by careful microbiological cultivation of the proper flora for each cheese variety. Many fruit flavours are produced in milk products by addition of fruit purees or juices.

Enzymes in milk. Milk contains many enzymes, and others are produced in milk as a result of bacterial growth. Enzymes are biologic catalysts elaborated by a living cell, in the case of milk the mammary tissue. Enzyme action in milk systems is extremely important in its effect on flavour and body of the different milk products. Lipases and other fat-splitting enzymes, oxidation catalysts, and protein-splitting proteases and starch-splitting amylases are among the more important enzymes naturally occurring in milk. These classes of enzymes and others are produced in milk by microbiological action. The proteolytic enzyme rennin, obtained from calves' stomachs, is used to coagulate milk for cheese manufacture.

Milk fat. Fat exists in milk as an emulsion in a water phase, that is, as finely dispersed globules that are stabilized by a milk protein membrane that permits clumping and gravity rise of the fat clumps. This is called creaming and it is expected in all pasteurized milk sold in bottles. In the United States, when paper cartons supplanted glass bottles, consumers stopped the practice of skimming cream from the top; processors then introduced homogenization by forcing the milk through a very small opening under pressure. This reduced the fat globules to a tenth their original size, and prevented their rapid gravity separation. Homogenization is used in many dairy processes to improve physical properties of products.

Coagulation. Coagulation of milk is an irreversible change of the milk protein from a soluble or disperse state to an agglomerated or coagulated condition. Its appearance is commonly associated with spoilage, but coagulation is a necessary step in some processing procedures. Milk may be coagulated by several agents: heat, acid, alcohol, rennet, metallic salts, certain gums, and other precipitants. Milk that naturally sours is coagulated by the lactic acid formed by the lactose-fermenting bacteria that it usually contains. When milk is pasteurized and held refrigerated for two or three weeks it will generally be spoiled if not coagulated by psychrophilic, proteolytic organisms.

Milk varies in its resistance to heat coagulation and may require many hours at 240° F (116° C) before coagulating. If the milk is concentrated, as in evaporated milk manufacture, it usually coagulates in 15 to 30 minutes at this temperature. Resistance of the concentrate against coagulation can be increased by prewarming the raw milk to a boiling temperature before concentration or by adding a minute quantity of a stabilizing salt before sterilization. Sodium phosphate is generally used for stabilization of evaporated milk.

Acidity and coagulation

Development of acid in milk by bacterial growth or addition of acid causes it to coagulate quickly during heating. Coagulation by acid is an important aid in production of special forms of milk for distribution under refrigeration in fluid markets. Cultured buttermilk is soured by controlled growth of bacteria to produce desired flavour and thick body. Yogurt is made by growing acid-forming yogurt organisms until a gellike structure is attained. Bulgaricus, koumiss, and kefir cultures make the coagulated milks named for them. Milks designed to produce a soft curd in the stomach, especially for children, can be made by treatment of the milk by a proteolytic enzyme. Some cows naturally produce soft curd milk, and the homogenization of milk will give it mild soft curd characteristics.

Processing milk. *Pasteurization.* Steps in the pasteurization of milk or cream are shown in the flow chart in Figure 2. All approved pasteurization systems are equipped with a flow diversion valve to return improperly heated milk to the raw milk side of the system in case of malfunction. The liquid-vapour separator is under partial vacuum and deodorizes the milk by removing volatile off-flavours. The bottling or packaging machine fills and seals the milk in retail containers.

Milk pasteurizers were originally of the vat type that held the milk at 145° F (63° C) for 30 minutes. Modern

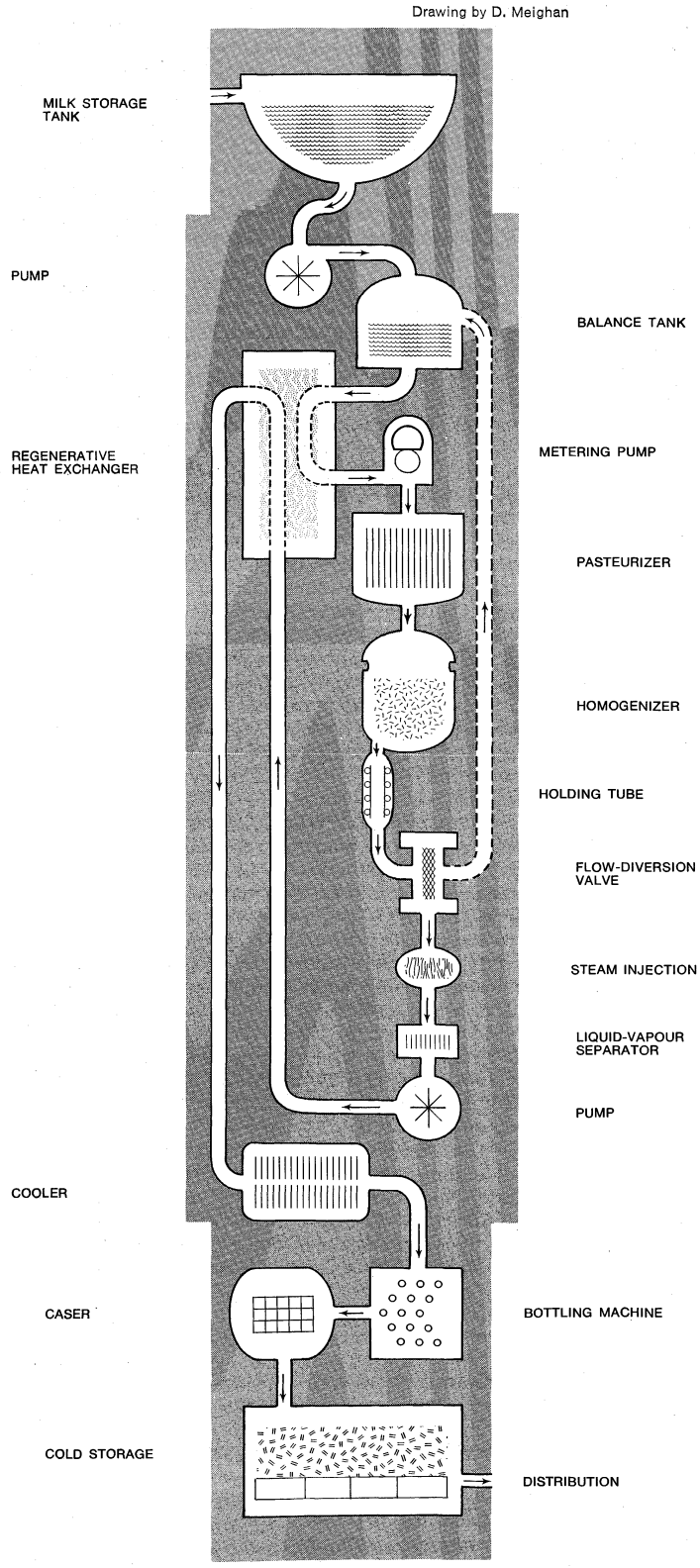


Figure 2: Milk pasteurization process.

pasteurizing equipment is designed for high temperature, short time, continuous operation between 162° F (72° C) for 16 seconds and up to 185° F (85° C) for no holding time. Pasteurizers are of the regenerative type, the hot pasteurized milk being cooled by the incoming raw product. Cream, skim milk, and other liquid products are pasteurized in the same equipment.

Separation of cream. The cream separator produces cream and skim milk, from which many dairy products are made. This machine, a high-powered centrifuge, is the successor of the gravity creaming pan. When subjected to the centrifugal force the milk-fat globules leave the plasma phase of the milk and emerge as cream from the separator bowl. Optimum separation temperature is 104° F (40° C). Efficiency is greatest when the fat globules are large, those under one micron in diameter usually remaining in the skim milk, which should contain less than 0.02 to 0.05 percent fat.

Clarification of milk. Milk clarifiers are centrifugal machines that remove extraneous matter from milk. They are built much like separators, but all the milk flows out of a single outlet, foreign material remaining in the bowl.

Quality of fresh milk products. Sanitary production and regulatory control of fresh dairy products are usually under the direction of municipal or national agencies. Inspectors regularly visit farms, plants, and distributing centres to ensure that the products are safe and nutritious.

Conditions of storage

Fresh milk, buttermilk, yogurt, cottage cheese, and allied products are very sensitive to conditions of handling and storage. Their storage life at 45° F (7° C) is usually considered to be about 15 days, but recent studies have shown that milk held at 33° F (1° C) has a storage life double that time. Thus the best practice is to hold dairy products at a temperature just above freezing.

Less rigid but safe requirements regulate the production of manufacturing grade milk, which is used in making all products except those in the fluid milk class.

Fat content may be considered an aspect of quality and, in most countries in which milk is an article of commerce, standards of composition of milk for retail sale are established by law or local regulations. The minimal standard for fat is usually some value between 3.0 and 3.5 percent, that for nonfat solids usually 8.5 percent.

Packaging fluid milk products. Milk containers in the U.S. are largely paper cartons, automatically filled and sealed by continuous machines. In countries where glass bottles are used, each bottle makes 20 to 40 trips. Returnable glass bottles are washed in automatic bottle-washing machines.

Sweetened condensed and evaporated milk. Sweetened condensed and evaporated milks are old forms of concentrated milks prepared and canned for consumer use. Their compositions are tabulated in Table 3. Sweetened condensed milk is a favourite product in many tropical countries, where it is made in modern recombining plants from dry skim milk and butter oil shipped in from dairy countries. Modified processes have been developed for the powder oil mixtures.

Sweetened condensed milk is usually made from fresh milk by adding sugar to the milk, prewarming and concentrating the mixture under high vacuum, cooling the sirupy milk so that the lactose crystallizes as very fine crystals, then canning the product. Although sweetened condensed milk thickens to a gel after long storage, it has excellent keeping qualities, being preserved by the sugar even after opening. The canned product is a consumer item used often where milk is scarce, refrigeration lacking, and temperatures such that unsweetened milk does not keep well.

Condensed milks without added sugar and without sterilization are made by pasteurizing fresh whole or skim milk and concentrating it under vacuum to about 30 to 40 percent solids. The products are cooled and held refrigerated until sold. Candy manufacturers, bakers, and ice cream processors are large users. The condensed products often are tailored to the specifications of the purchaser.

Production of evaporated milk involves several important steps. The raw milk is warmed to 203° F (95° C) for ten minutes to give its concentrate the stability needed to withstand sterilization in cans without coagulating. The milk is concentrated to a 2:1 solids content in a multiple-stage evaporator. The composition of evaporated milk is given in Table 3. The concentrated milk is homogenized, standardized, canned, and sterilized at 240° F (116° C) for 16 minutes in a high-capacity continuous sterilizer. Before sterilization the heat stability of the concentrate is adjusted with a stabilizing salt so that the body of the finished product is smooth and creamy. Improper stabilization can result in an unacceptable coagulated product.

New ultrahigh-temperature sterilization methods can be used to produce evaporated milk with only slightly more cooked flavour than pasteurized milk. In storage of such milk, flavour and body changes occur that reduce its consumer acceptance. More research is needed to improve the product.

Evaporated milk is usually fortified during manufacture with the addition of vitamin D; evaporated skim milk is fortified with vitamins A and D.

ICE CREAM AND OTHER FROZEN DAIRY PRODUCTS

Commercial production of ice cream originally grew out of the discovery that a mixture of ice and salt could produce lower temperatures than salt alone. In the late 19th century, development of mechanical refrigeration laid the basis for the modern ice cream industry.

Ice cream products are prepared for consumption in the frozen state, but many other dairy foods are frozen as a means of preservation. Frozen milk, frozen cream, frozen concentrated milk, and frozen concentrated skim milk are especially prepared for preservation by freezing and holding for future wholesale and retail use. Frozen cream and concentrated milks act as an industry balance wheel, since they are held from seasons of high production and used in seasons of scarcity.

Composition and properties. The compositions of ice cream, ice milk, and sherbet are shown in Table 3 and the quantities of the important vitamins in ice cream are given in Table 5. Fruit sherbet contains only 1 to 2 percent fat and 2 to 5 percent milk solids. All contain a maximum of 0.5 percent stabilizer and 0.2 percent emulsifier. The blend of milk fat and nonfat solids with sugar must result in a product of pleasing taste and one which is smooth and creamy. Composition is important but the most critical stage of ice cream manufacture is the mechanical blending, freezing and hardening of the frozen dessert. Even water ices on a stick, although quiescently frozen, must be frozen rapidly to prevent coarseness of body.

Imitation ice creams known as mellorine are made in various parts of the world where economic conditions favour them. Mellorine is cheaper than ice cream because inexpensive vegetable fats and oils are substituted for milk fat. Other than this change mellorine has approximately the same composition as ice cream. There is still no satisfactory, cheap substitute for milk protein, although some vegetable proteins, particularly soy, have been prepared with improved flavour in recent years.

Use of vegetable fats and oils

Manufacture of ice cream. Ice cream is a complex system in which the stable mixed emulsion of a four phase system, fat–water–ice–air, must be balanced and protected from breaking or separating. The manufacture of ice cream starts with making the mix. This is composed of a combination of suitable dairy products such as cream, milk, or skim milk either concentrated or dry. At least two mixing tanks and three aging tanks are needed to feed mix continuously to the freezer. The complete mix must be homogenized, pasteurized, and aged for at least four hours to condition it for freezing. The ice cream is extruded from the freezer into packages that are conveyed to the hardening and storage rooms.

The mix is compounded and frozen in such a way that the final product is smooth, homogeneous, and free of coarse ice crystals. Emulsifiers and stabilizers are used to finely distribute the milk fat, the ice crystals, and the air

bubbles of the frozen product. Sugar in the mix is essential not only to give sweetness but also to lower the freezing point. As the water freezes the syrup becomes more concentrated, and this finally creates an unfrozen syrupy vehicle to give added smoothness and palatibility.

Incorporation of air in a mix during freezing (overrun) can double the volume. Hand freezers beat air into the mix as the water froze, but commercial freezers meter the air in under controlled conditions. The air must be incorporated and distributed as very fine bubbles.

Freezing is done with great rapidity and under severe agitation. Formation of the maximum amount of ice during freezer agitation produces smoothness. As ice forms the mix temperature is lowered until the heavy or molten product is discharged.

The common drawing temperature for batch freezers is 24° F (−5° C), when about 38 percent of the water in the mix is frozen. In the newer continuous freezers 54 percent of the water is frozen and the ice cream is drawn at 21° F (−6° C).

Ice cream quality. Quality implies a cleanly produced product of acceptable flavour, body, and texture. Defects that develop in ice cream have been a challenge to manufacturer and distributor. One scoring system for ice cream reflects the problem areas. Of 100 points, 45 are given for flavour, 30 for body and texture, 15 for manufacturing facility, 5 for melting qualities, and 5 for packaging and colour. Flavour acceptability is governed by the quality of the flavouring ingredient used; *e.g.*, fruit, chocolate, or nuts. The basic flavour must come from high quality milk and cream. Body and texture are affected by physical characteristics such as freezing and storage conditions. Two principal defects are sandiness, caused by formation of noticeably large lactose crystals, and shrinkage of the product away from the sides of the package. Both of these develop rather slowly in hardening rooms or cabinets. An icy texture may result from partial melting and refreezing.

Scoring system for ice cream

Grades and standards for ice cream permit it to be made to suit different tastes, weather conditions, and selling prices. Richness, reflected in fat content, affects both flavour and price. The usual fat content of ice cream is about 10 percent, but richer and more expensive products, such as French ice cream, may contain 12 or even 14 percent butterfat.

Ice milk contains between 2 and 7 percent fat. But with the use of stabilizers and rapid low-temperature freezing techniques ice milk can be made almost as smooth and creamy as ice cream.

Frozen dairy products. Milk and cream are sometimes frozen as a means of preserving them. The freezing point of milk is quite constant at 31.05° to 30.98° F (−0.53° to −0.58° C). A check of freezing point is a reliable method for detecting addition of water to milk. Freezing tends to destroy the fat emulsion of whole milk so that on thawing an oily fat layer forms. Homogenization before freezing almost entirely eliminates fat separation except in creams containing more than 30 percent fat. Very rapid freezing, and increase in nonfat solids, or addition of sugar also retards the amount of fat that "oils off" after thawing. Cartons of homogenized milk (not glass containers) may be quickly frozen for later use without adverse effects. Cream is frozen commercially for future use.

Fresh concentrated milk may be frozen without emulsion damage, but prolonged storage may cause the casein gradually to become insoluble.

Freezing has no observable effect on the characteristics or quality of butter and it is commonly held in cold storage at −4° F (−20° C). The freezing point of unsalted butter is 32° F (0° C) and that of salted butter −4° F (−20° C).

Because of its high sugar and lactose content sweetened condensed milk is not damaged by freezing temperatures. Its freezing point is in the vicinity of 5° F (−15° C).

The freezing points of cheeses vary from about 30° to 3° F (−1° to −16° C), depending upon moisture content. When cheese undergoes extensive freezing the body and texture become more crumbly and mealy after thawing. High-moisture cheeses such as cottage, Neufchâtel, and cream are usually seriously damaged by freezing.

BUTTER AND BUTTERFAT

From a quarter to a third of the world's milk production is used to make butter. Butter is a concentrate of butterfat or milk fat. Creams of various fat contents are intermediate products in butter manufacture. When the cream emulsion is broken, free milk fat is released and this is called butter oil or, if entirely dehydrated, anhydrous milk fat. Plastic cream is a milk fat concentrate containing up to 80 percent fat, the same fat content as butter, but with the emulsion unbroken. Plastic cream is produced by centrifugally concentrating milk fat, while butter requires not only concentration but also breaking of the fat emulsion. Various kinds of butter-like spreads containing some milk fat have been made but do not have wide recognition as standard products.

Composition. The composition of butter, butter oil, and the creams from which they are prepared is given in Table 3. Most butter contains at least 80 percent fat, not more than 16 percent water, about 2 percent added salt, and 1 percent milk curd. It is a stable mixed emulsion of fat and water but stability is lost and it oils off when the fat is melted. The fat consists largely of mixed triglycerides of fatty acids and its composition tends to vary with many factors.

Vitamins E and A and carotenoids are present in microgram quantities and these are responsible for the natural golden colour of milk fat. Vegetable colour is sometimes added in commercial production.

Physical properties. The physical properties of butter and the high-fat creams and spreads stem from the unusual characteristics of milk fat as it occurs in milk. Manufacturing processes are built around the physical behaviour of the individual fat globule. Most of the globules are smaller than 4 microns (1,000 microns equal one millimetre) and they seldom exceed ten microns in diameter. The globules are covered by a protective membrane that must be disrupted to obtain butter or butter oil from cream. When the globules are destabilized by churning and, at high-fat concentrations, by homogenization, the membrane is disrupted. The rising of fat globules in milk and the formation of a cream layer represent basic properties of the fat emulsion. During rising the smaller globules clump together and the aggregates rise rapidly but without disturbance to the membrane. The centrifugal separator hastens this process by producing a plasma and a cream phase rapidly and much more efficiently than older gravity systems. The separator bowl is filled with a series of disks to channel the skim milk phase to the outside of the bowl while the lighter cream phase moves toward the centre. Separator efficiency depends upon fat content and size of the globules and the temperature of the milk, which should be 104° F (40° C).

Characteristics of fat globules

In churning, about half the fat globule membrane material is liberated into the buttermilk. During homogenization there is a fourfold to sixfold increase in new fat surface and at high-fat levels, above about 50 percent fat, there is insufficient membrane material to cover the new surface. The system becomes destabilized and may oil off. Complete destabilization occurs in creams of 70–80 percent fat, and homogenization is used under such conditions to break the emulsion as the first step in the continuous churning process.

The hardness of the butter is affected by the physical state of the fat. When secreted by the cow in milk the fat is in a liquid state. It is partly solidified for churning by a temperature adjustment and aging at about 52° F (11° C). Temperature manipulation of the cream before churning affects the consistency of the butter. If milk fat or butter is melted and gradually cooled a large number of crystalline fractions can be obtained. Melted fat cooled to room temperature (70° F or 21° C) contains about half solid and half liquid fat. A variability in hardening temperature may be used to produce a more spreadable or an excessively hard butter at refrigeration temperature. The change in hardness is subject to further change by subsequent shifts in storage temperature.

Production of whipped butter

Whipped butter has been developed to furnish a soft, easily spread butter at refrigeration temperatures. Air or nitrogen gas is whipped into softened butter by giant whippers, the product is extruded into consumer-size cartons, tubs or serving dishes and hardened in cold storage. Volume is increased about 50 percent.

Manufacture of butter. Butter was first made by separating cream from milk by gravity, and then subjecting it to mechanical agitation. Invention of the cream separator made it possible to gather large amounts of cream in one place and this moved buttermaking from the home to the factory. Here the cream is churned and worked in wooden or metal churns (see Figure 3). Such butters are relatively uniform in appearance and physical characteristics. Continuous buttermaking, introduced after World War II, achieves increased manufacturing efficiencies and plant output. The early continuous processes tended to produce butters that lacked uniformity, but newer modifications of the process are succeeding in overcoming this difficulty.

By courtesy of F.A.O.; photograph, Prabha Art Studio, Bombay

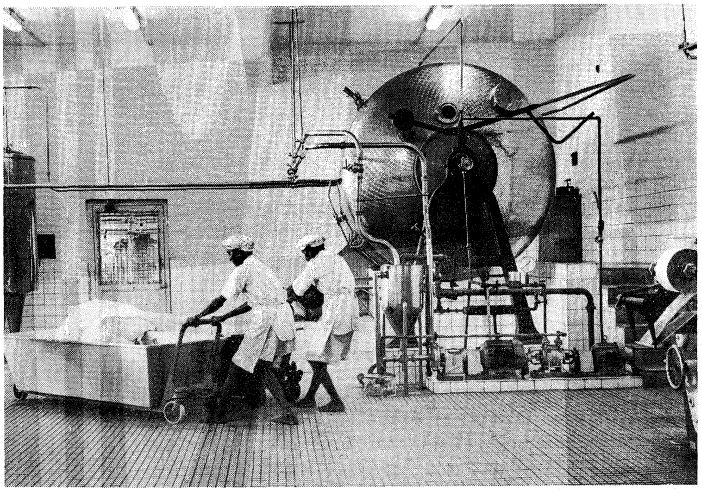


Figure 3: Bulk butter taken from a stainless steel butter churn (India).

Churning by either batch or continuous method must involve a destabilization of the fat emulsion by mechanical agitation. In the batch churning process small air bubbles are incorporated in the cream by agitation. The fat globules gather at the surface of the bubbles (at the fat–plasma interphase) until the foam collapses when butter granules are formed. The buttermilk is drained from the churn and the butter is worked to incorporate the correct amount of moisture. The butter is then chilled for 48 hours before grading.

Continuous buttermaking may be classed in two categories: methods involving the accelerated churning of cream of normal composition, and methods that utilize reseparated high-fat cream. In the first process, cream of about 36–40 percent fat moves in a thin layer through a churning cylinder. The violent agitation in the air-cream layer breaks the emulsion in a matter of minutes, compared with up to one and a half hours in the conventional churn. The air content of the butter and fat losses in the buttermilk have been troublesome but are being corrected. Factory capacity using this system has been increased from 500 to 5,000 kilograms (about 1,100 to 11,000 pounds) per hour.

The second continuous method appears to be less popular at present, but has been used extensively to prepare butter oil. The cream is pasteurized and concentrated from about 35 percent to 80 percent fat; it then goes to an emulsion breaker, usually a homogenizer. The oil is skimmed off to 98 percent fat in a second separator. If anhydrous milk fat is desired the oil goes to a vacuum dehydrator, which removes the last water to yield a 99.9 percent oil. If butter is desired the 98 percent fat product goes to a standardization tank where the required moisture, curd, and colour are added to ready the mixture for the butter worker and chiller from whence it is ready for extruding into packages.

Butter quality. Butter should be uniformly firm, waxy, and of good spreading quality, the granules close knit to cut clean when sliced. The water droplets should be well distributed throughout the mass. Some terms for body defects are "crumbly," "leaky," "sticky," and "weak." Of lesser importance in grading are colour and salt distribution.

The yellow colour of butter reflects its carotene content. The amount varies, winter butter produced on dry feed being lower in carotene than summer butter. Vitamin A is also substantially higher in summer butter. Ninety to 100 percent of the vitamin A in the milk goes into the butter. In the U.S., a harmless yellow vegetable colour, annatto, may be added to butter to improve colour uniformity throughout the year. No other additive is permitted. Vitamin D content of butter is low.

Butter from periods of high production is often held in dry storage for later movement into consumer channels at 0° F (−18° C).

Packaging. A large part of the butter produced at country plants is shipped to primary receivers who "print" and package it for distribution. Printing is the term used for cutting and wrapping the butter in consumer units. Butter is usually sold at retail wrapped in tight, moisture-proof packages.

CHEESE

An ancient legend attributes the invention of cheese to an Arabian merchant who filled his pouch made of sheep's stomach with milk and, after travelling all day, found that the milk had separated into curds and whey. The lining of the pouch contained rennet which, with the sun's warmth, caused the milk to coagulate and the whey to separate just as it does in the modern cheese vat. Some such accident may indeed have occurred; in any case, the art of cheesemaking spread from western Asia to Europe in ancient times. Special varieties developed in isolated communities in the Middle Ages, such as Gorgonzola in the Po Valley in Italy and Roquefort in the Roquefort caves of France. In the mid-19th century cheesemaking changed from a farm product to a factory industry.

Composition and properties. Cheese is a complex product and its composition is determined by a number of interrelated factors. The percentage of fat in the milk and the method of manufacture are most important. Moisture content, related to hardness, strongly influences keeping quality. Compositions of some common cheeses are shown in Table 4.

Part of the calcium and phosphorus of the milk is retained in the cheese. Essentially all of the fat and casein of milk goes to the cheese while the whey proteins (globulins and albumins), lactose, and soluble salts remain in the whey.

Varieties of cheese

There are several hundred varieties of cheese, all with somewhat different composition and properties. Most were developed gradually over a long period as a result of accidental or intentional modifications of the cheesemaking process. Today, as cheesemaking passes from an art to a science with exact chemical and microbiological control, there is better standardization and uniformity of composition.

Classification of cheese. The cheese-manufacturing process is designed to produce cheese in one of four classification groups. Group 1 is very hard grating cheese ripened by bacteria, such as Parmesan, Romano, and sapsago. Group 2 is hard cheese ripened by bacteria but without eyes, as Cheddar, or with eyes, as Swiss and Gruyère. Group 3 cheeses are semisoft and of three kinds: (1) ripened by bacteria, as brick and Muenster; (2) ripened by bacteria and surface micro-organisms, as Limburger and Port du Salut; (3) ripened by mold in the interior, as Roquefort, blue, Gorgonzola, and Stilton. Group 4 is the soft cheese of two kinds, ripened, as Bel Paese, Camembert, and Neufchâtel; and unripened, as cottage, pot, cream, mysost, primost, and fresh ricotta. Examples of the manufacturing processes for one cheese in each class are given below.

Manufacture of cheese. Cheese is made by coagulating milk, cutting and heating the curd to express the whey, then pressing and ripening the cheese. The process is a microbiological one in which enzymes produced by bacteria develop the desired body and flavour. The initial coagulation may be brought about by addition of rennet, by addition of bacterial starter that develops acid by fermentation of the lactose in the milk, by direct addition of acid, or by a combination of these. To hasten curd formation the milk is held quiescent and warm for a number of hours. When congealed the curd is cut and cooked by mildly heating it to release the whey. The warm curd, containing the bacterial flora needed to produce the desired cheese variety, is pressed in suitable molds, hoops, or boxes. Some varieties (*e.g.*, cottage) are sold in the fresh state. Ripened cheeses must be held under critical temperature and moisture conditions to permit bacterial enzyme action, which creates the kind of cheese desired.

Curds and whey

Parmesan cheese. Parmesan was first made in the vicinity of Parma, Italy. Milk or partly skimmed milk is warmed to 90°–98° F (32°–37° C) in copper kettles, a heat-resistant lactobacillus culture, plus rennet, is added to produce a firm curd in 20 to 30 minutes. The curd is placed in a hoop 10 inches (25 centimetres) deep and 18 inches (46 centimetres) in diameter and pressed for 20 hours. It is taken to a salting room and held at 62° F (17° C) for three days. The cheese is then removed from the hoop and held in brine for about 15 days, after which it is dried for 10 days. The first stage of curing consists of holding the cheese at 60° F (16° C) and 80–85 percent humidity for one year. The second stage is usually in the dealer's curing room at 54°–60° F (12°–16° C) and 90 percent humidity. It can now be sold, but it will keep almost indefinitely. Its most important use is for grating.

Cheddar cheese. Cheddar cheese was first made in the 16th century in the village of Cheddar in England. In the U.S., where almost 70 percent of the cheese made is Cheddar (often called American cheese or American Cheddar), the term "Cheddar" is used to describe both a type of cheese and a step in the manufacturing process. The most common form of the cheese is 14½ inches in diameter by 12 inches thick (37 by 30 centimetres) and weighs 75 pounds (34 kilograms).

A hard, white-to-yellow coloured cheese, Cheddar is usually made from pasteurized milk (see Figure 4). Starter, rennet, and colour are added to the milk, which sets to a firm curd in about 30 minutes. The curd is cut into small cubes and stirred continuously with warming to 100° F (38° C). About 2½ hours after the rennet is added the whey is drained and the cubes of curd are piled along the sides of the vat, where they fuse together. When the curd is firm enough to be turned without breaking, it is "cheddared" or matted; that is, cut into slabs to be turned and piled into layers. The slabs of curd are cut in a curd mill, salted, drained, and placed in cloth-lined metal hoops for pressing. After pressing the cheeses are dressed and dried for three to four days at about 55° F (13° C) before dipping them into wax. Rindless cheese can be made by wrapping it in heat-sealing plastic film. Cheddar cheeses are usually cured at about 45° F (7° C) for several months, or sometimes as long as a year.

By courtesy of the New Zealand Consulate General, New York

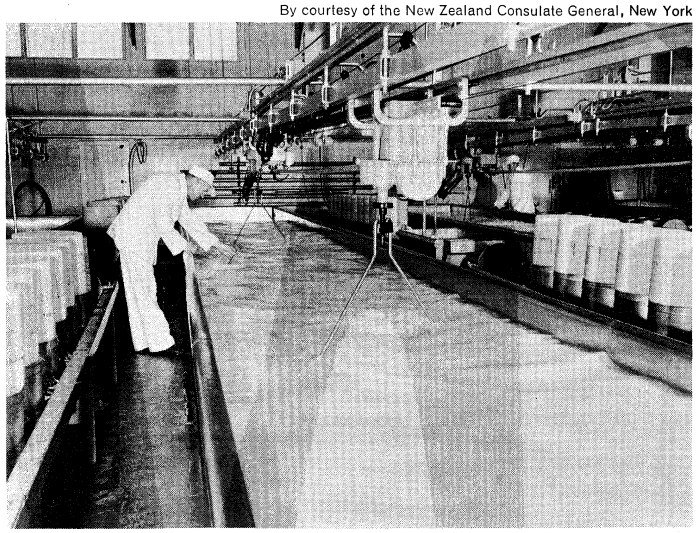


Figure 4: Milk in a 2,600-gallon vat being cooked and agitated during a cheese-making process (New Zealand).

Much effort in the U.S., Europe, and Australia has been directed toward automation of the Cheddar-making process, but it has proved difficult to mechanize the various steps without altering the sequence of chemical and microbiological changes that must occur to produce an acceptable product. Every step of the process is critical. All must be accomplished at the proper time, temperature, and acidity. The four basic steps suitable for mechanization are: (1) coagulating the milk or forming the curd; (2) draining the whey and cheddaring the curd; (3) cutting and salting; (4) pressing and handling for ripening. Two types of equipment for steps 2 and 3 have been developed in Australia.

Roquefort or blue-veined cheese. The name Roquefort is limited by French regulation to cheese made in the Roquefort area from ewe's milk; much more common is a similar blue-veined cheese made in many countries from cow's or goat's milk and often called blue cheese. Blue cheese is usually made from cow's milk by setting and cutting the curd, draining the whey, and placing the cheese in hoops. Blue mold powder grown from the mold *Penicillium roqueforti* is mixed with the curd. After 24 hours, the cheeses are removed from the hoops and dry salted over a seven to ten day period at a temperature of 48° F (9° C) and 95 percent relative humidity. A week after salting, each cheese is pierced with 40 small holes to permit air to reach the interior, air being essential for mold growth. The cheese is cured for three months and scraped and cleaned every three weeks during this time. The natural caves at Roquefort provide ideal curing conditions, 48° F (9° C) and 95 percent humidity. Suitable caves exist in other countries and artificial conditions have been successful.

Cottage cheese. Cottage is sometimes called pot or Dutch cheese, or schmierkase. It is soft, uncured, high in moisture, and perishable. Cottage cheese is made from pasteurized skim milk to which rennet and bacterial starter are added to coagulate the milk and produce flavour. The curd is cut when firm and whey is expelled from it by heat. The whey is drained from the firmed curd particles, which are then salted, creamed, and packed in 50-pound (23-kilogram) tubs or consumer-size cartons.

Process cheese. Process cheese is made by grinding fine and mixing together by heating and stirring one or more cheeses of the same kind, or two or more varieties. Soft cheeses such as cottage, cream, or skim-milk cheeses are not used. Vinegar, other organic acids, colour, spices, and flavouring may be added. Emulsifying agents help to reduce the mixture to a homogeneous mass which is then stirred and cooked at about 155° F (68° C). When ready to be packaged it is extruded into cartons lined with a transparent film that acts as a sealer to exclude air. The packaged cheese is then cooled and held under refrigeration. As made and packaged, process cheese is practically sterile and does not ripen further. Fruits, vegetables, or meats, or mixtures of these, may be added to process cheese.

Quality of cheese. Cheese is a product of fermentation except for certain fresh curd varieties made by direct addition of acid to coagulate the milk. This is a new procedure and is not yet acceptable in many areas. Fermentations constantly undergo changes that affect cheese flavour, body, texture, and colour. These are influenced by the quality of the milk, the techniques of manufacture, and the temperature and time of curing. To manufacture cheese of uniformly good quality, it is essential to use good quality milk, sanitary and adequate equipment, uniformly active starter, a standardized and proven manufacturing procedure, and controlled temperature and length of curing time.

Microbiological control to suppress unwanted bacterial

and mold growth during cheese ripening is important in production of any specific variety. The making and ripening processes have been developed to encourage favourable chemical and physical changes of fermentation, proteolysis, and slight fat breakdown, and to discourage growth of organisms that produce objectionable characteristics.

Microbiological control

DRIED MILKS

Skim milk, whole milk, and certain other milk products are dried to reduce weight and shipping charges, to provide a means of handling surpluses, to prolong keeping quality, and to reduce the products to a more useful form. Besides skim milk and whole milk, dairy products or by-products that are dried include buttermilk, malted milk, sweet or sour cream, ice cream mix, whey, coffee creamers, cheese mixtures, chocolate milks, skim milk-vegetable fat mixtures, and sodium or calcium caseinates. Lactose is always prepared in dry form. Skim milk, whole milk, malted milk, and buttermilk are major dry products; the others may be considered as by-products for which there are many special uses.

Composition and properties. Composition of dry whole, dry skim, dry malted milk, and dry cream are shown in Table 3. Composition, especially fat content, may be varied over a wide range to suit special uses in food manufacture. The content of vitamin A and the water-soluble B vitamins are not seriously changed by processing into dry form. The objective of drying is to produce a dry product without seriously impairing the solubility, colour, or flavour of the original material. This can be done to a remarkable degree and a freshly dried product when remixed with water cannot usually be distinguished from its fresh pasteurized counterpart.

Manufacture of dried milks. Two general types of milk dryers are used today. The simplest and cheapest is the drum or roller dryer. A common form consists of two large metal drums that turn toward each other and are heated from the inside by steam. The concentrated liquid is applied to the hot drum in a thin sheet which is dried during one revolution of the drum then scraped off by a steel blade. The process produces a flake-like powder particle that is easily dispersed in water but that dissolves poorly. It is used in food manufacture or for animal feed.

Spray drying

The spray dryer is widely used to dry liquid materials by spraying them in finely atomized form into a stream of hot air. The large droplet surface causes rapid evaporation of moisture and provides a highly soluble powder. The drying chamber may be conical, rectangular, or silo-shaped. The dry powder passes from this chamber through a series of cyclone collectors for separation from the drying air. About three pounds of steam are required to evaporate each pound of water. The product to be dried is heated, condensed to 45 to 50 percent solids, sprayed into the dryer, and collected as powder. The process may include foam-spray drying. This is a new modification and involves injection of gas, usually air, into the high-pressure line before the spray. The gas expands the droplet and causes it to dry easily into a light low-density particle.

Spray-dried milk is difficult to mix with water, and to improve its dispersibility a process called agglomeration or instantizing has been developed. This consists of quickly moistening the fine powder particles and causing them to stick together. The instantizing process produces only limited improvement in dispersibility of fat-containing dairy products, and it is not used for dry whole milk. Most powders can be made to dissolve readily when they contain sugar, particularly if they are dry-blended with sugar.

Buttermilk is either spray- or roller-dried. Malted milk consists of approximately half milk solids and half solids from a mash prepared from barley malt and wheat flour. It is usually made by evaporating the mixture to dryness under a vacuum.

Whey is roller- or spray-dried, but, because of its high lactose content, it presents special drying problems. The milk sugar is best crystallized after concentration and before drying since its presence in the dry powder in a non-crystalline condition is conducive to moisture pickup and caking of the powder.

Quality of dried milks. Unusual sanitary precautions are taken to avoid contamination of the powder with salmonella organisms. All skim-milk powders are examined and must be free of these organisms. Whey is a perfect growth medium as it is drained warm from the cheese curd and for food use must be pasteurized and cooled at once, after which it can be held for condensing and drying.

Packaging, marketing, and distribution. Dried milks must be packaged to preserve flavour and physical characteristics. When packaged and distributed under favourable conditions dried whole milk and buttermilk will keep six months, while the skim milks and malted milks may be held up to 18 months. Stored milk fat in any form is subject to flavour changes caused by oxidation. Dried whole milk is packed in tins in an oxygen-free atmosphere and should be stored in a cool place, preferably under 45° F (7° C). Dried buttermilk contains traces of an unstable fat fraction and is best handled like whole milk. It is used in wholesale food preparation.

Keeping characteristics

The dried milks and wheys readily absorb moisture, and packaging and storage must ensure against water pickup. The products contain about 3 percent moisture as made but readily absorb water to 5 or 6 percent, which causes rapid deterioration of flavour, caking, and lowered solubility of the milk proteins. Instant dried nonfat milk is packed in tight paper boxes or envelopes; it may be quickly dispersed in water with little or no stirring or shaking, and there should be no settling of insoluble protein on refrigerated storage. Once mixed with water the milk must be held under refrigeration.

Malted milk contains a natural antioxidant, and chocolate malted milk, especially, will keep in a dry environment at room temperatures for at least 18 months.

DAIRY BY-PRODUCTS

Skim milk, buttermilk, casein, whey, and whey components including lactose or milk sugar, are by-products of the dairy industry. They have high nutritive value and are used for human and animal food and to a lesser extent for industrial purposes. Most of the price paid for milk is for the fat and thus by-product prices are relatively low, the major items contributing to their cost being transportation, manufacture, distribution, and selling. Early in the 20th century farmers were separating their milk, selling the cream, and feeding the skim milk to farm animals. Today skim milk is valued as an important human food supplement and as a source of casein and lactose, which are prepared from it. Whey is a by-product of cheese and casein manufacture and is an important source of whey protein and lactose. Buttermilk is a by-product of the churning of cream. In the past the cream often was sour, leaving a buttermilk fit only for animal food. Today, most butter is made from sweet cream and the buttermilk has become a valuable food component.

Composition. The composition of skim milk, buttermilk, and whey are shown in Table 3. Extra grade casein contains 95 percent protein, a maximum of 10 percent moisture, 1.5 percent fat, and 2.2 percent ash. Lactose or milk sugar is the carbohydrate obtained from whey, and it should be at least 99 percent pure. Casein and whey protein are sometimes prepared by a single precipitation, and the composition of the coprecipitate may be 83 percent protein, 1 percent lactose, 1.5 percent fat, 10.5 percent ash, and 4.0 percent moisture.

Physical properties. The physical properties of the by-products differ greatly, depending upon their composition. The state of the protein and lactose is of major interest to the food manufacturer and the consumer since this affects the physical properties of the food. The casein of skim milk is uniformly dispersed but it can be flocculated or coagulated by acid, enzymes, or heat. Coagulation is accompanied by an uptake of water, which causes thickening and the building of body or structure in the food. The coagulum is unstable and syneresis or wheying off will occur with time or heating, similar to the whey separation in cheesemaking. Casein is coagulated to re-

Casein of skim milk

move it from the milk. To make it redispersable for use as food it is prepared as sodium caseinate, which will reabsorb water and rebuild a structure. Whey protein differs from casein in that it is acid soluble but very sensitive to heat, coagulating at temperatures below the boiling point of water.

Lactose, which makes up half the solids of skim milk and three fourths of the solids of whey, is much less soluble than most other sugars. At room temperature lactose is saturated in water at a 17 percent concentration. On a scale of sweetness on which sugar is 100, lactose is only about 20. If lactose crystals form in a food such as ice cream or sweetened condensed milk they are kept to a very small size to avoid giving a sandy texture.

Manufacture of by-products. Manufacture of skim milk, whey, and sweet cream buttermilk for food use follows conventional lines of vacuum concentration and drying. Sometimes these products are used in condensed form and the lactose is crystallized with stirring to produce fine crystals. Drying may be either by the spray or roller methods, the spray system producing the higher quality product. Buttermilk as a by-product of buttermaking is not to be confused with cultured buttermilk, which is sold fresh, never in the dried form. Cultured buttermilk, once dried, will not reabsorb water. Dried sourcream buttermilk is suitable only for animal feed.

Processing of whey presents a special problem because small cheese factories cannot afford concentrating and drying equipment, and whey cannot economically be hauled over long distances. Reverse osmosis or membrane filtration to concentrate whey to 25 percent solids, thus reducing hauling costs, is being advanced as a solution. In any processing scheme for whey the extreme perishability of the product must be considered.

Casein is made by precipitating it from skim milk by the addition of an acid, usually hydrochloric. The precipitated casein is washed, dried, and ground. Skim milk can be naturally soured, and the lactic acid thus formed acts as the acidulant.

Lactose is produced from whey by heat coagulating and removing the whey protein, concentrating the clear whey, then crystallizing the lactose and removing it by centrifugal separation. A pure sugar is obtained by redissolving the sugar, clarifying, decolourizing, and concentrating the syrup, then centrifuging it.

Quality of by-products. The world shortage of food and a realization of the high nutritional value of milk components have led to the development of rigid quality standards for the sanitary production of milk by-products. Rigid bacteriological control is necessary since skim milk, buttermilk, and whey all contain 90 percent or more water and are excellent media for bacteria growth. By-products are usually packaged in multiwall paper bags, often with a moisture-tight lining.

DAIRY PRODUCTS IN HUMAN NUTRITION

Milk and dairy products are accepted in many countries as basic foods, some of which persons of all ages should have every day. The precise role of the milk components —protein, fat, lactose or carbohydrate, and minerals—in human nutrition is still incompletely understood. But these essential elements of a balanced diet are combined in the proportions and amounts needed for the growth of infants and children and for the dietary well-being of an adult. Indeed, the milk of each species is a complete food for its young. One pint of summer milk contributes about 90 percent of the calcium, 30 to 40 percent of the riboflavin, 25 to 30 percent of the protein, 10 to 20 percent of the calories and vitamins A and B, and up to 10 percent of the iron and vitamin D needed by an adult. Although milk supplies a higher proportion of the daily needs of a five-year-old child for calories, protein, vitamins A, B_1, and B_2, the contribution of the calcium needs is reduced to about 70 percent because of the higher calcium requirements of a child.

Food values of milk

Milk protein is of high nutritional value, since it contains all the essential amino acids; *i.e.*, those that cannot be synthesized in quantity. Eighty-two percent of milk protein is casein and 18 percent whey protein. The nutritional response to casein or to whey protein is quite uniform, and these proteins, especially casein, are used as a protein reference standard in feeding experiments. Because cow's milk contains proportionally more casein and less lactose than human milk, it is usually recommended that, for babies' formulas, water and lactose be added to cow's milk.

The nutritional value of milk fat is still unclear. It is the most complex of the natural fats, containing at least 142 fatty acids. Its composition varies widely, depending upon such factors as the cow's intake of unsaturated fatty acids and the levels of dietary fat, protein, and roughage eaten. Considerable research has been done on the possible involvement of milk fat in cardiovascular disease in man but no conclusions can yet be drawn as to what effects this or other fats may have on this ailment.

Lactose, or milk sugar, is a product of mammalian metabolism secreted in milk for nourishment of the young. Its exact nutritional function is not known, but its role in a number of metabolic processes has been studied. Lactose is hydrolyzed in the body to glucose and galactose. Glucose is absorbed directly. Galactose is considered a dietary essential because of its occurrence in cerebrosides and mucopolysaccharides. A deficiency of these is thought possibly to lead to diseases of structural and nervous tissue in later life. There is increasing evidence that a lactose intolerance produces mild or even severe digestive disturbances and diarrhea. Symptoms have been produced in individuals, most often of non-Caucasian races, by feeding 50 grams of lactose in water or even in milk or whey. It is not known whether the intolerance is genetic or whether it has been acquired by long omission of milk from the diet. Intolerance appears to be caused by the absence of the enzyme lactase in the intestine.

There are interrelationships between milk minerals and other food nutrients that are still not clear. Many minerals are involved in maintaining the balance of mineral ions in body fluids, in regulating the metabolism of enzymes, in keeping an acid-base balance, and in facilitating membrane transfer of essential compounds. The mineral content of milk includes calcium and phosphorus adequate for normal skeletal development. The dietary essential minerals that occur in major amounts in milk are, in grams per quart: potassium 1.31, calcium 1.18, chlorine 0.97, phosphorus 0.91, sulfur 0.28, and magnesium 0.11.

The nutritionally essential elements that occur in milk in minor amounts, in milligrams per quart, are: zinc 3.6, iron 0.95, copper 0.28, iodine 0.20, fluorine 0.15, manganese 0.019, molybdenum 0.069, cobalt 0.0006. Other minerals in milk occur only in trace amounts.

Milk contains all of the vitamins known to be required by man. The quantities of water soluble vitamins in a quart of milk are shown in Table 6. Milk contains the fat-soluble vitamin A and carotene, its precursor, but the amount varies considerably with the food of the lactating animal. Since green food is the main source of this vita-

Table 6: Water-Soluble Vitamins in Milk

	mg/qt
Choline	123
Inositol	123
Ascorbic acid	15
Pantothenic acid	3.31
Riboflavin	1.49
Niacin	0.80
Vitamin B_6	0.45
Thiamine	0.40
p-Aminobenzoic acid	0.1
	μg/qt
Biotin	33.10
Vitamin B_{12}	5.28
Folic acid	2.17

Source: M.F. Brink in B.H. Webb and E.O. Whittier, *Byproducts from Milk*, 1970.

min in the diet of cows, summer milk usually has more than winter milk. In recent years some commercial producers have fortified milk with the addition of vitamins and sometimes of iron.

DAIRY INDUSTRY ORGANIZATIONS

Types of organizations

The dairy industry is organized in rural and urban areas on a local, national, and international basis. The organizations represent farmers, processors, distributors, suppliers, engineers, educators, industrial and research scientists.

The International Dairy Federation, with headquarters in Brussels, consists of member countries throughout the world. It sponsors an International Dairy Congress at four-year intervals, the 18th such congress having been held in Sydney, Australia, in 1970. The 19th congress is scheduled for New Delhi, India, in 1974. The congresses include reviews of research and industrial progress and exhibits of new dairy industry equipment.

The United Nations Food and Agriculture Organization (FAO), Rome, and United Nations Children's Fund (UNICEF), New York, have made significant contributions to international dairying. Their attention has been directed especially toward increasing the production and utilization of milk in the less developed countries. FAO has sponsored studies in production, processing, and distribution of milk and its products. UNICEF has been the motivating force for establishing a dairy industry in many underdeveloped countries. One of these is India, where large, modern processing plants have been set up to process locally produced milk or to reconstitute milk from donated or purchased milk fat and powder. Domestic milk production is increasing in India, and part of the pasteurized milk is provided free to children in the larger cities through UNICEF auspices.

An outstanding example of a UNICEF-aided activity in India is the Bombay Milk Scheme, which furnishes pasteurized milk for the people of Bombay. Two hundred miles (about 300 kilometres) north of Bombay at Anand is Amul Dairy, a milk-processing cooperative that has developed a daily milk intake of 500,000 litres milked from water buffalo. Under agreement with UNICEF, 200,000 litres of this milk is shipped to Worly Dairy in Bombay for distribution in the city milk stations, the remainder being processed into powder, concentrated milks, and cheese. Amul Dairy operates a feed mill and offers complete veterinary services for its members. A successful example of a complete food-producing unit, Amul Dairy is being copied in other parts of India with assistance from UNICEF, the Indian government, and some of the large milk-producing countries of the world. These organizations have sponsored the Indian Dairy Development Board (Anand) and the Indian Dairy Corporation (Baroda), which derive support from the sale of milk from donated butter oil and milk powder.

The kind of national and local dairy organizations within a country vary widely. They are sponsored by governments, farmer cooperative groups, industrial processors, suppliers, or marketers, and institutional research and development organizations.

The first cooperative artificial breeding association was organized in Denmark in 1936. There are now many such associations, which use a few highly selected bulls to breed large numbers of cows.

Regulatory agencies operate in all advanced countries, supported by either local or national governments. The dairy industry has always been strictly regulated because it is concerned with a basic but highly perishable food in which contaminating organisms can grow quickly to dangerous numbers. Most municipalities require inspection of the cattle, the farm, trucks, and factories.

BIBLIOGRAPHY. Works treating production, processing, plant operation, and marketing include: A.W. FARRALL, *Engineering for Dairy and Food Products* (1963); H.S. HALL, YNGVE ROSEN, and HELGE BLOMBERGSSON, *Milk Plant Layout* (1963); P.M. REAVES and H.O. HENDERSON, *Dairy Cattle Feeding and Management*, 5th ed. (1963); and the NATIONAL COMMISSION ON FOOD MARKETING, *Organization and Competition in the Dairy Industry*, Technical Study No. 3 (1966).

Works treating scientific aspects include: B.H. WEBB and A.H. JOHNSON (eds.), *Fundamentals of Dairy Chemistry* (1965); and COMMONWEALTH BUREAU OF DAIRY SCIENCE AND TECHNOLOGY, *Dairy Science Abstracts* (monthly).

Nutritional aspects are discussed in: A.M. HARTMAN and L.P. DRYDEN, *Vitamins in Milk and Milk Products* (1965); and in the U.S. DEPARTMENT OF AGRICULTURE, *Composition of Foods*, USDA Handbook No. 8 (1963).

Works treating specific dairy products include: B.H. WEBB and E.O. WHITTIER (eds.), *Byproducts from Milk* (1970); W.S. ARBUCKLE, *Ice Cream* (1966); H.E. WALTER and R.E. HARGROVE, *Cheese Varieties and Descriptions*, USDA Handbook No. 54 (1969); and F.H. MCDOWALL, *The Buttermaker's Manual*, 2 vol. (1953).

(B.H.We.)

Dakar

Dakar, capital of Senegal, is situated close to Africa's westernmost point. Its strategic and impressive site, and the importance of its port and airport, make it one of Africa's major cities. From 1904 to 1959 it was the capital of French West Africa, and from 1959 to 1961 of the Mali Federation. The name comes from *dakhar*, a Wolof name for the tamarind tree, as well as the name of a coastal village inhabited by members of the local Lebu tribe, which lies south of what is now the first pier. The population of the metropolitan area is about 660,000.

History. European settlement in the Dakar area began when the Dutch bought the islet of Gorée, near Dakar Point, in 1617; it was captured by the French in 1677. In Anglo-French wars, Gorée was taken by the British five times; in peacetime it was a calling point for French East Indian ships, a centre of the slave trade, and, finally, a base for the suppression of the slave trade.

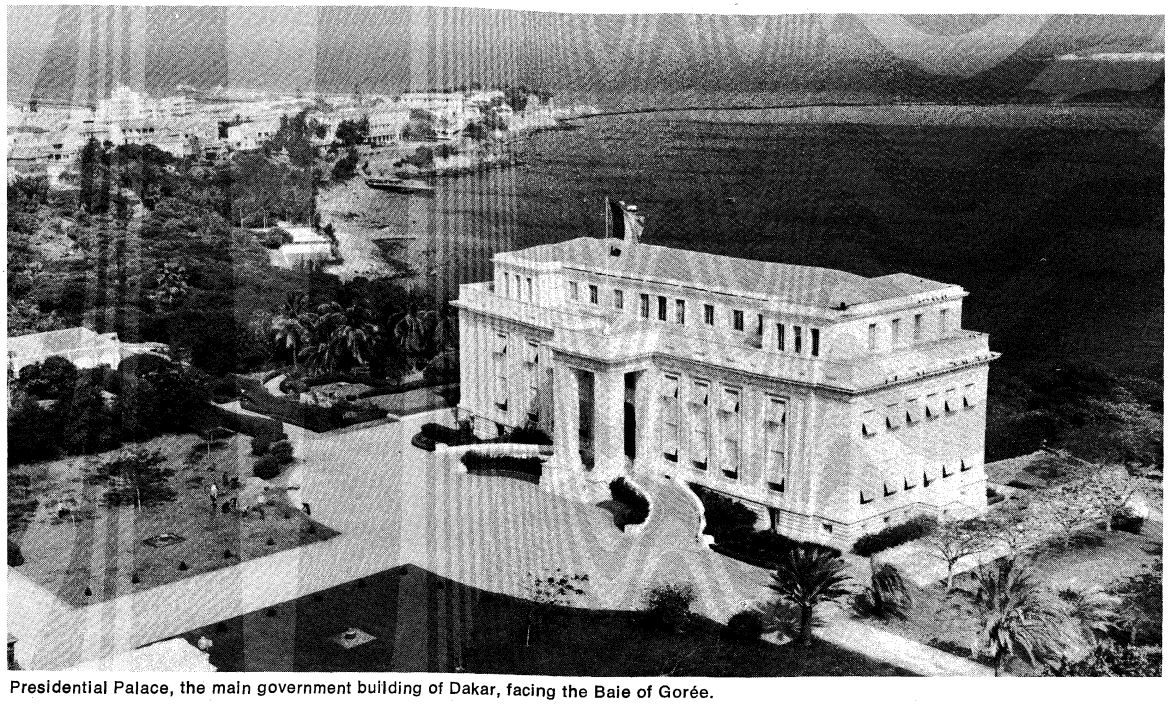
French occupation

The mainland was occupied by France in 1857. A pier was built on Dakar Point, and in 1866 French steamships serving South America began to call there to take on coal. The next impetus to development came with the opening in 1885 of West Africa's first railway, from Saint-Louis to Dakar. The object was to replace Saint-Louis, the port for the then important Senegal Valley, with the better port of Dakar. The railway achieved this, but unexpectedly did far more for Dakar and Senegal by stimulating the cultivation of peanuts (groundnuts) in the vicinity of its track. The increase in trade led to an extension of the jetty in 1892 and to the building of the port's first breakwater.

Anglo-French rivalries in Africa and British troubles in South Africa caused France in 1898 to decide to establish a naval base at Dakar. As a result, a northern breakwater was built to enclose a large deepwater harbour. The southern jetty was again extended, and a dry dock provided. In 1904 Dakar replaced Gorée as the federal capital of French West Africa. Two more southern piers were built between 1904 and 1910, and other facilities were improved. By 1914 Dakar was a well-equipped port and pleasantly planned town, with a population of almost 24,000.

World War I brought a great increase in the tonnage of shipping using the port; this increase was partly maintained after the war by the opening in 1924 of the Dakar-Bamako-Koulikoro railway line to the French Sudan (now Mali). The railway brought new transit trade to the port, and peanut cultivation was again stimulated in both countries. Between 1926 and 1933 two piers were added near the landward end of the northern breakwater especially for peanut export, supplemented in the later 1930s by the installation of pipelines for the export of peanut oil. These improvements killed the port of Rufisque, about 17 miles to the east, as a peanut-shipping competitor to Dakar. A fuelling pier was also built on Dakar's northern breakwater. Just before World War II the original pier on the southern jetty was further improved, and the breakwater was equipped to discharge oil tankers and refuel other vessels. By 1936 the town's population was almost 93,000.

During World War II Dakar, like all of French West Africa, recognized the authority of the Vichy administration of France in 1940, and the efforts of the Free French, under the leadership of General de Gaulle, to



Presidential Palace, the main government building of Dakar, facing the Baie of Gorée.
Harold M. Lambert Studios

secure the town in the same year failed badly. Further development of Dakar was delayed until French West Africa rallied to the Allies in 1943. Yof Airport was then constructed with American aid, and has since become of intercontinental significance. Another pier for peanut exports was added on the northwestern side of the harbour, the northern jetty and the westernmost southern pier were extended, and a cold-storage plant was provided. The storage plant encouraged Japanese and other tuna fleets to call at Dakar and led to a local fleet being established. It also made possible the import of more varied foodstuffs from temperate climates.

Postwar development

World War II gave a great impetus to peanut-oil refining because of local and North African needs for vegetable oil, previously refined mostly in France. Industries manufacturing cigarettes, textiles, shoes, soft drinks, and other consumer products began to be established to supply the entire French West African market. When, from 1959 to 1961, French West Africa split into eight independent states, each seeking its own industries, markets were reduced. The threat to Dakar's industries was less than feared, however, because the local market increased through the continued presence of about 28,000 French residents, by the relative prosperity of the port at times when the Suez Canal was closed, by the advent of cruise liners, and, finally, by increasing air traffic.

The years after World War II saw the greatest urban expansion, in the course of which the area of Dakar doubled. New quarters were built in the 1950s to the north and northwest of Dakar beyond the Médina, at Fann, Point E (with the campuses of the university and of the Fann Hospital Centre between them), Mermoz, and in the 1960s at many estates of Grand Dakar in the north. The first of these new quarters were originally European, but are no longer exclusively so. The northward spread of Dakar is also the consequence of an expansion in both industry and transport, including an increase in ownership of small French cars. The islet of Gorée has become a charming tourist centre, as well as a retreat for the diplomatic and richer commercial communities.

The contemporary city. *The site.* The narrow, westward-tapering Cape Verde Peninsula is in part a tombolo or former island that has been joined to the land by coastal drift and windblown sand on the north and by backwash on the south. Cape Manuel to the south and the offshore islets are Miocene basalts (from 7,000,000 to 26,000,000 years old). The centre of the town is built on Cretaceous or Eocene limestone (from 38,000,000 to 136,000,000 years old) on some of which there is laterite (red, leached, ferruginous soil and rock), while to the northwest are the two conical Pleistocene hills (from 10,000 to 2,500,000 years old) called Les Mamelles (breasts), the westerly and higher being Cape Verde itself. Basalts from these form the Almadies Point, Africa's western extremity. All these rocks provide building stone, and compose the low cliffs from the harbour to Capes Manuel, Verde, the Almadies, and Yof. To the northeast there is sand, on which the harbour and the industrial and northern housing estates have been built. The urban area extends about three miles, the limit of Dakar proper coinciding with the transverse avenue El Hadji Malik Si. Beyond are the peripheral arrondissements (quarters). Dakar Commune includes Dakar proper, the peripheral arrondissements, and the suburbs of Rufisque, Bargny, and Sebikhotane.

Climate and soils. The climate, although humid, is usually pleasant because northern winds and sea breezes sweep across the narrow peninsula, helped by the street alignments. Mean maximum temperatures vary from 79° F (26° C) in January to 89° F (32° C) in September to October, and the mean minimum temperatures from 63° F (17° C) in February to 76° F (24° C) from July to October. Annual rainfall is 23 inches, falling between late June and early October. Originally, the volcanic areas, such as the island of Gorée, were devoid of surface water, while the sandy areas were often marshy.

The city's districts

The city plan. The shape of the peninsula, its geology, history, colonial and other policies, such as subsidiary houses for civil servants, have created contrasted districts. In the southern district are public buildings, hospitals, the Pasteur Institute, and embassies. North is the business district, which is focussed on the central Place de l'Indépendence. North and east lie the quarters associated with the port, such as the port proper, the naval arsenal, the fishing harbour, and the peanut export sector. Near the latter, and close to the railway, are the older peanut-crushing plants and other factories, and further north is the industrial estate of Hann. The central business district and its northwestern fringes were until 1939 bordered by open space. Industry, markets, and sports stadiums later occupied this area, to the north of which lies the Médina, the first out-of-town Senegalese quarter, designed by the French in 1916. Beyond are the quarters built since World War II.

Transport. Roads are generally adequate for traffic needs; a motorway leads north and a good road north-

east. Buses serve the main areas, and both buses and the railway provide limited suburban service.

The port is fully sheltered, and always easy for ships to enter. There are five miles of quays, with depths of up to 36 feet; about 30 vessels can be accommodated. The port facility is equipped for repairs, refuelling, and the provision of water and victuals. In 1969 more than 5,300 vessels entered Dakar; goods exported totalled 1,559,000 tons (of which 1,048,000 tons were phosphates), and imports amounted to almost 1,973,000 tons (of which hydrocarbons to fuel shipping and generate electricity accounted for 1,001,000 tons). The modern airport is important for European-South American, American-African, European-African, and local services. In 1969 more than 8,000 commercial aircraft were handled.

Demography. The last sample census, in 1960–61, put the population of Dakar and Gorée at about 266,000 Africans and about 32,000 non-Africans. In 1968 the population of Dakar proper was estimated as more than 550,000. In addition, there are the peripheral arrondissements, which had about 65,000 Africans and about 3,000 non-Africans in 1961. As to tribal affiliation, the 1960–61 census revealed that the Wolof comprised 46 percent of the Africans, the Tukulor 11 percent, the Lebu 8 percent, and the Serer 6 percent. The total metropolitan area population was estimated at 661,000 in 1968. The annual growth of Dakar proper is about 6 percent, an increase that is now more by natural growth than by immigration. In religion the Africans are mostly Islāmic, while the non-Africans are mostly Catholic Christians.

Housing. Types of housing range from ultraluxurious mansions or apartments in the south and in the Fann area to shacks made of flattened containers or boxes in the north-central quarters near the motorway. The northern estates have been built by commercial housing societies. Public buildings are imposing, and include the Presidential Palace, the Administrative Building, the National Assembly, and the buildings around Place de l'Indépendence and at the university.

Manufacturing industries

Economic life. Dakar is one of tropical Africa's leading industrial and service centres. Industries not already cited include fish canning, flour milling, brewing, truck assembly, and oil refining. Commercial activities are usually prosperous and international in their range.

Services and institutions. Public services, by the standards of tropical Africa, are good—especially the educational and medical services. Dakar has several newspapers and many cinemas. There are excellent museums of the sea and of history on Gorée, and of ethnography and archaeology in Dakar; there is also a village of working craftsmen. There is, however, no permanent theatre nor are there any significant public libraries. There is a radio station and one government-sponsored educational television station. The only park and the zoo are at Hann; other open spaces are few. The corniche road (cut into the cliff) round Cape Manuel affords fine views of the harbour and the islands. On this road is a large swimming pool, and there are some good beaches.

Dakar is neither purely Senegalese nor typically West African; rather it is southern French in character and international in function.

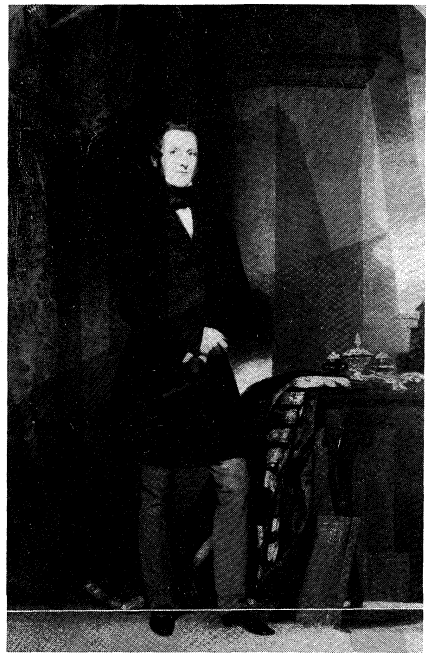
BIBLIOGRAPHY. RICHARD J. PETEREC, *Dakar and West African Economic Development* (1967); ASSANE SECK, *Dakar: Métropole Ouest-Africaine* (1971), and "The Changing Role of the Port of Dakar," in B.S. HOYLE and D. HILLING, *Seaports and Development in Tropical Africa* (1970); and R.J. HARRISON-CHURCH, *West Africa*, 6th ed., pp. 193–201 (1968), are the most recent sources. Older studies are D. WHITTLESEY, "Dakar and the Other Cape Verde Settlements," *Geogrl. Rev.*, 31:609–638 (1941), and "Dakar Revisited," *Geogrl. Rev.*, 38:626–632 (1948); LEON COURSIN, "Dakar: Port Atlantique, *Cah. d'Outre-Mer*, 1:275–285 (1948); J. DRESCH, "Villes d'Afrique Occidentale," *Cah. d'Outre-Mer*, 3:217–222 (1950), and "L'Agglomeration Dakaroise," *Etud. Senegal.*, no. 5 (1955); C. TOUPET, "Dakar," *Tijdschr. Econ. Soc. Geogr.*, 49:35–40 (1958); and A. SECK, "Dakar," *Cah. d'Outre-Mer*, 14:372–392 (1961). For statistics, see *Guid'Ouest Africain* (annual); *Situation Economique du Sénégal* (annual); and *Bulletin Statistique et Economique Mensuel* (monthly).

(R.J.H.-C.)

Dalhousie, James Ramsay, 1st marquess of

Governor general of India from 1847 to 1856, James Andrew Broun Ramsay, 1st marquess of Dalhousie, added to the territories under British administration by many conquests and annexations. At the time, confidence in British rule and its benefits was at its peak, and Dalhousie, who fully shared this confidence, accelerated the pace of westernization. So radical were the resulting changes and so widespread the resentment they caused that his policies were frequently held responsible for the Indian Mutiny in 1857, one year after his retirement.

By courtesy of the National Portrait Gallery, London



Dalhousie, oil painting by Sir John Watson-Gordon, 1847. In the National Portrait Gallery, London.

He was born on April 22, 1812, the third son of George Ramsay, the 9th earl of Dalhousie, at Dalhousie Castle, Midlothian, Scotland. His family had traditions of military and public service but, by the standards of the day, had not accumulated great wealth, and, consequently, Dalhousie was often troubled by financial worries. Small in stature, he also suffered from a number of physical infirmities. Throughout his life he derived energy and satisfaction from the thought that he was achieving public success in spite of private handicaps.

After an undistinguished career as an undergraduate at Christ Church, Oxford, he married Lady Susan Hay in 1836 and entered Parliament the following year. From 1843 he served as vice president, and from 1845 as president, of the Board of Trade in the Tory (conservative) ministry of Sir Robert Peel. In that office he handled a number of railroad problems and gained a reputation for administrative efficiency. He lost his post when Peel resigned in 1846. In the following year he accepted the new Whig ministry's offer of the governor generalship of India, becoming the youngest man ever appointed to that post.

Role in the Punjab rebellion of 1848

When Dalhousie arrived in India in January 1848, the country seemed peaceful. Only two years earlier, however, the army of the Punjab, an independent state founded by the religious and military sect of the Sikhs, had precipitated a war that the British had won only with great difficulty. The discipline and economy enforced by the new Sikh regime, sponsored by the British, aroused discontent, and in April 1848 a local rebellion broke out at Multan. This was the first serious problem faced by Dalhousie. Local officers urged immediate action, but he delayed, and Sikh disaffection spread

throughout the Punjab. In November 1848, Dalhousie dispatched British troops; and, after several British victories, the Punjab was annexed in 1849.

Dalhousie's critics maintained that he had allowed a local rebellion to grow into a national uprising so that he could annex the Punjab. But the commander in chief of the British army had warned him against precipitate action. Certainly, the steps Dalhousie eventually took were somewhat irregular; the uprising at Multān had been directed not against the British but against policies of the Sikh government. In any event, he was created marquess for his efforts.

In 1852, commercial disputes in Rangoon prompted new hostilities between the British and the Burmese, rival powers in the area for almost a century, prompting the Second Burmese War. It was settled within the year with little loss of life and with the British annexation of Rangoon and the rest of the Pegu province. Dalhousie was again criticized for aggressive diplomacy, but Britain profited from the installation of a new Burmese government that was less aggressive abroad and less oppressive at home. Another advantage was that Rangoon, Britain's most valuable acquisition from the war, became one of the biggest ports in Asia.

Dalhousie also took advantage of every opportunity to acquire territory by peaceful means. The East India Company (no longer an independent corporation but largely under the control of the British government) was rapidly becoming the predominant power in India. It had concluded alliances with Indian rulers, promising to support them and their heirs in return for various concessions, including the right to keep a British resident and a military force within the state. Although this type of agreement gave the British an effective influence over general policy, Dalhousie sought to acquire even more power. It was customary for a ruler without a natural heir to ask the British government whether he could adopt a son to succeed him. Dalhousie concluded that if such permission were refused, the state would "lapse" and thereby become part of the British possessions. On these grounds, Sātāra was annexed in 1848 and Jhānsi and Nāgpur in 1854. Dalhousie maintained that there was a difference in principle between the right to inherit private property and the right to govern, but his main argument was his own belief in the moral and material benefits of British rule.

Policy of "lapse" and annexation

His annexation of Oudh in 1856, however, entailed grave political danger. Here there was no question of lack of heirs; the Nawab (ruler) was simply accused of misgovernment, and the state was annexed against his will. The transfer of power over the Nawab's protests offended the Muslim elite. More dangerous was the effect on the British army's Indian troops, many of whom came from Oudh, where they had occupied a privileged position before its annexation. Under the British government, however, they were treated as equals with the rest of the population, which represented a loss of prestige. Moreover, after Dalhousie's departure in 1856, the landed aristocracy of Oudh lost many of its privileges. In these various ways, the annexation of Oudh contributed to the mutiny and rebellion of the following year.

Dalhousie's energy extended beyond the mere acquisition of territories. His greatest achievement was the molding of these provinces into a modern centralized state. His confidence in Western institutions and his ability as an administrator immediately led him to attend to the development of a communication and transportation system. He gave much attention to the planning of the first railways. Drawing on the knowledge he had acquired in London at the Board of Trade, he laid the foundation of future railway development, outlining the basic concept of trunk and branch lines and making provisions to safeguard both the railway workers and the property owners affected by railway construction. He planned and instituted a network of electric telegraph lines, promoted the completion of the Grand Trunk Road between Calcutta and Delhi and its extension into the Punjab, and instituted a centralized postal system, based on a low uniform rate paid in advance by the purchase of stamps, thus replacing a variety of methods characterized by uncertainty of delivery and high rates. His social reforms included strong support for the suppression of female infanticide in the Punjab and in the northwest generally and the suppression of human sacrifice among the hill tribes of Orissa. Besides encouraging the use of the vernacular languages in schools, he gave particular encouragement to the education of girls.

He left India in 1856, and the controversies aroused by his policy of annexation, which were widely—and justly—criticized as contributory factors to the mutiny and rebellion of 1857, overshadowed his achievements in modernization. Vexed by such arguments and exhausted by his years of overwork and ill health in India, he died on December 19, 1860, at Dalhousie Castle.

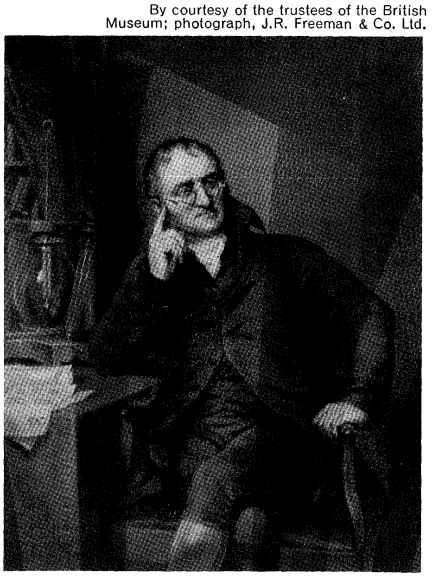
BIBLIOGRAPHY. WILLIAM LEE-WARNER, *The Life of the Marquis of Dalhousie*, 2 vol. (1904), the standard biography; M.N. DAS, *Studies in the Economic and Social Development of Modern India, 1848–56* (1959), a careful account of Dalhousie's economic and social policies; M.A. RAHIM, *Lord Dalhousie's Administration of the Conquered and Annexed States* (1963), an argument for the position that the advantages Dalhousie expected from annexation were precluded by the difficulties of introducing British rule; S.N. PRASAD, *Paramountcy Under Dalhousie* (1964), an analysis of Dalhousie's policy toward Indian rulers.

(K.A.B.)

Dalton, John

John Dalton was a self-tutored English chemist and physicist who crystallized modern scientific thought about the nature of matter with his development of the atomic theory, one of the basic theories of all sciences and the one that marked the establishment of chemistry as a true science.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Dalton, engraving by W. Worthington, after a portrait by William Allen, 1814.

Dalton was born in the village of Eaglesfield, in the North of England, on Sept. 6, 1766, the son of a Quaker weaver. When only 12 he took charge of a Quaker school in Cumberland and two years later taught with his brother at a school in Kendal where he was to remain for 12 years until he became a teacher of mathematics and natural philosophy at New College in Manchester, a college established by the Presbyterians to give a first class education to both laymen and candidates for the ministry, the doors of Cambridge and Oxford being open at that time only to members of the Church of England. He resigned this position in 1800 to become secretary of the Manchester Literary and Philosophical Society and served as a public and private teacher of mathematics and chemistry. In 1817 he became president of the Philo-

sophical Society, an honorary office that he held until his death in 1844.

Early scientific research

In the early days of his teaching, his way of life was influenced by a wealthy Quaker, a capable meteorologist and instrument maker, who interested him in the problems of mathematics and meteorology. His first scientific work, which he began in 1787 and continued until the end of his life, was to keep a diary—which was ultimately to contain 200,000 entries—of meteorological observations recording the changeable climate of the lake district in which he lived. In 1793 he published *Meteorological Observations and Essays*. He then became interested in preparing collections of botanical and insect species. Stimulated by a spectacular aurora display in 1787, he began observations about aurora phenomena—luminous, sometimes coloured displays in the sky caused by electrical disturbances in the atmosphere. His writings on the aurora borealis reveal independent thinking unhampered by the conclusions of others. As he notes, "Having been in my progress so often misled by taking for granted the results of others, I have determined to write as little as possible but what I can attest by my own experience." In his work on the aurora he concluded that some relationship must exist between the aurora beams and the earth's magnetism: "Now, from the conclusions in the preceding sections, we are under the necessity of considering the beams of the aurora borealis of a ferruginous (iron-like) nature, because nothing else is known to be magnetic, and consequently, that there exists in the higher regions of the atmosphere an elastic fluid partaking of the properties of iron, or rather of magnetic steel, and that this fluid, doubtless from its magnetic property, assumes the form of cylindric beams." Some of his studies in meteorology led him to conclusions about the origin of trade winds involving the earth's rotation and variation in temperature—but he may not have been aware that this theory had already been proposed in 1735 by George Hadley. These are only some of the subjects on which he wrote essays that he read before the Philosophical Society: others included such topics as the barometer, thermometer, hygrometer, rainfall, the formation of clouds, evaporation and distribution and character of atmospheric moisture, including the concept of the dew point. He was the first to confirm the theory that rain is caused not by any alteration in atmospheric pressure but by a diminution of temperature. In his studies with water he determined the point of the maximum density of water to be 42.5° F (later shown to be 39.16° F). Along with his other researches he also became interested in the theory of colour blindness, a condition that both he and his brother shared. The results of this work were published under the title "Extraordinary Facts Relating to the Vision of Colours."

Research in chemistry

Dalton had a gift for recognizing many unsolved scientific problems around him as well as a tremendous capacity to organize research on a wide variety of problems. An indefatigable investigator or researcher, he had an unusual talent for formulating a theory from a variety of data. The mental capacity of the man is illustrated by his major work that was to begin at the turn of the century—his work in chemistry. Although he taught chemistry for six years at New College, he had no experience in chemical research. He embarked on this study with the same intuitiveness, independence of mind, dedication, and genius for creative synthesis of a theory from the available facts that he had demonstrated in his other work. His early studies on gases led to development of the law of partial pressures (known as Dalton's law), which states that the total pressure of a mixture of gases equals the sum of the pressures of the gases in the mixture, each gas acting independently. These experiments also resulted in his theory according to which gas expands as it rises in temperature (the so-called Charles' law, which should really be credited to Dalton). On the strength of the data gained in these studies he devised other experiments that proved the solubility of gases in water and the rate of diffusion of gases. His analysis of the atmosphere showed it to be constant in composition to 15,000 feet. He devised a system of chemical symbols and, having ascertained the relative weights of atoms (particles of matter), in 1803 he arranged them into a table. In addition, he formulated the theory that a chemical combination of different elements occurs in simple numerical ratios by weight, which led to the development of the laws of definite and multiple proportions. Dalton discovered butylene and determined the composition of ether, finding its correct formula. Finally, he developed his masterpiece of synthesis—the atomic theory, the thesis that all elements are composed of tiny, indestructible particles called atoms that are all alike and have the same atomic weight.

All of Dalton's studies and writings, many included in his *New System of Chemical Philosophy* (part I, 1808; part II, 1810), cast light on the man—dedicated to scientific research, independent in his approach, often diffident in seeking help in scientific papers that would aid him, or misguide him as he often thought—a genius in synthesizing facts and ideas. Almost a recluse, with few friends, and unmarried, he was deeply dedicated to a search for the answer to scientific problems. His homemade equipment was crude, his data not usually exact, but good enough to give his alert and creative mind clues to the probable answer. A scientist exclusively dedicated to the pursuit of new knowledge, Dalton remained a man of simple wants and uniform habits; his dress and manners were consistent with his Quaker faith. He died at Manchester on July 27, 1844.

Dalton's record keeping, although remarkable for quantity, often lacked exactness in dating, probably because he revised his manuscripts as secretary of the Philosophical Society between the time of the oral presentation and the publication. The exact date of some of his work, especially the atomic theory, is still in doubt because he had the opportunity to revise it before publication. His documents were destroyed during the bombings of England in World War II. A fellow of the Royal Society, from whom he received the Gold Medal in 1826, and a corresponding member of the French Academy of Sciences, John Dalton was also cofounder of the British Association for the Advancement of Science. A giant among early scientists and at the same time a quiet, modest man, he was highly revered and respected by his countrymen as well as by scientists who benefitted from his discoveries.

BIBLIOGRAPHY. H.E. ROSCOE, *John Dalton and the Rise of Modern Chemistry* (1895), the most authoritative biography, and with A. HARDEN, *A New View of the Origin of Dalton's Atomic Theory* (1896), original material on Dalton's research; D.S.L. CARDWELL (ed.), *John Dalton and the Progress of Science* (1968), J.B. CONANT, and L.K. NASH (eds.), *Harvard Case Histories in Experimental Science*, vol. 1 (1957), probably the most critical recent analysis of Dalton's work; FRANK GREENAWAY, *John Dalton and the Atom* (1966).

(A.B.Ga.)

Dam

A dam is a structure built across a stream, river, or estuary to retain water. Its purposes are to meet demands for water for human consumption, irrigation, or industry; to reduce peak discharge of floodwater; to increase available water stored for generating hydroelectric power; or to increase the depth of water in a river so as to improve navigation. An incidental purpose can be to provide a lake for recreation.

Auxiliary works at a dam may include spillways, gates, or valves to control the discharge of surplus water downstream from the dam; an intake structure conducting water to a power station or to canals, tunnels, or pipelines for more distant use; provision for evacuating silt carried into the reservoir; and means for permitting ships or fish to pass the dam. A dam therefore is the central structure in a multipurpose scheme aiming at the conservation of water resources. The multipurpose dam holds special importance in the underdeveloped countries, where a small nation may reap enormous benefits in agriculture and industry from a single dam.

Dams fall into several distinct classes, by profile and by building material. The decision as to which type of dam to build depends largely on the foundation conditions in the valley and the construction materials available. Broadly, the choice of materials now lies between concrete, soils, and rockfill. Though a number of dams were built in the past of jointed masonry, this practice is now largely obsolete. The monolithic form of concrete dams permits greater variations in profile, according to the extent water pressure is resisted by the deadweight of the structure, is transferred laterally to buttresses, or is carried by horizontal arching across the valley to abutments formed by the sides of the valley.

HISTORY

Ancient dams. The earliest recorded dam is believed to have been on the Nile River at Kosheish where a 15-metre-high (49-foot) masonry structure was built about 2900 BC to supply water to King Menes' capital at Memphis. Evidence exists of a masonry-faced earth dam built about 2700 BC at Sadd-el-Kafara, about 30 kilometres (19 miles) south of Cairo; this dam failed shortly after completion when, in the absence of a spillway, it was overtopped by a flood. The oldest dam in use is a rockfill structure about six metres (20 feet) high on the Orontes (Nahr al-'Āṣī) in Syria, built about 1300 BC.

The Assyrians, Babylonians, and Persians built dams between 700 and 250 BC for water supply and irrigation. Contemporary with these was the earthen Sudd-al-Arim Dam, built near Marib in the Yemen, 14 metres high and nearly 600 metres long. Flanked by spillways, this dam delivered water to a system of irrigation canals for more than 1,000 years. Other dams were built in this period in Ceylon (modern Sri Lanka), India, and China.

The Romans. Despite their undoubted skill as civil engineers, especially in the field of water supply, the Romans' role in the evolution of dams is not remarkable for quantity or for advances in height. Their skill lay in the comprehensive collection and storage of water and in its transport and distribution by aqueducts. Remarkably, at least two Roman dams in southwestern Spain, Proserpina and Cornalbo, are still in use, although a third, the Alcantarilla Dam, has overturned, and the reservoirs of some others have filled with silt. The Proserpina Dam, 12 metres (40 feet) high, has a masonry-faced core wall of concrete backed by earth; it may be regarded there-

Proserpina Dam

By courtesy of (B) Henri Goblot; from (A) *Revista de Obras Publicas* (June 1961); (C) E. Wegmann, *Design and Construction of Dams* (1922); John Wiley & Sons

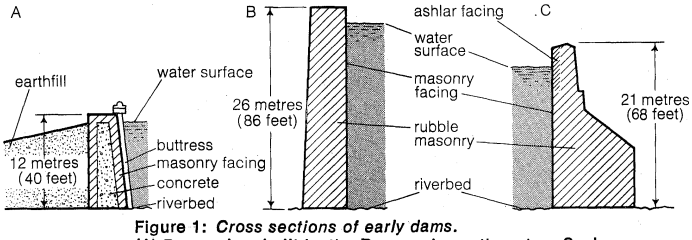


Figure 1: *Cross sections of early dams.*
(A) Proserpina, built by the Romans in southwestern Spain about AD 100. (B) Kebar, in Iran, constructed about AD 1300. (C) Almanza, in Spain, completed about AD 1586.

fore as a forerunner of the modern earth dam. The Proserpina is strengthened on the upstream face by buttresses (Figure 1A). Of similar construction, 14 metres (46 feet) high and 550 metres (1,804 feet) in length, Alcantarilla Dam was supported by a great weight of earth, which eventually caused failure of the wall. Cornalbo represented a further advance in design; the masonry wall was constructed of cells, which were filled with stones or clay, and faced with mortar. It differs from Proserpina and Alcantarilla in having a sloping upstream face and in being straight in plan. Proserpina and Alcantarilla were polygonal in plan. The merit of curving a dam upstream was not apparently fully appreciated by the Romans until such a curved structure, the forerunner of the modern "arch-gravity" dam, was built in AD 550 at Dâra on the present Turkish-Syrian border by Byzantine engineers.

Early dams in the Orient. Quite independently, dam construction evolved in the East. In 240 BC a stone crib was built across the Gukow River in China; this structure was 30 metres (98.4 feet) high and about 300 metres (984 feet) long. Many earth dams of moderate height (in some cases, of great length) were built by the Sinhalese in Ceylon after the 5th century BC to form reservoirs or tanks for extensive irrigation works. The Kalabalala Tank (formed by an earth dam 24 metres [79 feet] high and nearly six kilometres [3.7 miles] in length) had a perimeter of 60 kilometres (37 miles) and helped store monsoon rainfall for irrigating the country around the ancient capital of Anuradhapura. Many of these tanks in Ceylon are still in use today.

The Kalabalala Tank

In Japan the Diamonike Dam reached a height of 32 metres (105 feet) in AD 1128. Numerous dams were also constructed in India and Pakistan. In India a design employing hewn stone to face the steeply sloping sides of earth dams evolved, reaching a climax in the 16-kilometre-long (ten-mile) Veeranam Dam, Tamil Nadu, built from AD 1011 to 1037.

In Iran the Kebar, a pioneer arch dam, was built early in the 14th century (Figure 1B). Spanning a narrow limestone gorge, it reached 26 metres (86 feet) high with a thickness of less than five metres (16 feet). The central curved portion, 38 metres (124.6 feet) in length and radius, was supported on two straight abutments.

Fifteenth to 18th century. In the 15th and 16th centuries dam construction resumed in Italy and, on a larger scale, in Spain, where Roman and Moorish influence was still felt. Of these dams, the Tibi (1579–89) was an arch-gravity structure 42 metres (137.8 feet) high; this height was not surpassed in western Europe until the building of the Gouffre d'Enfer Dam in France, almost three centuries later. Almanza Dam in Spain is illustrated in Figure 1C. An attempt to build a dam 52 metres (170.6 feet) high near Lorca, Spain, at the end of the 18th century failed disastrously in 1802, when earth and gravel below the piled structure washed out. In Europe, where rainfall is ample and well distributed throughout the year, dam construction before the Industrial Revolution was on only a modest scale and was restricted to forming water reservoirs for towns, driving water mills, and making up water losses in navigation canals. An exception was the 36-metre-high (115.2-foot) earth dam built in 1675 at St. Ferréol, near Toulouse, France, to supply the Canal du Midi; for more than 150 years it was the highest earth dam in existence.

Nineteenth century. Up to the middle of the 19th century, dam design was entirely empirical. Knowledge of the properties of materials and structural theory had been accumulating for 250 years; Galileo, Newton, Leibniz, Hooke, Daniel, Bernoulli, Euler, de la Hire, and Coulomb had made outstanding contributions. In the 1850s W.J.M. Rankine, professor of civil engineering at Glasgow University, successfully demonstrated how applied science could help the practical engineer. Rankine's work on the stability of loose earth for example, provided a better understanding of the principles of dam design and performance of structures. This led in turn to improved construction techniques and larger dams. Furthermore, in certain countries, Rankine's work encouraged acceptance of civil engineering as a subject for university study and added to the status of civil engineers. Much remained to be learned of soils and natural rocks in the 100 years after Rankine. Many scientists and engineers made, and continue to make, noteworthy contributions.

Rankine's contributions

Development of modern structural theory. Concrete dam design is based on conventional structural theory. In this relationship two phases may be recognized. The first, extending from 1853 until about 1910 and represented by the contributions of a number of French and British engineers, was actively concerned with the precise profile of gravity dams in which the horizontal thrust of water in a reservoir was resisted by the weight of the dam itself and the inclined reaction of the dam's foundation. Starting about 1910, however, engineers began to recognize that concrete dams were monolithic three-dimensional structures in which the distribution of stress and the deflec-

tions of individual points depended on stresses and deflections of many other points in the structure. Movements at one point had to be compatible with movements at all others. Owing to the complexity of the stress pattern, model techniques were gradually employed. Models were built in plasticine, rubber, plaster, and finely graded concrete. In recent years the digital computer has facilitated use of the analytical method of finite elements, by which a monolithic structure is divided into an assembly of separate blocks. Study of both physical and digital models permits deflections of a dam's foundations and structure to be taken into account.

During the 100 years up to the end of World War II, experience in design and construction of dams advanced in many directions. In the first decade of this century many large dams were built in the U.S. and western Europe. In succeeding decades, particularly during the war years, many impressive structures were built in the U.S. by federal government agencies and private power companies. Hoover Dam, built in 1936, is an outstanding example of an arch-gravity dam built in a narrow gorge across a major river and employing advanced design principles. It has a height of 221 metres (726 feet) from its foundations, a crest length of 379 metres (1,243 feet), and reservoir capacity of 37×10^9 cubic metres.

Among earth dams, Fort Peck Dam, completed in 1940, contained the greatest volume of fill, 96,000,000 cubic metres. This volume will not be exceeded until the completion in 1975 of Tarbela Dam in Pakistan (145,000,000 cubic metres). The table lists the greatest dams in the world.

BASIC PROBLEMS IN DAM DESIGN

Most modern dams continue to be of two basic types: masonry (concrete) and embankment (earthfill). Masonry dams are typically used to block streams running through narrow gorges, as in mountainous terrain; though such dams may be very high, the total amount of material required is limited. Embankment dams are preferred to control broad streams, where only a very large barrier, requiring a great volume of material, will suffice. The choice of masonry or embankment and the precise design depend on the geology and configuration of the site, the functions of the dam, and cost factors.

Site investigation and testing. Investigation of a site for a dam includes sinking trial borings to determine the strata. The borings are supplemented by shafts and tunnels which, because of their cost, must be used as sparingly as possible. In the shafts and tunnels, tests can be made to measure strength, elasticity, permeability, and prevailing stresses in strata, with particular attention given to the properties of thin partings, or walls, between the more massive beds. The presence in groundwater of chemical solutions harmful to the materials to be used in the construction of the dam must be assessed. Sources of construction materials need exploration. As dams continue to increase in height, the study of foundation conditions becomes increasingly critical.

Model tests play a major part in the structural, seismic, and hydraulic design of dams. Structural models are particularly useful in analysis of arch dams and in verifying analytical stress calculations. Various materials have been used for model tests; on some early tests for Hoover Dam, rubber was employed. The need for accurate reproduction of stress patterns in complex models is met by using material of low elasticity. In a sense, dams themselves are models for future design. The instruments built into them to record movements under load, strains within materials after construction, temperature and pressure changes, and other factors are installed primarily to study the performance of the structure and to warn of possible emergencies, but their value in confirming design assumptions is important.

The digital computer has permitted considerable advance in analytical methods of design. Its ability to handle a great volume of data and to solve large sets of simultaneous equations containing many variables has made practicable the method of Finite Element Analysis. In this method a complicated structure is divided into a number of separate equilibrium conditions, and strains are rendered compatible, thus leading to a complete analysis of stress and strain distribution throughout the structure. Finite Element Analysis

Problems of materials. Each of the two basic dam materials, concrete and earth or rock fill, has a weakness that must be overcome by the proper design of the dam.

Weaknesses of concrete. Concrete is weak in tensile strength; that is, it can be pulled apart easily. Concrete dams must therefore be designed to place minimum tensile strain on the dam and to make use of concrete's great compressive strength, or ability to support vertical loads. The chief constituent of concrete, cement, shrinks as it sets and hardens, due to water absorption in the crystalline structure, to evaporation of water to the atmosphere, and to cooling from the higher temperatures reached when the chemical reactions in the cement are in progress during hydration. Because of the large volume of concrete in a dam, shrinkage presents a serious cracking hazard.

Various expedients are used to overcome the problem. Concrete is usually cast in separate blocks of limited height. Gaps may be left to permit heat losses and filled in later. Low-heat cements may be used; these are specially blended so that rates of heat evolution are retarded. Cement content can be safely reduced in the interior concrete in the dam, in which strength and resistance to climatic and chemical deterioration are less important. The cement content, and therefore the heat caused by hydrating, can also be reduced by using aggregate (the other major constituent of concrete) of larger stones. Another expedient is to use other fine-grained materials, such as fly ash (pulverized fuel), as filler, reducing the total cement volume in the concrete. Another is to use certain additives, surface-active agents, and air-entraining agents that permit using a lower water-to-cement ratio in mixing the concrete. Techniques used to speed the cooling process include replacing some of the water in the mix by ice, circulating water through pipes laid in the concrete, and extracting excess water from surfaces by vacuum.

Weaknesses of earth and rock fill. Soils and rock fragments lack the strength of concrete, are much more permeable, and possess less resistance to deterioration and disturbance by flowing water. These disadvantages are compensated for by a much lower cost and by the ability of earth fill to adapt to deformation caused by movements in the dam foundation. This assumes, of course, sufficient usable soil available close to the dam site. In bare mountain country it may be necessary to quarry rock and construct a rockfill rather than an earthfill dam. Earth fill is of course more economical, and often a suitable borrow area can be found close to the site.

Soil consists of solid particles with water and air in between. When the soil is compressed by loading, as occurs in dam construction, some drainage of air and water takes place, causing an increase in pressures between the solid particles. When there is a high rate of seepage, the soil tends to develop differential pressures and reach a condition called quick, in which it behaves as a fluid. Even if it does not reach this condition, there is often some weakening of its structure, and steps must be taken to counter this.

The earthquake problem. Many large dams have been built in the seismically active regions of the world, including Japan, the western United States, New Zealand, the Himalayas, and the Middle East. In 1968 the Tokachi earthquake damaged 93 dams in Honshu, the main Japanese island; all were embankment dams of relatively small height.

Despite a great deal of work on the distribution of seismic activity, the measurement of strong ground motions, and the response of dams to such motions, earthquake design of dams remains imprecise. The characteristics of strong ground motions at a given site cannot be predicted, and all types of dams possess some degree of freedom, imperfect elasticity, and imprecise damping. Never-

World's Greatest Dams Completed by End of 1973

name	type*	completed	river	country	height (metres)	volume (000 cubic metres)	reservoir capacity (000 cubic metres)	authorized power capacity (000 kilowatts)
by height								
Nurek	E	1972	Vakhsh	Soviet Union	317	58,000	10,400,000	2,700
Grande Dixence	G	1962	Dixence	Switzerland	285	5,957	400,000	840
Rossella	GA	1965	Rossella	Italy	265	400	17,200	...
Mica	E	1972	Columbia	Canada	242	32,111	24,669,800	1,800
Mauvoisin	A	1957	Drance de Bagnes	Switzerland	237	2,030	180,000	...
Oroville	E	1968	Feather	U.S.	236	59,639	4,298,500	440
Contra	A	1965	Verzasca	Switzerland	230	658	86,400	...
Bhakra (Gobind Sagar)	G	1963	Sutlej	India	226	4,130	9,868,000	1,050
Hoover (Boulder)	A	1936	Colorado	U.S.	221	3,364	36,703,000	1,354
Dworshak (Bruces Eddy)	G	1972	N. Fork, Clearwater	U.S.	219	4,970	4,277,778	1,060
by volume								
Fort Peck	E	1940	Missouri	U.S.	76	96,034	23,600,000	165
Oahe	E	1963	Missouri	U.S.	75	70,343	29,100,000	595
Gardiner (South Saskatchewan)	E	1968	South Saskatchewan	Canada	68	65,553	9,867,920	800
Mangla	E	1967	Jhelum	Pakistan	116	64,991	7,250,000	0
Oroville	E	1968	Feather	U.S.	236	59,639	4,298,500	440
by size of reservoir								
Owen Falls	G	1954	Victoria Nile	Uganda	31	...	204,800,000	120
Kariba	A	1959	Zambezi	Rhodesia–Zambia	128	1,065	181,591,500	705
Bratsk†	E	1961	Angara	Soviet Union	40	9,563	169,270,000	4,500
Aswān High	G	1970	Nile	Egypt	111	42,620	164,000,000	2,100
Akosombo	R	1965	Volta	Ghana	141	7,900	148,000,000	786
by power capacity								
Grand Coulee	G	1942	Columbia	U.S.	168	8,093	11,795,000	6,180
Krasnoyarsk	G	1967	Yenisey	Soviet Union	124	4,350	73,300,000	6,000
Bratsk†	E	1961	Angara	Soviet Union	40	9,563	169,270,000	4,500
John Day (Lake Umatilla)	EG	1968	Columbia	U.S.	71	2,650	3,256,440	2,700
Nurek	E	1972	Vakhsh	Soviet Union	317	58,000	10,400,000	2,700

*Key: A, arch; E, earth fill; G, gravity; R, rock fill.
†A concrete gravity dam of different size, referred to in text, also forms part of the Bratsk complex.

theless, the digital computer and model testing have given promise of considerable progress. It is now possible to calculate the response of a concrete dam to any specified ground motion; this has been done for the Tang-e Soleyman Dam in Iran and the Hendrik Verwoerd Dam in South Africa.

There has also been considerable advance in the theoretical estimation of the effects of ground motion on embankment dams.

THE MODERN CONCRETE DAM

Concrete gravity dams. Concrete gravity dams share certain features with all types of concrete dams. Running in virtually a straight line across a broad valley, they resist the horizontal thrust of the retained water entirely by their own weight; at each level in their height the water's thrust is deflected down toward the foundation by the weight of the concrete. In this action their purpose resembles that of the abutment of an arched bridge or the buttresses and pinnacles of a church. A gravity dam is a right-angled triangle; its hypotenuse forms the sloping downstream face. The base width is approximately three-quarters the height of the dam.

The three main forces acting on a gravity dam are the thrust of the water, the weight of the dam, and the pressure, or reaction exerted by the foundation, which is necessarily inclined in respect to the superstructure. It is also essential to consider the thrust exerted on the upstream face by silt deposited in the reservoir or by ice on the water surface, the inertia forces that can be caused by seismic action, and, in particular, the buoyant uplift force of water seeping under the dam or into the horizontal joints.

Uplift pressure

Uplift due to seepage has caused sustained discussion among engineers. It calls for the greatest of care in design and construction. Where a dam is founded on solid rock, a simple downward projection of concrete into the rock will generally suffice to cut off seepage and eliminate uplift pressures. Usually, however, the rock foundation is permeable, sometimes to considerable depths, so construction of an absolutely reliable cutoff is either difficult or impossible. Reliance must then be placed on an extensive system of grouting the fissured rock and on relieving uplift pressures by means of drainage. Many dams possess both cutoffs and underdrainage.

Post-tension construction

A relatively new development in the construction of gravity dams is incorporation of post-tensioned steel into the structure. This helped reduce the cross section of Allt Na Lairige Dam in Scotland to only 60 percent of that of a conventional gravity dam of the same height. A series of vertical steel rods near the upstream water face, stressed by jacks and securely anchored into the rock foundation, resists the overturning tendency of this more slender section. This system has also been used to raise existing gravity dams to a higher crest level, economically increasing the storage capacity of a reservoir.

Of special interest are three concrete gravity dams all of which feature a straight sloping downstream face. Bratsk, built across the Angara River at Irkutsk, in the Russian S.F.S.R., and completed in 1964, stands 125 metres (410 feet) above foundation level and, excluding the earth side dams, is nearly 1,525 metres (5,000 feet) in length; it contains 4,500,000 cubic metres (159,000,000 cubic feet) of concrete. Grand Coulee Dam, completed in 1942 across the Columbia River, Washington, is 168 metres high, 1,280 metres in length and contains 8,100,000 cubic metres of concrete. Grande Dixence Dam in Switzerland, completed in 1962 across the narrower valley of the Dixence, has a crest length of 700 metres and contains approximately 5,957,000 cubic metres of concrete; at 285 metres (935 feet) it was the highest dam in the world until the Nurek was completed. By comparison, the Pyramid of Khufu contains 2,600,000 cubic metres (92,000,000 cubic feet) of masonry.

Concrete buttress and multiple-arch dams. Unlike gravity dams, buttress dams do not rely entirely upon their own weight to resist the thrust of the water. Their upstream face, therefore, is not vertical but inclines about 25–45 degrees, so the thrust of the water on the upstream face inclines toward the foundation. Embryonic buttresses existed in some Roman dams built in Spain, among them the Proserpina. As technology advanced, dams with thin buttresses of reinforced concrete supporting inclined panels of similar construction were built. In today's buttress dams, less account is taken of effecting maximum economy in the use of concrete. The trend is to reduce the area of costly formwork necessary and to avoid use of steel reinforcement.

With greater heights, modern buttress dams are inevitably less slender.

Several variations are possible in the design of the junction between each buttress at the water face. Where no relative movement in the buttress foundations is anticipated, the design can link individual buttress heads rigidly, by means of arches, to form a multiple-arch dam. A recent Canadian example of this type is the 214-metre-high (703-foot) multiple-arch Daniel Johnson Dam on the Manicouagan River, Quebec. The dam has a total of 14 buttresses used in its crest length of 1,310 metres (4,297 feet); two very much larger buttresses support the structure over the original riverbed.

Where buttress foundations might yield, the design must allow some freedom of movement between the heads of the buttresses. This is normally achieved by enlarging the heads until they are almost in contact, then joining them with flexible seals. Thus joined, they present a solid face to the water. Such a design was used in the construction in the Farahnaz Pahlavi Dam in Iran. Built for the Tehrān Regional Water Board, this dam has a maximum height of 107 metres (351 feet) and a crest length of nearly 360 metres (1,181 feet).

A comparison between the Daniel Johnson multiple-arch dam and the Farahnaz Pahlavi buttress dam shows that the buttresses have to be placed much closer together than is necessary with a multiple-arch dam. This allows each buttress to be more slender, however, and spreads the load more easily over the foundation. The detailed design at the bottom of the Farahnaz Pahlavi buttresses was necessitated by weak foundation conditions at the site and by the need to limit the length of each buttress to reduce its response to seismic action. By contrast, the Daniel Johnson buttresses could be founded individually, exploiting fully an important advantage of buttress dams over gravity dams, that of smaller uplift forces.

Arch dams. The advantages of building a dam curved in plan, utilizing the water pressure to keep the joints in the masonry closed, was appreciated as early as Roman times. An arch dam is a structure curving upstream, where the water thrust is transferred either directly to the valley sides or, indirectly, through concrete abutments. Theoretically, the ideal constant angle arch in a V-shaped valley has a central angle of 133 degrees of curvature. This leads to the development of the cupola (or variable radius) arch dam with the crest portion overhanging downstream (Figure 2). The constant radius arch dam generally has a vertical upstream face. There are many other factors, however, to take into account, including fixity at the abutments at the upper levels and the vertical cantilever effect of the arch at the riverbed level.

From W. Creager, J. Justin, and J. Hinds, *Engineering for Dams* (1945); John Wiley & Sons, Inc.

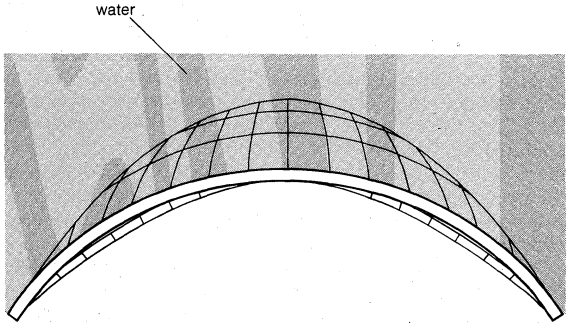


Figure 2: Plan view of an arch dam showing double curvature.

An arch dam is therefore a shell structure, admittedly sometimes of significant thickness, that owes its strength essentially to its curved profile but that is supported at the riverbed and up the valley sides by constraints that cause both flexure and shear on the membrane. Dependent for its strength upon effective support at its abutments, its very strength and rigidity make it sensitive to movements at the abutments. Only favourable sites providing sound rock are suitable for arch dams.

The great reserves of strength inherent in an arch dam were dramatically displayed in 1963 when the reservoir behind Vaiont Dam in Italy was virtually destroyed by a landslide. Vaiont, at that time the second highest dam in the world, was built across a narrow gorge on limestone foundations so that the crest 262 metres (858 feet) above the valley bottom was only 190 metres (623 feet) in length. Some large-scale instability in the mountainside above the reservoir had been observed earlier by the engineers during filling; they were allowed to proceed very slowly and three years later on October 9, 1963, with filling still incomplete, about 240,000,000 cubic metres of soil and rock slid down into the reservoir, sending a tremendous volume of water to a height of 260 metres (853 feet) on the opposite side of the valley. The flood overtopped the dam to a depth of 100 metres (328 feet) and surged down the valley, causing a major tragedy, the destruction of several villages with a large loss of life. Yet only superficial damage was caused to the dam, which, at its crest, is about 3.4 metres (11.2 feet) thick.

Vaiont Dam disaster

EMBANKMENT DAMS

General characteristics. Earlier embankments were undoubtedly built as simple homogeneous structures, with the same material used throughout. No effort was made at first to subdivide the dam into separate zones with the best suited material in each zone. The homogeneous dam nicely illustrates the general behaviour of an embankment dam and demonstrates the reasons for the rather baffling pattern of heterogeneous dam profiles employed.

Like a concrete gravity dam, the weight of an embankment dam deflects the horizontal thrust of the water pressure down to the foundation. The resultant pressures on the foundation must not cause excessive deformation or collapse.

Unlike concrete, embankment dam materials possess only limited resistance to water penetration. The rate of penetration depends on the pressures exerted by the water in the reservoir, the length of seepage paths through the dam, and the permeability of the material of construction. Soils and rock range from substantially impermeable clays, through silts and sands, to coarse-graded gravels and rock fragments that possess little resistance to the movement of water. The range is extremely wide; the seepage rate through clean gravel is 10,000 times that through sand, 10,000,000 times that through silt, and 100,000,000 times that through dense clay.

Water penetration

An embankment dam must be stable in itself. Its side slopes must not slip or slide; liquefaction of the soils must not occur; erosion of the soils, as the result of water overtopping the crest, by wave action on the upstream face, or by seepage washing out the fine material through the coarser, must be avoided. As with a concrete dam, seepage of water from the reservoir through the foundation and under the actual embankment also must be controlled.

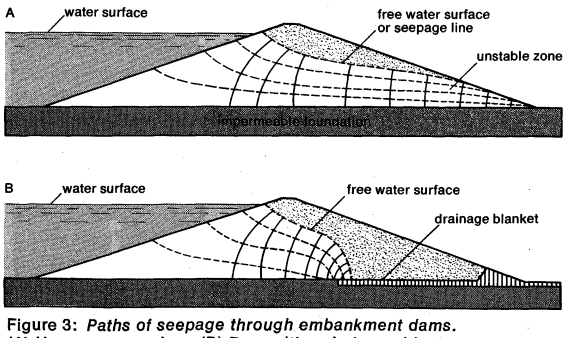


Figure 3: *Paths of seepage through embankment dams.* (A) Homogeneous dam. (B) Dam with a drainage blanket.

Potential weakness. There are three parts of a dam where weakening of the soil structure and liquefactions can occur. In Figure 3A the pattern of seepage through a homogeneously filled dam is shown. Near the downstream toe the gradient of the pore-water pressures is

steep and constraints holding the soil structure together are low; this is one area of weakness in an embankment dam. One solution is to introduce drainage as indicated in Figure 3B where the area of steep seepage gradients has been moved to where the soil is constrained near the centre of the dam. Seepage gradients at the vulnerable downstream toe are eliminated.

A second area of potential weakness is the upstream face when the water in the reservoir is rapidly drawn down. If the pore-water pressures cannot adjust themselves fast enough to this change in the free-water surface in the reservoir, severe seepage gradients begin; these can cause failure. A zone of freely draining fill of coarser grading can be placed on the upstream face to counter this.

Water seepage from the reservoir through the foundations under the dam is another potential weakness. Owing to their great widths, embankment dams can be constructed on unfavourable sites, such as open-joined rock or weaker and possibly locally permeable clay. It is necessary, however, either to check or to drain away harmlessly the seepage water that would otherwise weaken the downstream parts of the dam, in extreme cases causing it to fail. Several countermeasures, possibly in combination, can be employed: the foundation can be grouted or a cutoff trench excavated and backfilled with an impermeable material; a drainage blanket can be constructed at the base of the downstream part of the dam, or individual drainage wells or galleries can be excavated; the length of the seepage paths under the dam can be extended by means of an impermeable blanket laid on the upstream side of the dam, or additional free-draining fill can be placed at the downstream toe of the dam.

Construction techniques. *Use of cores.* Today, all large embankment dams have a core of lower permeability built near their centre. Suitable materials, such as a plastic clay, are weaker than more permeable soil. The width of the core is restricted to that necessary to lower sufficiently the pore pressures in the downstream part of the dam. Though the top of the core must be at the crest of the dam, the core itself need not be vertical. On some rockfill dams the core can slope forward to an extreme position where it lies on the upstream face. Usually a sloping core occupies an intermediate position so it can be constructed on a sloping face of a partially built dam.

Where seepage is inevitable, the use of finely graded core material in proximity to coarser material is avoided. Bands of intermediately graded material must be inserted to prevent the finely graded material from leaching through the coarse zones. Filter zones are graded so each band is four to five times coarser than the preceding band.

Aswān High Dam

Figure 4A shows a typical section of the Aswān High Dam, an embankment 111 metres (365 feet) high, built of dune sand and rock fill on a very permeable foundation of deep alluvium. Here, the central clay core is vertical; this barrier to seepage is extended to the original riverbed as grouted sand and below the riverbed to a depth of 225 metres (740 feet) as a grout curtain. A corrugated blanket of clay extends upstream within the dam from the base of the core. Within the upstream and downstream cofferdams, partly of rock fill, much of the filling is of compacted sand. Filter layers separate the cofferdam filler from the outer layers of freely draining rock fill. Drainage wells will be observed below the downstream toe. The early stages of construction were carried out under deep water; hence the grouted coarse sand between the clay core and the grout curtain.

The rather simpler section of Oroville Dam, in California, is shown in Figure 4B. Until the 317-metre-high (1,040-foot) Nurek Dam in the Soviet Union was completed, Oroville (236 metres) was the highest embankment dam in the world. Unlike the Aswān High Dam, Oroville was not built on deep permeable alluvium, nor was it necessary to place part of the fill under water. Unusual is the concrete block at the base of the sloping core designed to fill in the incised gorge of the Feather River Canyon. The grout curtain, compared with that of

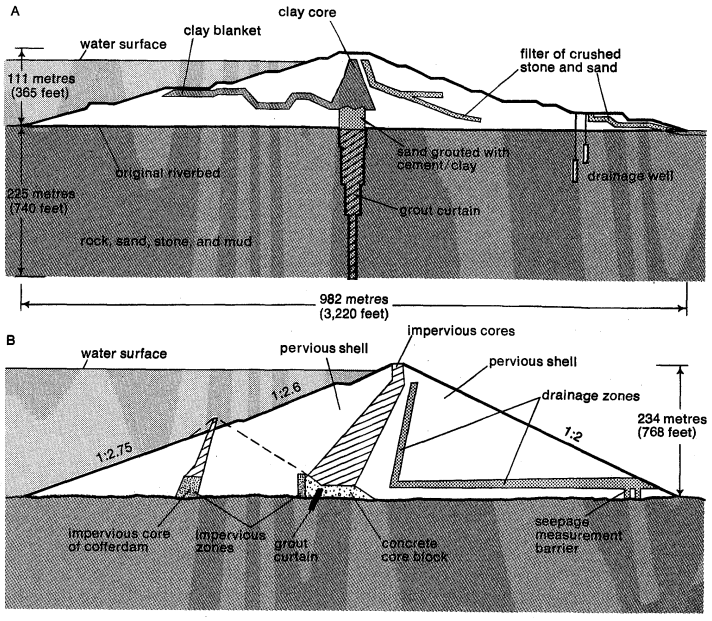


Figure 4: Sections of (A) the Aswān High Dam, Egypt, and (B) Oroville Dam, California.
From (A) 9th International Congress on Large Dams; (B) *Civil Engineering*, A.S.C.E. (June 1969)

the Aswān High Dam, is of nominal depth. On each side of the sloping core, transition zones separate the core from the main mass of more pervious filling. The downstream transition zone is backed by a curtain drain of selected pervious material connected to a drainage blanket on the downstream side. The upstream face of the dam is protected against wave action by a one-metre layer of broken stone (riprap).

Placing and compacting materials. Efficient compacting of soils requires maximum density of dry particles consistent with an economic number of passes of the compacting plant. The process of compacting a soil by kneading it involves expelling as much of the air as practicable; water content is not normally much reduced. The optimum water content for maximum dry density—which results in maximum strength—can be achieved for a given amount of work done on the soil in compaction. In arid climates water must often be added to excavated soils. In temperate climates, however, water content is usually too high except in deeply excavated and well-drained soils.

Normally, soils are placed in embankment dams in thin layers individually compacted by rolling. Finer soils, such as those used in cores, may be harrowed before rolling. Coarser soils, including rock fragments, are compacted by vibration and rolled. Coarse rock fragments (rock fill) are compacted to a limited extent by impact on being dumped from the construction plant; compaction of smaller fragments is assisted by sluicing with water.

Hydraulic filling

In the process of hydraulic filling, sands are dredged from borrow pits, transported in water by pipelines to the filling area, and deposited there by draining off the surplus water. Hydraulic filling is widely practiced in maritime works, if sand is the only construction material available. It has also been used in the construction of embankment dams, although on some inland sites too much water would be required to transport the material. The practice has tended to fall out of favour for dams, but renewed interest in hydraulic filling has been taken in the Soviet Union.

AUXILIARY STRUCTURES

Spillways. Serious consequences follow if a dam is overtopped. Disaster is likely in the case of an embankment dam not designed to permit uncontrolled flow of water on its downstream slope. In March 1960 the partially completed embankment dam at Orós, Brazil, was accidentally overtopped during a period of unexpectedly heavy rainfall. Despite heroic efforts to avert disaster, the

water level rose nearly one metre above crest level, eroded about half the fill in the dam, and cut a deep breach about 200 metres (660 feet) wide in the structure. Although there was time to evacuate 100,000 people living downstream, half were subsequently rendered homeless and about 50 perished. Spilling over a concrete gravity dam is also serious since the floodwater would erode the foundations at the downstream toe. Arch dams possess greater resistance to failure after overtopping.

Flood hydrology is a difficult subject. Much effort is being made to establish relationships between rainfall and river discharge, both in quantity and in time lag. Such statistical methods cannot estimate the maximum possible flood. At best they indicate only the probability of a specified flow being exceeded in a particular period. In constructing the Kariba Dam on the Zambezi, analyses of the available records of river discharge yielded the estimate that a flood of 7,600 cubic metres per second should be expected once in four years. During the first year of construction on the riverbed a flood of 8,500 cubic metres per second was experienced, and in the second year the Zambezi discharged 16,200 cubic metres per second.

In these circumstances civil engineers attach much importance to the design of spillways on dams. Inadequate spillway capacity caused failure of many older earth dams built before modern flood data became available.

Four general aspects of spillways should be emphasized. First, the uncontrolled discharge of surplus water past the dam should be automatic and, like a safety valve on a steam boiler, not under human control. Second, the spillway intake should be wide enough so that the largest floods can pass without increasing the water level in the reservoir enough to cause a nuisance to riparian inhabitants and property upstream. Third, the rate of floodwater discharge should not increase much above that experienced before the construction of the dam. An increase creates a flood nuisance downstream. A dam usually reduces the peak discharge rate due to the lag effect caused by a flood passing through the reservoir. Fourth, floodwater discharged over the height of a dam can be destructive to the dam structure itself and to the riverbed unless its energy is controlled and dissipated in harmless turbulence.

With embankment dams, a separate spillway structure is normally constructed to one side of the dam itself. With concrete gravity dams the sloping downstream face of the structure serves as the spillway. The water travels at very high speeds (about 160 kilometres [100 miles] per hour in the case of a dam 100 metres [328 feet] high), forms a standing wave where it enters the riverbed and proceeds downstream at lower mean velocity, but in a highly turbulent state. Grand Coulee Dam utilizes a spillway of this type. An obstruction known as a kicker, placed at the toe of the dam to project the water slightly upward, can move further downstream the area in which erosion of the riverbed is most intense. With higher dams, it is possible to deflect the jet of spilling water from a level above the base of the dam; this is known as a ski-jump spillway.

Ski-jump spillway

Spillways need not be open to the atmosphere. Shaft and tunnel spillways can destroy the energy of the water at a predetermined point downstream of the dam. At the upstream end, the intake can be self-priming siphons or bell-mouthed drop shafts; the latter are also known as morning-glory spillways.

With arch dams it is convenient to construct gated openings in the shell structure at some distance below the crest of the dam, ensuring that the discharging jets fall well clear downstream. A line of six such gates is used in the design of Kariba Dam.

Spillways constructed to one side of earth dams are featured in the design of Oroville Dam, California, and of Mangla in Pakistan. The spillway at Mangla discharges 28,000 cubic metres of water per second; the upper stilling basin has the dimensions of an Olympic Games stadium with its grandstands.

Rockfill dams, specially designed to be overtopped in times of flooding, are a new development. The first such permanent dams have been constructed in Australia. For temporary works the technique has been used on the Blue Nile.

Gates, sluices, and outlets. In addition to spillways, openings through dams are also required for drawing off water for irrigation and water supply, for ensuring a minimum flow in the river for riparian interests downstream, for generating power, and for evacuating water and silt from the reservoir. These gated openings normally are fitted with coarse screens at the upstream ends to prevent entry of floating and submerged debris. Provision for cleaning these screens is essential.

Several forms of gates have been developed. The simplest and oldest form is a vertical-lift gate that, sliding or rolling against guides, can be raised to allow water to flow underneath. Radial or tainter gates are similar in principle but are curved in vertical section better to resist water pressure. Tilting gates consist of flaps held by hinges along their lower edges that permit water to flow over the top when they are lowered.

Drum gates can control the reservoir level upstream to precise levels, automatically, and without assistance of mechanical power. One drum gate design consists of a shaped steel caisson held in position by hinges mounted on the crest of the dam and supported in a flotation chamber constructed immediately downstream of the crest. Water pressure in the reservoir and buoyancy of the caisson in the flotation chamber hold the caisson in rotational equilibrium. Raising or lowering the water level in the flotation chamber causes the caisson to rotate in the same direction, thus reducing or increasing flow from the reservoir over the gate. This action can be linked to and operated automatically by a float control device in the reservoir. Two drum gates are installed at Pitlochry Dam in Scotland.

Reservoirs. Modern engineers have learned the value of giving attention early to potential problems in reservoir maintenance. Sediment in rivers seriously influences the effective life of a reservoir and therefore the financing of a dam. Some modern dams have been rendered useless for storing water because the reservoir has filled with silt. In many others effective storage capacity has been seriously reduced. At the Nile barrages, the heavy silt-laden floodwater is allowed to pass through the sluices and only the cleaner water at the end of the flood season is stored.

Fish passes. For many years hydroelectric dam design has taken into account the need to conserve certain species of migratory fish. Success has been achieved with salmon in Scotland and on certain rivers in the U.S. and Canada. Notable examples of conservation measures are to be found at Bonneville, Priest Rapids, and Wanapum dams and at many dams in Scotland.

Adult salmon striving to reach their spawning grounds upstream must be prevented by screens from entering the turbine tailraces at power stations, and induced instead to enter a fish pass that allows them to surmount the dam. Similarly, young salmon must be allowed to pass a dam safely on their journey downstream to feeding grounds in the ocean. Young salmon are remarkably insensitive to sudden changes of pressure and have been known to pass safely through turbines operating at heads of up to 50 metres (160 feet). Nevertheless, it is preferable to induce them to use the fish passes.

Fish passes usually take the form of fish ladders and fish locks. A fish ladder is utilized at Pitlochry Dam in Scotland; it consists of a series of stepped pools, through which water is continuously discharged during the migratory seasons. The individual pools may be separated by a series of low weirs or linked by short inclined underwater pipes to provide the necessary steps of 0.3 to 0.6 metres in water levels. Sometimes both weirs and pipes are provided.

Fish ladder

The Borland fish lock was developed in Scotland as an alternative to fish ladders. It operates on the same intermittent principle as a ship lock but is constructed as a closed conduit. Intermittent closure of the gates at the bottom causes the continuous flow through the lock to

fill the conduit at intervals, and thus allows fish waiting in the bottom chamber to be raised through the height of the dam. The lock also serves at other seasons to flush young salmon down past the dam.

BIBLIOGRAPHY. E. WEGMANN, *The Design and Construction of Dams*, 8th ed. (1927), a general textbook on dams containing much information of historical interest; W.P. CREAGER, J.D. JUSTIN, and J. HINDS, *Engineering for Dams*, 3 vol. (1945), a textbook dealing principally, but not exclusively, with American practice; J. HINDS, "Continuous Development of Dams Since 1850," *Trans. Am. Soc. Civ. Engrs.*, CT:489–520 (1953), a history of the development of dam design and construction over 100 years giving some selected examples to illustrate design principles; J. GUTHRIE BROWN (ed.), *Hydro-Electric Engineering Practice*, 3 vol., 2nd rev. ed. (1964), a comprehensive modern textbook dealing principally with British and European practice; G.F. SOWERS and H.L. SALLY, *Earth and Rockfill Dam Engineering* (1962), on the design and construction of embankment dams with strong emphasis on soil mechanics problems; J.L. SHERARD *et al.*, *Earth and Earth-Rock Dams* (1963), on the design and construction of foundations and embankments after site selection; C.V. DAVIS (ed.), *Handbook of Applied Hydraulics*, 3rd ed. (1969), a classic work on the basic principles of hydraulic engineering and the design of hydraulic structure; AMERICAN SOCIETY OF CIVIL ENGINEERS, *Symposium on Arch Dams* (1957), *Symposium on Rockfill Dams* (1958), collections of papers, mostly case histories, by eminent engineers from many countries on the design and construction of arch dams and rockfill dams; INTERNATIONAL COMMISSION ON LARGE DAMS, *World Register of Dams*, 4 vol. (1963, updated to December 1968), a register listing basic statistical data about large dams in countries that are members of the commission; U.S. COMMITTEE OF THE INTERNATIONAL COMMISSION ON LARGE DAMS, *Register of Dams in the United States* (1963), basic statistical details of dams in the U.S., with some photographs; CANADIAN NATIONAL COMMITTEE OF THE INTERNATIONAL COMMISSION ON LARGE DAMS, *Register of Dams in Canada* (1964), basic statistical details of Canadian dams, with some photographs and a useful bibliography; ACADEMY OF SCIENCES OF THE GEORGIAN S.S.R., INSTITUTE OF POWER ENGINEERING, *The High Dams of the World* (Eng. trans. 1967), basic statistical details of world dams, with an extensive bibliography; NORMAN SMITH, *A History of Dams* (1971), an outstanding detailed record of ancient dams.

(J.G.B.)

Damascus

Damascus (in Arabic, formally Dimashq, colloquially ash-Shām), the capital of Syria, is one of its two largest cities, the other being Aleppo. Situated at the base of Jabal Qāsiyūn (Qāsiyūn Mountain), among the orchards of al-Ghūṭah oasis, it has been called "the pearl of the east," "the city of many pillars," and "the gate of Mecca." It is probably the oldest city in the world.

History. The historical origins of Damascus are unknown, but excavations in the court of the Umayyad Mosque have unearthed potteries belonging to the 3rd millennium BC. The earliest historical reference to it is found in the hieroglyphic tablets of Tell el-Amarna in Egypt, where "Dimashqa" is listed as one of the cities that were conquered during the 15th century BC by Thutmose III, who subdued all Syria after defeating its Aramaean kings. The Aramaeans had established a number of kingdoms in Syria, of which Damascus was the most famous and powerful. Damascus had many conquerors, including King David of Israel, the Assyrians under Tiglath-pileser III (732 BC), and the Chaldeans under Nebuchadrezzar II (around 604 BC). The Persians conquered it around 530 BC under Cambyses; the kings of Damascus continued to reign under the protection of the Persian Empire until 333 BC, when Syria was conquered by Alexander the Great.

The ancient city

The year 333 BC marked the beginning of the Hellenistic Age, during which Damascus came under the rule of the Seleucids, the Romans, and the Byzantines, in succession. Around 90 BC many Greeks came to live in it, thus beginning an important cultural contact between the Aramaean and the Hellenistic civilizations. The native population assimilated the language, arts, and beliefs of the Greek community. A Greek city eventually sprang up close to Aramaean Damascus, and remains of it still exist in the form of a porched street east of the Umayyad Mosque and an agora that used to occupy the centre of the city. Greek influence began to dwindle in 85 BC, when the city was occupied by the Nabataeans under Aretas III. The Nabataeans established themselves in a quarter of their own, east of the Greek. Western influence made a second return to Damascus in 64 BC, when it was captured by the Romans, who declared Syria a Roman province. The Nabataeans were allowed to reoccupy the city in 37 AD, and they remained there until 54 AD. It was during this period that Saul of Tarsus came to Damascus in pursuit of Christians and was miraculously converted to become St. Paul the Apostle. A chapel relating to his visit, known as Ananias' Chapel, is still in good preservation.

Damascus profited considerably from the Pax Romana. Under the emperor Diocletian it became an important military centre for the armies of the empire in their wars with the Persians. It was given the status of a metropolis under Hadrian (2nd century) and became a Roman colony under Alexander Severus (222–235). The advancement of the city allowed it to play an active role in Roman life. Its celebrated architect Apollodorus of Damascus built the famous stone bridge over the Danube, the Trajan Forum in Rome, and many public buildings, baths, and triumphal arches. Damascus flourished as a trade centre on the desert caravan routes. The Romans took special care to fortify the city, enclosing it with a rectangular-shaped wall, of which some of the remains have been preserved. At the northwest corner was a castle that has since been restored and is known as the Citadel. Some of the Roman aqueducts were in use until the early years of the 20th century.

With the division of the Roman Empire toward the end of the 4th century, Damascus became an important military outpost of the Byzantine Empire. Most of the population adopted Christianity. The Temple of Jupiter was turned into a cathedral dedicated to St. John the Baptist, whose head is said to lie in a crypt within the Umayyad Mosque.

Damascus under Islām

In 635 Damascus fell to the Arabs, and from 661 to 750 it was the capital of the Islāmic empire. During this period, the city began to spread outside its walls. The Christian cathedral was purchased by the caliphate and turned into a mosque, which was restored in the 11th, 15th, and 19th centuries and is now known as the Umayyad Mosque. With the fall of the Umayyad dynasty in 750, insurgent armies ravaged the city and destroyed all within it that they considered to stand for Umayyad dynasticism: buildings, palaces, fortifications, and even cemeteries. The new ʿAbbāsid dynasty removed the seat of the empire from Damascus to Baghdad. As the unity of the caliphate dissolved and local dynasties arose, Damascus changed hands often; but for most of the 200 years after the middle of the 9th century it was held by those who ruled Egypt. According to the 12th-century historian ʿAbd ar-Raḥmān ibn al-Jawzī, the population of Damascus was reduced to 5,000 in the year 1075. Only two bakeries were left out of 240, the marketplaces were empty, and a house that normally cost 3,000 dinars failed to find a buyer willing to pay 10 for it.

In 1076 Damascus was occupied by the Seljuq Turks. It became the capital of a province extending eastward to the Euphrates, westward to Tiberias, and northward to Homs. During the Second Crusade it was besieged unsuccessfully in 1148 by Conrad III of Germany and Louis VII of France. It was finally taken by Nureddin (Nūr ad-Dīn), the Egyptian general, who entered it in 1154 and united Syria and Egypt in 1168. By this time the city plan of Damascus had become chaotic. The Muslims had the western part of the city around the citadel, as well as the Umayyad Mosque, the Christians had the northeastern, and the Jews, the southeastern parts. Under Nureddin, Damascus regained its status as a capital. His successor, Saladin, reunited Syria with Egypt. Many remains in Damascus date back to this period, among them Bāb-al-Faraj in the northern part of the city wall, and the Ṣāliḥīyah quarter at the foot of Jabal Qāsiyūn (started in 1161). After a brief occupation by the Mongols, Damascus was included in the Mamlūk

state of Egypt and Syria in 1260. This period saw a good measure of progress in the sciences, arts, and industries. Damascene goods acquired a world reputation, attracting merchants from Venice and Genoa. The sultans regarded Damascus as their second capital, had palaces built for them, and occasionally lived there. The administrative province of the city extended to the Euphrates in the east, to the Mediterranean in the west, and to the river Rastan in the north. The Mamlūk period was interrupted by a tragic interlude when the Mongols under Timur levelled Damascus and deported its scholars and artisans to their capital in Samarkand. The Mamlūks returned the following year and set out to reconstruct the city.

In 1516 Damascus was captured by the Ottoman sultan Selim I and remained the capital of a Turkish province for 400 years. Religion apart, there was little in common between the Arabs and their Ottoman rulers. Damascus was reduced to the status of a small *paşalik* within an immense empire run by a powerful central authority. The governors of the city succeeded one another at such a rate that there were 133 of them between 1517 and 1697. Economic life suffered under their exactions, and trade was further diminished by the discovery of an alternative commercial route between Europe and the east toward the end of the 15th century. But Damascus still traded with the rest of Syria and with neighbouring countries, such as Iraq, Anatolia, Persia, and Hejaz; and French and Venetian commercial agents brought their silks, woollens, sugar, and gold in exchange for raw materials and spices. Damascus also became the point of departure of the pilgrim caravan to Mecca under the governor's protection and patronage. A new quarter grew up on the pilgrimage route in the south of Damascus, called as-Sinānīyah after the Turkish governor Sinān Pasha.

Entry of the West

One of the most important events in the 19th century was the coming to power in Egypt of Muḥammad ʿAlī Pasha and the incorporation of Syria in his empire. During the Egyptian period (1831–40) Damascus once again became the capital of Syria. After 1878 its population grew to 150,000. Europeans came in increasing numbers—consuls, merchants, and tourists. The imperialistic ventures of the great powers in the Middle East opened the door wider to European cultural and economic influence.

In 1878 an enlightened Turkish governor, Midhat Paşa, was appointed for Damascus. He succeeded in improving its sanitary conditions and its system of transport. He dug new roads in the Old City and widened some of its dark alleys, such as that connecting Sūq al-Ḥamīdīyah with the Great Umayyad Mosque. New government buildings rose up west of the Old City around al-Marjah Square (the Meadow), which has become the city centre. In 1894 Damascus was linked by railway to Beirut and Ḥawrān and in 1908 to Medina in the Hejaz.

Two new quarters of Damascus were added as Muslim immigrants arrived from Crete, settling on the slopes of Jabal Qāsiyūn in what has come to be known as al-Muhājirīn, the immigrants' quarter, and as Kurds coming in from the north found space just to the east of al-Muhājirīn in al-Akrad, the Kurds' quarter. New buildings were put up near Bāb Tūmah (Thomas' Gate) for European consuls, merchants, and missionaries.

In World War I Damascus was the headquarters of Turkish–German forces. The Arab Legion entered Damascus in October 1918, followed by the Allies. The first Syrian government was set up in 1919, with Damascus as its capital. But the French, having obtained a mandate from the League of Nations, occupied the city in July 1920. The French presence (1920–45) encountered much resistance in the form of strikes, riots, and uprisings, the most serious of which occurred in October 1925 and was not suppressed before the French had twice bombarded the city.

The period of the French mandate was one of great urban development, often at the expense of the orchards on the left bank of the Baradā River. The new quarters of al-Jisr, al-ʿArnūs, and ash-Shuhadā' accommodated various elements of the population without any apparent segregation. Northeast of the city the quarter of Qaṣṣāʿ sprang up adjacent to Bāb Tūmah and the eastern city wall. Wide, tree-lined avenues were laid out, and the city was provided with a modern water supply.

During World War II Allied troops entered the city in June 1941, driving out the Vichy French, who were pro-German. After four years of tension over the issue of independence, British and French troops were evacuated in April 1946, and Syria became an independent country with Damascus as its capital.

The city and its environment. The early site of Damascus was a terrace on the right bank of the Baradā River at the point where it emerges from its canyon before losing itself in the desert. The terrace, consisting of compact gravels, has an elevation of about 2,250 feet above sea level. The growth of the city has forced it up the slopes of Jabal Qāsiyūn, northwest of the original terrace, to an elevation of almost 3,000 feet. The main expansion of the city, however, has been to the northeast and south, where the terrain is gentler and the elevation does not exceed 2,300 feet.

Topography

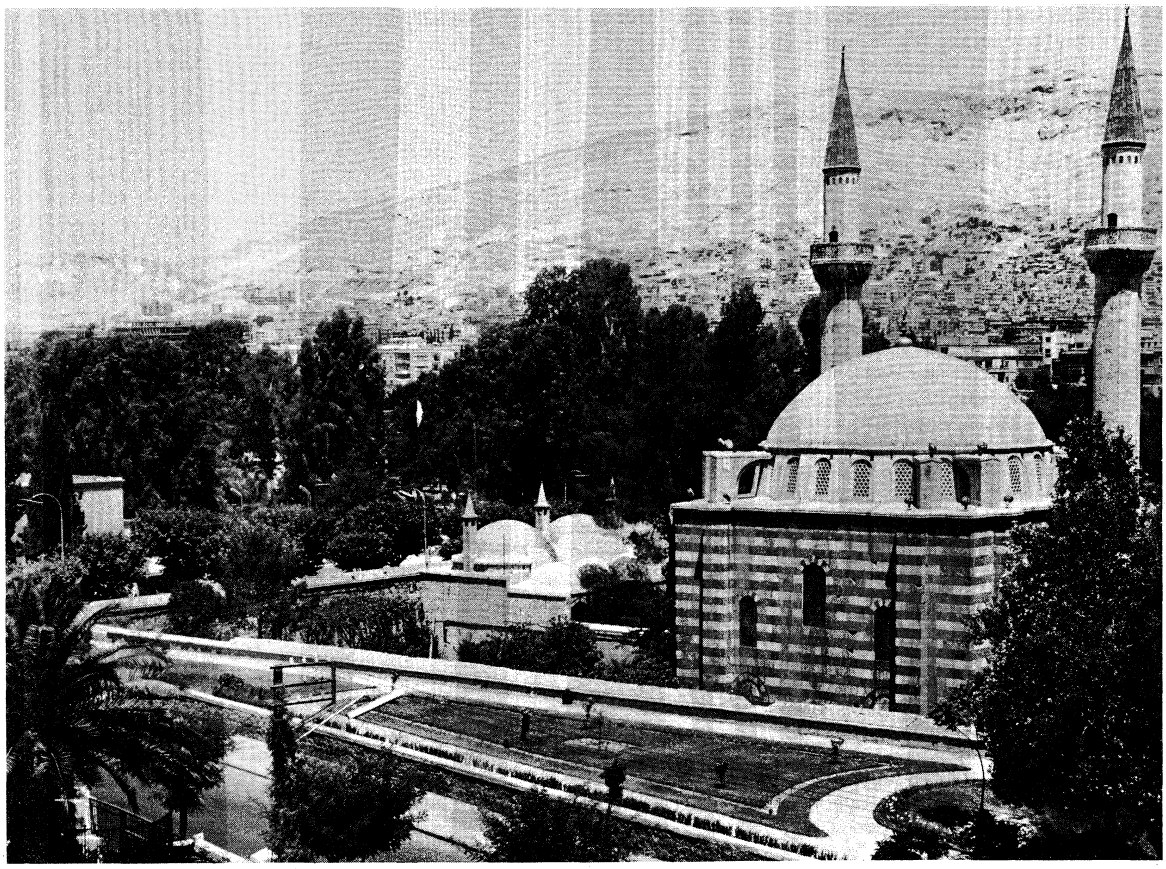
Damascus is open in all directions except the west, which is blocked by Jabal Qāsiyūn. The fertile alluvial plain extending 12 miles eastward supplies the city with vegetables, fruit, timber, and some cereals. The principal cereal requirements are met from the plains of Ḥawrān to the south.

Greater Damascus covers an area of about 39 square miles (100 square kilometres), including all of Jabal Qāsiyūn, Mt. Mazzah, and the townships of Dummar, Barzah, al-Qābūn, Jawbar, al-Qadam, al-Mazzah, and Kafr Sūsah; *i.e.*, about five times the area of the city proper, which is eight or nine square miles. The city has sent out tentacles in all directions except the west. The most important of these elongations is that extending along the slopes of Jabal Qāsiyūn to the villages of Barzah in the northeast and al-Mazzah in the southwest; a second extends northeast to Jawbar and al-Qābūn; the third and oldest has already swallowed up the village of al-Qadam. Al-Mazzah, al-Qābūn, and Barzah are considered suburban areas because they are still separated from Damascus by green areas that the municipal government is anxious to preserve.

Climate. Although Damascus is only 50 miles from the sea, the Lebanon and Anti-Lebanon mountain ranges cut off the moisture-laden winds and make the climate a subdesert one. The mean annual precipitation is 6.89 inches, as compared with a mean of 33.15 inches in Beirut, Lebanon, on the other side of the mountains. Most of the precipitation is in November, December, January, and February. In the dry months, desert winds blow from the south and east, bringing dust with them. Variations of temperature are sudden and capricious. The January average is often about 45° F (7° C) but may occasionally fall to 18° F (−8° C) and rise to 72° F (22° C). August is the hottest month in Damascus, with an average of 81° F (27° C), but the temperature often rises to 99° F (37° C) and may reach 111° F (44° C). Summer is the longest season. Spring lasts a few weeks between the second half of March and the first half of April, while autumn may not come until November. Variations in rainfall are extreme. In some years it may reach as much as 13 or 14 inches, which is sufficient for dry farming, while in others it may fall below six inches or even as low as two inches or less, as it did in 1931–33 and 1959–60.

Plant and animal life. The desert surrounding the oasis has a variety of dry and thorny shrubs, thistles, and herbs. The Ghūṭah, however, has a temperate and subtropical vegetation, including such trees and plants as the mulberry, apricot, olive, vine, apple, peach, and nut; aspens and poplars grow thickly near the streams. Truck gardening is intensive, and the fringes of the Ghūṭah contain cereals such as wheat, barley, and lentils grown in the shade of olive trees.

Animals have been driven out by man. Gazelles, wolves, and hares have disappeared. The drying up of marshes has eliminated many migratory birds, such as ducks, water hens, and storks. Cows, goats, and sheep are bred



The Citadel of Damascus, with Jabal Qāsiyūn in the background.
Carl Frank

in increasing numbers to meet the milk and meat requirements of the city, while transport animals—horses, mules, and donkeys—are rapidly decreasing.

The changing city

City plan. A modern city plan was laid out for Damascus in 1929 by a French expert, providing for its orderly growth. A number of wide streets and avenues were built subsequently, notably in the modern residential quarter of Abū Rummānah, which sprang up at the expense of the old village of Nayrab. This quarter, which has many embassies, has continued to extend westward and northward, threatening to engulf all the orchards in the northwestern parts of the city. Other residential quarters have also sprung up north of Baghdad Street, spreading toward the al-Akrad quarter on the slopes of Jabal Qāsiyūn. The city's accelerated growth since World War II has required modifications in the plan, which were made by a Japanese planner in 1964.

Most of the streets in Damascus are straight, and they cross at right angles. There are some diagonals. Widening operations have been carried out in the Old City to give it better ventilation and open it to modern transport, but its most prominent features remain. A typical house in the Old City opens onto a yard lined with trees and containing a water fountain in the centre. Damascus is zoned into residential, commercial, and industrial areas, the latter being located to the east of the city. Governmental ministries and departments are scattered about in various locations west of the Old City. Many are concentrated at Marjah Square, the area of banks and fashionable shops.

Streetcar lines were built in Damascus in 1906, but between 1963 and 1970 they were discontinued. Intracity transport is provided by buses and taxis. Traffic lights have been introduced since 1966. The total number of motor vehicles in 1970 was about 20,000, including motorcycles. Damascus has an international airport about 19 miles east of the city.

Demography. The city's population in 1971 was more than 1,000,000, as compared with 130,000 in 1922 and 529,963 in 1960. The present figure includes 50,000 refugees from Palestine since 1948, and 75,000 from the Golan Heights since June 1967. Muslims constituted more than 91.2 percent of the population, Christians 8.4 percent, and Jews 0.04 percent. The natural increase was estimated at the very high rate of 28 per thousand in 1968, with a birth rate of 40 per thousand and a death rate of 12 per thousand. Those below 15 years of age comprised 42.8 percent of the population; only 11.2 percent were over 50. There were 313 families with 10 children or more, and some had as many as 17.

The traditional Damascene house with its patio and water fountain is no longer being built. Cement, which was not introduced as a building material until 1918, has brought an architectural revolution that has completely transformed the city. European design has become the rule, usually in the French and Italian Mediterranean style. Reinforced concrete is the principal material of residential buildings. Their upper stories are often uninhabitable in the heat of summer.

Major industries

Economic and cultural life. Damascus is a centre of government and light industry. Once celebrated for its fine silken textiles and ornamental fabrics, as well as for steel and swords, it still has a fairly active and prospering group of handicraft industries. Notable among their products are mother-of-pearl mosaics, silk brocades, copper goods, blown glass, and cabinetry. The more modern industries include furniture, clothing, shoes, textiles, food, leather, and printing. These industries started to emerge in 1929 when the first wool-cloth factory was built, followed by a cement factory in 1930 and a food-preserving factory in 1935. The cotton textile industry began in 1946, although factories using imported yarns date back to 1937. Damascus now has 80 textile establishments with more than 1,800 modern looms. The first glass factory was completed in 1945. Most of the major factories are now state owned.

Damascus plays an important role in transit trade between Beirut, Amman, Baghdad, Kuwait, and the countries of the Arabian Peninsula. It is also a distributor of imported goods to various parts of central and southern

Syria. Wholesale trade is now largely a state monopoly. An international trade fair is held at the river every autumn. The fair is visited by about 1,500,000 people. Damascus has an excessive number of retail shops and small tradesmen, attributable to a lack of regular employment in other fields. There is a surplus of manpower because of heavy migration to the city from the countryside.

Public utilities. Damascus draws its water supply from the Baradā River's source at ʿAyn al-Fījah through underground canals. The system, initiated by the Turkish governor Nāẓim Pasha in 1908, was completed in 1932. Electric power consumption in 1968 was 206,109,000 kilowatt-hours, of which one-fourth went to industry and over one-half to illumination. Sewage is discharged into the lower Baradā River or into the Yazīd and Tawrāt streams that issue from the Baradā.

Public health. In 1969 Damascus had 10 state hospitals and sanatoria, with 2,676 beds, and 17 private hospitals, with 403 beds, affording one bed for every 421 people. There were 61 state and private outpatient clinics, 13 maternal and child-welfare centres, one centre for tuberculosis control and another for malaria eradication, two schools of nursing and midwifery, and 700 physicians—45 percent of all the country's doctors. Endemic diseases are generally absent from Damascus with the exception of tuberculosis, which attacks the poor.

Education. Damascus had 385 public elementary schools in 1968 (150 for boys, 120 for girls, and 115 mixed) along with 80 private schools and 50 schools run by the United Nations for Palestine refugees. The number of teachers in all these was about 4,000. Preparatory and secondary education was provided by 60 state, 55 private, and 18 United Nations' schools employing about 2,000 teachers. There were four industrial schools with 3,745 pupils, a domestic science school with 267 girls, two commerce schools, and one agricultural school. The University of Damascus had about 31,000 students, about half of whom were registered in the faculty of literature. Illiteracy among persons over the age of 15 was about 20 percent.

Culture and recreation. Most formal cultural activities in Damascus are directed and controlled by the Ministry of Culture. The National Museum, founded in 1921, superintends the excavations carried out by foreign archaeological missions (Italian, French, American, Danish, and Polish) in various parts of the country. The Qaṣr al-ʿAẓm Museum, close to the Great Umayyad Mosque, specializes in the history of folklore and popular arts. The Arab Academy, founded in 1919 by the scholar Muḥammad Kurd ʿAlī, contains the aẓ-Ẓāhirīyah national library, with 40,000 manuscripts and 100,000 publications.

Mass media

All mass media are controlled by the Ministry of Information. Syria's two national daily newspapers, *al-Baʿth* and *ath-Thawrah*, are published in Damascus, as well as two weeklies—one for workers and the other for farmers and peasants. The Damascus radio broadcasts 14 hours daily in Arabic, English, French, Turkish, Russian, German, Hebrew, Armenian, and Spanish. There are also daily television broadcasts.

The cinema is one of the most popular amusements among Damascenes, who are presented with a choice of films from Arab countries, the United States, Europe, India, and Pakistan. The first cinema theatre in Damascus was built in 1925; in 1971 there were 13 of them. Damascus has night clubs and numerous discotheques. The annual international fair, held August 25 to September 20, offers a great variety of recreation and entertainment.

There are some public gardens and parks, most of them fairly small. The great orchards of the Ghūṭah are open to the public, but there is no zoological garden.

The most popular sport in Damascus is association football, followed by basketball and swimming. The main stadium accommodates 50,000 spectators. Damascus also has a number of excellent swimming pools in its suburbs. Other sports are wrestling, boxing, and tennis.

BIBLIOGRAPHY

History: PHILIP K. HITTI, *History of Syria, Including Lebanon and Palestine* (1951); J.L. PORTER, *Five Years in Damascus* (1855); NICOLA A. ZIADEH, *Damascus Under the Mamlūks* (1964).

Climate and geography: CHARLES COMBIER, *Apérçu sur les climats de la Syrie et du Liban* (1945); RENE DUSSAUD, *Topographie historique de la Syrie antique et médiévale* (1927); RICHARD THOUMIN, *Géographie humaine de la Syrie Centrale* (1936).

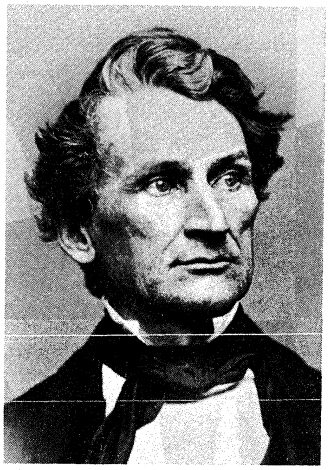
General: M. ECOCHARD and G. BENSHOYA, *Ville de Damas, études preliminaires au plan d'urbanisme* (1946); N. ELISSEEFF in the *Encyclopaedia of Islam*, new ed., vol. 2 (1965); M. FONTAINE, *Ville de Damas: alimentation en eau* (1963); JEAN SAUVAGET, *Les Monuments historiques de Damas* (1932).

(A.-R.H.)

Dana, James Dwight

A versatile naturalist and an eminent student of the Earth, James Dwight Dana shaped his country's geological thought for over half a century. Explorer, investigator, writer, editor, and teacher, Dana's productive career spanned 60 years, from his first publication in 1835, a study of Vesuvius, to the 4th edition of his *Manual of Geology* in 1895. A brilliant synthesizer and leading theorist of geology, he also made permanent contributions to the subjects of mineralogy and zoology.

By courtesy of the Yale University Library



Dana.

Born in Utica, New York, February 12, 1813, Dana attended Charles Bartlett's Academy, then entered Yale College as a sophomore in 1830. On graduation from Yale in 1833 he instructed midshipmen in mathematics on a United States Navy cruise to the Mediterranean; he returned to New Haven in 1836 as an assistant to his former teacher, Benjamin Silliman, professor of chemistry and mineralogy at Yale. Evidence of Dana's great productive energy came at age 24 with the publication in 1837 of *A System of Mineralogy*, a work of 580 pages that has persisted through numerous editions and remains the standard manual in its field. In 1838 Dana joined a United States exploring expedition to the South Seas under Charles Wilkes; he served four years as a geologist and also was responsible for much of the zoological work.

The Wilkes expedition

In 1844, two years after his return from the expedition headed by Charles Wilkes, Dana married Henrietta Silliman, the daughter of his mentor at Yale, and settled in New Haven. Dana spent his intense energy largely on science. From 1844 to 1854, his most productive years, he published about 7,000 printed pages in addition to hundreds of plates, most of which he drew. His writings on the Wilkes expedition include four illustrated quarto volumes, each over 700 pages—the *Zoophytes* (1846) describing the coral-forming animals and a two-volume work, *Crustacea* (1852–54) and the report on the *Geology* in 1849. Numerous short papers treated other aspects of his work with the expedition. The zoological work alone, in which he described more than 800 species of marine animals, was monumental and made Dana an authority on the groups.

The main thrust of Dana's effort was geological. Among

his many publications were the text *Manual of Mineralogy* (1848) and three editions of *A System of Mineralogy* (1st edition, 1837), including a complete revision in which he founded a classification of minerals based on mathematics, physics, and chemistry. More significant to Dana's impact on American geology during this decade was the start of his long association with the *American Journal of Science*, a leading organ of scientific inquiry founded by Benjamin Silliman. As an editor and contributor of critical reviews, original papers, and perceptive syntheses, Dana exerted a vitalizing influence on American geology. One of the best informed men of science in his day, his concern with the physical processes that produced geologic phenomena led to brilliant generalizations on such fundamental questions concerning the formation of the physical features of the Earth as the origin and structure of continents and ocean basins, the nature of mountain building, and volcanic activity. From his own studies and his mastery of the works of other American and foreign geologists, Dana constructed a view of the Earth as a geological unity developing through time. Adopting the theory of a contracting Earth cooling from a molten condition, he argued that the present continents mark areas that cooled first; subsequent contractions caused the intervening oceanic areas to subside. As settling oceanic crusts adapted periodically to a shrinking interior, pressure was exerted against continental margins causing upheaval of great mountain chains such as the Appalachians, Rockies, and Andes. Dana stressed the progressive change of the Earth's physical features but at first was reluctant to accept the idea of the evolution of living things.

Studies of oceanic crusts

By the early 1850s Dana had attained international recognition and was corresponding with other outstanding scientists of his day, among them Asa Gray, noted American botanist; Louis Agassiz, Swiss-born naturalist and teacher at Harvard; and Charles Darwin. All had measurable influence on his thinking. Spurred by tentative proposals from Harvard for Dana's services, friends at Yale established the Silliman Professorship of Natural History, which Dana accepted in 1856. But in 1859, the strain of his self-imposed overwork resulted in a physical breakdown from which he never fully recovered. During his remaining 35 years he was forced to live a secluded existence, largely withdrawn from the public. For a less modest man this would have been difficult, for many academic honours came to him during this period. Recognition included the presidencies of the American Association for the Advancement of Science and the Geological Society of America; he was also a founding member of the National Academy of Sciences.

Despite ill health Dana continued to publish: in 1862 his influential textbook, *Manual of Geology* (4 eds.); in 1864, *A Text-book of Geology*, a more elementary work; and in 1872 *Corals and Coral Islands*, which was the climax to his notable studies of coral reefs, begun on the Wilkes expedition. Dana investigated coral islands in greater detail than anyone before him, substantiating Darwin's observation that atolls were evidence of subsidence of the ocean bottom. Reef-building corals, Dana independently concluded, lived only in shallow, tropical waters on hard substrates, commonly forming fringing reefs about volcanic islands. Coral rock found at some depth on island flanks and atolls made only of reef rock indicated that extensive volcanic lands had disappeared beneath the Pacific leaving clusters of atolls to mark their former existence.

During his later years he wrestled with the challenge of organic evolution proposed by Darwin. Always a deeply religious man, Dana believed in the special creation of species, yet he was keenly aware of the intricate relationships between species and their environment. Darwin's impressive argument, coupled with Dana's own zoological knowledge, was persuasive in the end, however, and he adopted evolution theory in the last edition of his *Manual*. For Dana the natural and the divine had to be inseparable—all nature and the design of continual improvement of life that he read into it were a manifestation of the divine.

During Dana's lifetime, and largely under his leadership, American geology grew from a collection and classification of unrelated facts into a mature science.

Actively at work until the last, Dana died on April 14, 1895, in New Haven at the age of 82.

BIBLIOGRAPHY. D.C. GILMAN, *The Life of James Dwight Dana* (1899), a comprehensive account of Dana's life and work; L.V. PIRSSON, "Biographical Memoir of James Dwight Dana, 1813–1895," *Biogr. Mem. Natn. Acad. Sci.*, 9:41–92 (1919), a summary of Dana's scientific work, with references to other short biographical notices; WILLIAM STANTON, "Dana, James Dwight," *Dictionary of Scientific Biography*, vol. 3, pp. 549–554 (1971).

(K.M.W.)

Dance, Art of

The art of dancing is the art of moving the body in a rhythmical way, usually to music, to express an emotion or idea, to narrate a story, or simply to take delight in the movement itself. The dance impulse seems, so far as is known—for there is no concrete evidence—to date from the very beginning of man's existence on Earth and to begin with his own life. Dance may even have been the first means of communication; children today, even in sophisticated societies, will dance with delight before they know how to express their feelings in words.

Dance as an art, as distinct from a natural activity, can be traced from the moment when it was harnessed to a rhythm, probably the stamping of feet and clapping of hands. It is, however, an intangible art, existing in the bodies of the performers and dying with them. Only in recent years has it been possible to preserve on film or in precise notation the actions of the dancers.

There is evidence, throughout recorded history, of men dancing, but the evidence is static—in drawings, paintings, or sculpture. It is only possible to guess how the dancers portrayed may have moved from one pose to another. Not until the 15th century is there any written description of actual dances, and, although scholars have attempted to reconstruct the social dances of the 13th and 14th centuries, they are still reluctant to say more than "they may have danced in this manner." Yet, it is possible to trace in broad outline the origins and functions of dancing, its components, and its development in various ways of life and also to describe the kinds of dancing that can be seen today.

Since this article comprises a general survey of all forms of dancing, other articles should be consulted for greater detail on some of the major forms: *e.g.*, BALLET; FOLK DANCE; MODERN DANCE; and POPULAR DANCE. Various kinds of interaction between dance and drama are covered in such articles as KABUKI THEATRE; MUSIC, THEATRICAL; NO THEATRE; and POPULAR THEATRE. The article DANCE, WESTERN covers the history of dance in the West, but, for information on the character and the history of dance in non-Western cultures, see DANCE AND THEATRE, EAST ASIAN, as well as sections on dance or the performing arts in such articles as AMERICAN INDIAN PEOPLES, ARTS OF; CENTRAL ASIAN PEOPLES, ARTS OF; SOUTH ASIAN PEOPLES, ARTS OF; ISLAMIC PEOPLES, ARTS OF; JEWISH PEOPLES, ARTS OF; and OCEANIAN PEOPLES, ARTS OF.

ORIGINS AND FUNCTIONS OF THE DANCE

Primitive beginnings. It is not known how primitive man danced, but it can be deduced from the behaviour of primitive tribes surviving today that the rhythms that spurred on the dancers came mostly from the beat sustained through the stamping of feet upon the ground. Footwork probably would not have been intricate, and rhythms would have been simple—a regular recurrence of beat or shouts as the dancers roused themselves to a frenzy of activity. The earliest dances may have been simple expressions of pleasure or acts related to courtship, a theory supported by the fact that birds and animals, even spiders, dance to attract and impress a mate. Soon, however, dance was used for purposes affecting a whole tribe, which would, so far as can be deduced, dance and enact beforehand the results they wished to achieve on some expedition—be it a hunting dance to

depict the capture of prey or a warlike dance to show the defeat of an enemy.

There were surely dances of thanksgiving for harvest, for good weather, and for wanted rain. Rain dances have survived in parts of the world until today, and belief in "the rainmaker" persisted in parts of North America into the 20th century.

Emergence of religious motivations for dance

Supplication to the gods in early times was the beginning of the religious aspect of dancing. It has been used throughout history not only for glorification but also, before written books, as a means of communication between religious leaders and the people. Dances of worship continue today in every part of the world. In the East, temple dancers preserve something of the style and traditions of ancient cultures. In the West, the most sophisticated of professional dancers put their art at the service of religion, taking part in services of celebration in great cathedrals or humble churches.

Ritual dances of considerable complexity and beauty are believed to have been developed during the ancient civilizations of Egypt and Peru. Bas-reliefs of royal processions and religious ceremonies suggest a highly stylized kind of movement. Indeed, so far as anthropologists can establish, dancing has played an important role in the social life of all civilizations.

From antiquity to the Renaissance. It is certain that the Greeks considered the art of dancing an essential element in their ideal of creating a sound mind in a sound body. Plato demanded a place for the dance in his ideal republic, saying it would help toward acquiring "noble, harmonious and graceful attitudes." Such a statement implies that dance was already appreciated for its own sake as an expression of beauty of movement. This aspect of dance, as well as more grotesque and comical performances, was incorporated into the Greek theatre, which reached its highest achievement *c.* 500 BC. From that period the beginning of theatrical dance can be dated, but how the dancers performed can only be speculated upon: the attempts to revive the Greek dance by the U.S. dancer Isadora Duncan and by the British dancer Ruby Ginner are based on such conjecture.

The Romans, it is believed, developed earthier and more grotesque kinds of dancing in their plays and spectacles, bringing the theatrical form into disrepute. The social accomplishment of dancing, however, was recognized, and daughters of the nobility received dancing lessons. (There is no evidence of couples dancing together, however; that was to happen much later, probably in Provence in the 12th century.) The Romans also developed the art of mime, which was to lead, by the 16th century, to the improvised mime plays of the Italian commedia dell'arte.

Development of dancing as a social grace

The East. Whereas in the West various kinds of religious, social, and theatrical dance evolved slowly into the dances of today, quite different styles of movement were being perfected in the East. In general, Western dance techniques are based mostly on footwork, whereas Eastern forms of art dance, depending upon footwork to sustain the rhythm, derive their greatest beauty and subtle meanings from the infinite variety of delicate movements of the upper part of the body, especially the hands, neck, and head.

In the East as in the West, dance was indigenous to religious ceremony, was used for court entertainments, and, in a debased form, was practiced by courtesans. This debasing of the art has happened in all countries and all centuries. Because of its physical appeal, dance lends itself to erotic purposes and has been practiced to these ends by both sexes. It has, therefore, frequently incurred the displeasure of various ecclesiastical authorities who, though using dance for disseminating their own views, have condemned indulgence by the people.

The Renaissance. The domination of the Roman Catholic Church in Europe, after the fall of Rome, virtually extinguished theatrical dance, but it was reborn, together with the other arts, in Renaissance Italy. Emerging from the restrictions of the Middle Ages and intoxicated with a love of beauty and knowledge, the Italians, encouraged by their great princes, incorporated dance into their lavish spectacles and court entertainments. The earliest book from which scholars can endeavour to re-create their dances was appropriately called *De arte saltandi et choreas ducendi* ("On the Art of Dancing and Conducting Dances") by Domenico da Piacenza, written in the early 1400s. From Domenico's manuscript and from the published works of his successors can be traced, albeit with much conjecture, the dance forms of Europe that were to spread throughout the Western world.

Post-Renaissance developments. The art of dancing in modern times stems from these Italian dancing masters. The dances of the people, the folk dances, often yielded characteristics that were incorporated into the more formal measures of the nobility. The court dances were used in the court entertainments until the end of the 18th century, by which time professional dancers took over as the entertainers and techniques began to diverge. The courtiers, who loved dancing for the honoured guests, maintained the noble style, whereas the professionals, at first following the court's example, eventually developed a technique of their own that was eventually to be codified by the 19th-century Italian teacher and theoretician Carlo Blasis as the classical ballet. But, once the courtiers had handed over responsibility for staging and dancing to the professionals, the audience changed, for the ballets were mounted in the new public theatres, and the common people could watch as well as the nobility.

Court dance and its descendants

Social dances—as distinct from folk dances—became the ballroom dances of the upper classes, so, though class distinctions gradually were eroded, especially in the 20th century, such dances became popular among all classes. Dances with a prescribed technique were performed at formal occasions. The popularity of different dances changed with the discovery of new rhythms for which new steps were invented. Queen Elizabeth I of England delighted in a dance called the volta, which contained some rather daring "lifts" and doubtless horrified her churchmen. In the 19th century the waltz was a "craze" dance and was considered shocking; today, it is sedate yet charming. The 20th century has seen other temporarily fashionable dances—the Charleston boom of the 1920s, rock and roll in the 1950s. Common to all such forms of dance has been their reflection of the social structure and the "spirit of the time."

COMPONENTS OF THE DANCE

The dancer. It is through the body of the dancer that the art of dancing is portrayed, and the physical, emotional, and national characteristics of the dancer determine the quality and the nature of the dance. For example, women tend to be more supple than men, but they do not usually have equal strength and stamina. The way in which dancers perform depends greatly on the land of their origin, the dance styles, the national dress, even the climate. The dances of northern countries tend to be vigorous and those of southern lands more languid. Eskimos, bundled up in warm clothing, obviously will not attempt dainty movements. The small-boned, delicately built peoples of Southeast Asia, in their beautiful traditional costumes, perform dances that are exquisitely subtle. Conquest and occupation of a country by various invaders may alter and enrich dance styles. The Philippines, having been occupied by many races, have a dance culture that has borrowed from all of them.

Cultural and sexual dichotomies in dance

Temperament, a difficult word and idea to define, also affects the style of the dancer. Slavic peoples have a reputation for fiery manners, and the proud dignity of the Spaniard is reflected in his dances. Italian gaiety produced the lively tarantella. The dances of North American Indians and other tribally organized peoples are nearly all associated with matters of group concern.

Dance styles throughout the world are different for men and women. In most national dances the men demonstrate their strength and virility, the women their charm and femininity—characteristics that are well illustrated in Russian folk dances. In ballet, women's steps sometimes are performed by men for burlesque purposes, whereas some women have mastered dance styles that are really a male preserve. The great 20th-century Gypsy dancer Carmen Amaya could rival any man in the tem-

pestuous speed of her flamenco dancing. In India, Shanta Rao specialized in the *kathākali* style, which is usually associated with men.

Dance training

The dancer's instrument is the human body, and the ways in which it can be used depend almost entirely on how it has—or has not—been trained. It is possible to devise very beautiful and striking ensemble works with dancers who have had virtually no training but who can look and move well. A clever choreographer can manipulate them into sculptural poses or groups. Many Nativity plays, for example, have used untrained dancers to very good effect.

The trained body is, however, nearly always more eloquent and permits a much greater variety of steps, jumps, turns, lifts, and almost acrobatic feats. The classical ballet dancer (who must start training from about the age of 11 to achieve full technical mastery) has a remarkable range of movement that cannot be attempted by ordinary mortals. Similarly, the highly trained dancers of Bali and the men and women who have mastered the complex language of Indian dancing can use their bodies far more expressively than can the untrained.

The degree of movement of which the human body is capable depends on the flexibility and strength of the muscles and on the length of the bone-connecting ligaments. These tissues are not flexible, like muscles, but they can be lengthened by special training at an early age, before they have had time to harden. This is one reason why the professional dancer must start training young. In classical ballet, it is always possible to spot a "late starter" because the body—and especially the feet and legs—will not have the strength and flexibility of the dancer who has started young. The training of a dancer, in any style, is based on daily practice of exercises carefully planned and perfected over centuries, to give the greatest strength and awareness to every part of the body.

The least flexible part of the body is the skeleton, the bony frame. The structure of bones and joints governs the amount of bodily movement in any one direction. The ribs and chest can easily be bent to each side and forward but will not bend backward. The ball and socket structure of the shoulder and hip joints permits a small degree of movement. Movement from the hip is easier in a forward direction; it is more difficult to swing the leg up to the side or the back than in front of the body. The ballet dancer must practice until his legs can be raised high in all directions without loss of balance or control. A fundamental of dancing is the control of distribution of weight.

Although the ballet dancer trains each part of the body, performing exercises to acquire harmonious movements of the arms and changes of direction with the shoulders (épaulement), special care is given to the feet and legs. The hours of working in the classroom, at the barre or in the centre of the room, are designed to help the dancer to leap high in the air, to perform intricate beating movements of the feet (batterie), and also to alight softly and surely and with sufficient control to go straight into a fast series of multiple turns (pirouettes).

The Hindu dancer of India, performing barefoot and not attempting the virtuoso footwork of the ballet or other Western dancers, cultivates an equally complex technique. The *Bhārata Nāṭya-śāstra*, the 2,000-year-old manual of Hindu dance, lays down movements for all parts of the body, but it has been calculated that there are at least 4,000 *mudrā*s (picture patterns of the hands), so each finger must be eloquent. To the initiated, these *mudrā*s alone can tell detailed myths and stories. The 20th century has seen a revival of Indian dancing, and pioneers such as Uday Shankar and Ram Gopal have helped Western audiences understand their art.

Because the trained dancer has brought his body to a high state of flexibility, control, and alertness, he will always be a better and more expressive dancer than the untrained one. His body is probably more beautiful and his carriage more erect, so he can achieve the harmony of movement that is the essence of the art of dance.

The age at which a dancer retires depends entirely on the style of dancing and the amount of physical effort required. If the dance style can be adapted and refined down to a minimum of movement, then the dancer can continue until death. In nations in which age is revered, tribal elders will preside over dances until their deaths. In the West, the U.S. dancer Ruth St. Denis was still touring and appearing in some of her simple solo dances when over 80. The career of a classical ballerina, however, usually ends at 50, and male dancers tend to lose their virtuosity by their mid-40s. Character dancers in ballet can, of course, continue until old age. A man may lead a mazurka or czardas with authority until well into his 60s, while mime roles can be played by even older dancers. Ballroom dancing can be enjoyed so long as a man or woman is active, though perhaps without complicated variations.

Longevity of dancing skills

The choreographer. The choreographer, the arranger of dances either for a soloist or a group, is the most important figure in dance history. His is the creative mind that invents the order and combination of steps, the patterns of group dances, the communication of narrative through dance, and the extension of dance techniques by using familiar actions in new and more complex ways.

A choreographer, given trained bodies, can set them in motion in an almost unlimited number of ways. It is possible to construct "story ballets," grave or gay, in which the action is unfolded through dance and mime. Different combinations of steps can express different emotions. A quick, light, darting solo, for instance, may suggest the happiness of first love; heavier, earthbound, and slower dances may convey grief. Dance can be highly stylized or simple, almost naturalistic. Choreographers draw inspiration from many kinds of dancing, and each kind influences the others.

Unfortunately, nothing at all is known of the work of the first choreographers. Someone must have suggested the shapes of the traditional folk and group dances, but only the work of those who have tried to revive and preserve them can be studied.

Attempts have been made to write down dance patterns—social and theatrical—from the time of the first books about dancing. Western dance can be reconstructed from various systems of notation ever since Thoinot Arbeau, a canon of Langres in France, published his *Orchésographie* in 1588. The systems, however, although they can be deciphered by experts today, leave much open to conjecture. The most comprehensive system invented is that of Rudolf Laban, clearly set out in the textbook *Labanotation or Kinetography Laban* (1971) by Ann Hutchinson. Used in conjunction with film, this method records any kind of movement.

Methods of preserving choreographic works

Lacking such aids, choreographers have relied in the past on their works being preserved and handed down by generations of dancers, but inevitably changes will occur constantly. No producer can resist adding his personal stamp to a restaging of even the most cherished classic of the ballet repertory. It is possible to study the history of ballet, for example, from fragments of choreography surviving from as early as 1786, when the Italian ballet master Vincenzo Galeotti staged his *Whims of Cupid and the Ballet Master* for the Royal Danish Ballet in Copenhagen. The Danes have been great traditionalists in ballet, and their version of *La Sylphide*, originally staged in 1836, preserves most purely the Romantic style of ballet. In it can be found all characteristics of that period—the toe dancing of the supernatural Sylphide, the setting in a foreign country (Scotland), and the brilliant footwork for the male dancer that was derived from the example and teaching of Auguste Vestris, star performer of the French ballet in the late 18th and early 19th centuries.

In Russia, where technique was being enriched by teachers from France, Italy, and Scandinavia, the Romantic ballets were injected with the new virtuosity, and it is from Russia, through the Stepanov system of dance notation, that most of the ballets of the Romantic era have been restaged in the West. The 19tb-century Frenchman Marius Petipa was the architect of the pure classical ballet as used in such works as *The Sleeping Beauty*, *La Bayadère*, and the more brilliant parts of *Swan Lake*.

In the 20th century the influential choreographers have been the Russian reformer Michel Fokine, who returned to the landmark principles of unity of construction and expression articulated by the innovative dancer-choreographer Jean-Georges Noverre in 1760; the Russian dancer Vaslav Nijinsky, who in a short, tragic career created three works (none of which survives) that are considered of capital importance because of his novel uses of dance; the Russian Léonide Massine, who worked in the 1920s with Cubist and Constructivist painters and used every medium of theatre art; Nijinsky's sister, Bronisława Nijinska, who brought ritual and sculptural groupings to the ballet stage in *Les Noces* (1923; still danced by Britain's Royal Ballet); George Balanchine, who has shaped the development of classical dance in the United States; Sir Frederick Ashton, architect of the English, lyrical style of classical dancing; the Englishman Antony Tudor, who found new "psychological" undertones that could be expressed through movement; and the American Jerome Robbins, who has united nearly all forms of dance, writing with genius for classic ballet, musical comedy and drama, or motion picture. Some of their works have been recorded, but many have already been lost.

Accompaniments to the dance. Dance can be and has been accompanied by every conceivable kind of sound. On the other hand, it may need no sound, for the dancer, blessed with conscious rhythm, can perform highly complex and stylized movements in silence. It is reasonable to believe that the first accompaniments were stamping, clapping, and chanting and then possibly drumming. Song has also played an important part: at the time of the opera-ballets of the 17th and 18th centuries in western Europe, the arts of music, dance, and song were inseparably linked. Once divorced, theatre dance found its own composers, and the status of ballet has been almost entirely governed by the quality of its music. Dance is not dependent upon music, but it is usually enhanced by it; a gavotte danced to the music of Gluck is likely to be a more highly refined form of social dancing than a fox-trot done to a dance band.

Diversity of musical and natural sounds

The musical instruments used to play for dancing are almost invariably those of the country of the dance's origin—for example, the gamelan orchestra that accompanies the dances of Java and Bali. Sometimes, however, unusual juxtapositions can be effectively used in the theatre. Contemporary modern dancers in the West have found inspiration in Indian music. Music specifically written for the concert hall has often been used—with varying degrees of success—by choreographers working in the theatre or other mediums.

The spoken word can also be used as an effective accompaniment to dance. It may be written for the purpose, or a dancer may choose to "interpret" a poem through movement while it is being spoken on stage. Sometimes both the musicians and dancers will improvise, almost in competition with each other.

Theatrical effects. When dance is a spectacle, every device of costume, makeup, and theatrical effect can be used. The makeup and costumes for the ancient dances of Japan, China, and India are astounding in their complexity and the significance that each part of costume and the painting of the face plays in the dance-dramas. Kabuki, the popular Japanese dance theatre, places great emphasis on costume, and the personality of the dancer is molded, virtually extinguished, by the theatrical trappings. The spectacular Kathākali dancers of India spend up to four hours being made up for a performance. The makeup entirely transforms the human face to enable the dancer to play the superhuman characters of their dance-dramas.

The enhancement of visual effects

Even in the more realistic performances of the West, costume and decor play an important part. The entertainments of the Renaissance were of surpassing beauty, both in the costumes and in the scenic effects. Costume is nearly always influenced, in the theatre, by fashion. Dance techniques are encumbered or liberated by costume. Not until the invention of tights (French *maillot*) was the female ballet dancer free to leap high and throw her legs into the air. In recent years tights and leotards covering most of the body have given complete freedom of movement to both sexes.

The theatrical effects of the ballets staged in the 18th-century court theatres were elaborate. Gods would ascend or descend on chariots or clouds, and the stage machinery was remarkably effective. (A perfect survival is the court theatre at Drottningholm, Sweden, still used for the performance of 18th-century operas and reconstructions of the old ballets.)

Ballet design in the 18th and 19th centuries was usually entrusted to a resident designer, attached to the theatre. Diaghilev revolutionized this practice by bringing easel painters of such calibre as Pablo Picasso and Henri Matisse to design ballets.

Today, choreographers use every kind of lighting and mixed media in their productions. Film projections have long been used. The dance, especially among the modern dancers of the United States, and above all in the work of Merce Cunningham, keeps pace with every element of the theatre—and sometimes is ahead. The Alwin Nikolais Dance Company has probably had more influence on the theatre and television than on dance.

Narrative methods and materials. There are many things that can be said in dance, others that cannot be said. Passage of time in years can be suggested by aging the characters; lighting can indicate different times of day. Certain relationships cannot, however, be conveyed in dance or mime. It is easy to establish a love relationship through dancing, but it is quite impossible without program notes to convey blood relationships.

Dance as language and the languages of dance

The dancer communicates with the audience in two distinct ways, either through an outpouring of emotion through the body as well as the face or by a complex language of mime and gesture that can be fully understood only by spectators who know the "language."

The best known of these sign languages is the traditional mime of the classical ballet (now almost lost) and the *mudrās*—symbolic hand gestures—of Indian dancing. The classical ballet gestures are sometimes obvious to the point of naïvete—hand on heart means love; pointing to the fourth finger of the left hand means marriage, but these may be no more understood by the layman than the fact that a powerful downward movement of crossing the arms means death or that a circling of the arms above the head means to dance.

Expressive variety of Indian dance movement

The classical dancing of India uses an enormous amount of hand gestures that, to the initiated, can describe anything from a tortoise to a fish to a lotus flower. Although Indian dancing may be enjoyed by Westerners simply for its technical expertise and for the gorgeous costumes, a study of the basic *mudrās* is rewarding. In *kathākali* alone there are several hundred, built up from a basic 24. Good dancers are expected to have a vocabulary of 500.

The early ballets in the West were mostly based on stories from classical mythology, and, by the time ballet had reached the court of the 17th-century French king Louis XIV, it was being used for political purposes. The King himself danced only the most important roles, thus impressing upon the spectators his supremacy in his kingdom. (It was after he appeared as the Sun in the *Ballet de la nuit*, 1653, that he acquired the title Le Roi Soleil.)

Narrative and nonnarrative usages

A choreographer of genius can create works that are almost as eloquent in their storytelling as a play, and there have been examples of plays being translated into dance form and retaining much of their impact. The American Martha Graham has translated Greek tragedy into terms of contemporary dance, producing even a full-length version of *Clytemnestra* (1958). In England, the choreographer Sir Frederick Ashton used Shakespeare's *Midsummer Night's Dream* for his ballet *The Dream* (1964), contrasting the delicate dancing of the fairies with the rural frolics of the rustics and the comedy of manners of the lovers' quarrels. Quite different was Sir Robert Helpmann's version of *Hamlet* (1942), presented as a tormented and distorted flashback of events in the mind of the dying prince.

Such works, of course, presuppose a knowledge of the

play. Other narratives, such as *La Fille mal gardée*, exist only on the stage. First produced in Bordeaux in 1789, this charming pastorale—of a girl who escapes her mother's attempts to marry her to a rich simpleton and subsequently wins the young farmer she loves—has been, in varying styles of production, an enormous success. A narrative work gives pleasure from characterization and mime as well as from dance. Non-narrative works, such as Balanchine's *Jewels* and Ashton's *Monotones I and II*, rely upon pure dance images to make their effects.

KINDS OF DANCE

There are as many kinds of dancing as of any other human activity. The art of dancing stretches from the simplest forms of pure dance expression, performed by an individual for his or her own satisfaction, through the complexities and subtleties of Eastern dance, to the most sophisticated, modern, and large-scale productions in which dance is the dominant factor. The definitions must therefore be broad, and they must encompass both general ways of dancing as well as the recognized genres of the dance.

Natural, or untutored, dance. Being a natural impulse, dance can be enjoyed both by the trained and the untrained dancer. The untrained may derive the greatest personal pleasure, having to obey no rules, but will be restricted by the lack of a technique, a vocabulary of movement. Nearly every man and woman must at some time have danced for pure pleasure, spontaneously leaping and running to release an intoxication of high spirits. "Creative dance" is now encouraged in some schools that frown upon what they call the unnatural and rigid disciplines of formal dance training. Children are encouraged to "express themselves" in movement. In doing so, they simply release the natural instinct and, by giving it some semblance of "meaning," are veering toward the art of dancing.

The spontaneous play of children's dance

This activity may have therapeutic value, but it is arguable that it has any educational value, since the children submit to no discipline and are required to obey no laws of movement. From time to time, however, a great artist emerges who, through sheer force of personality and without formal training, can devise a form of dancing that may, in its own right, be described as true art. In the 20th century the supreme example is Isadora Duncan, but it is significant that her art died with her and that it was dependent to a great degree on the great music to which she danced.

A trained body is obviously a more articulate instrument than an untrained one, but these strictures apply mainly to the performing artist. For the individual to take part in group or couple dances, it is necessary to learn only a few basic steps.

Solo and group dance. Solo dance is usually performed for a specific purpose. Among primitive peoples, the incantations and frenzied movements of the shaman or witch doctor were designed to contact the spirits of nature and to drive out evil ones. Among sophisticated peoples, the solo dancers are usually those who perform in public recitals to demonstrate a particular technique or style of dancing in a program of their own devising. "Concert dancers" flourished in Germany in the 1920s and 1930s and among the modern dance pioneers in the United States. In all kinds of dancing, "stars" or virtuoso performers usually have a favourite dance of their own that they may perform within the context of a company. For example, the 19th-century Romantic ballerina Fanny Elssler was internationally celebrated for her cachucha dance, a Spanish dance accompanied by castanets. Anna Pavlova, a great Russian classical ballerina, is mostly remembered for her famous solo *The Dying Swan*. The supreme Spanish dancer of this century, Antonio, will be remembered for his solo *zapateado*, a flamenco dance of dazzling and intricate heel work, rather than for his lavish stage spectacles.

Group dances fall into several categories. Folk dance is almost always performed by groups with perhaps a leader to begin the dance. Ballroom dancing is likewise a group activity, with many couples taking the floor together and sometimes changing partners in the course of a dance. In classical ballet the corps de ballet, or chorus, functions essentially to sustain the action and to provide a framework for the principal dancers. Modern or contemporary dancers usually have their own groups for whom they choreograph, although they often take major parts themselves.

Ethnic dance. The concept of ethnic dance probably embraces the largest variety of dances, since it applies to all dance that is indigenous to a certain race or country. It applies to the primitive tribal dances, with their built-in purpose of religious supplication or thanksgiving, and it applies to the most highly trained court dancers used in religious ceremony, especially in the East. Ethnic dance includes the traditional styles of dancing developed all over the world, and excludes theatrical or spectator dance as well as folk dance.

Folk dance. The folk dances of the world are, as the name suggests, the dances of the peoples of all parts of the world that they have devised for their own pleasure. Essentially, they are for the enjoyment of the performers, and they are not spectator dances— although pleasure may be derived from watching them, and they have been, in recent times, artificially preserved and performed in places far from the village green or jungle clearing of their origin.

In the linked, or chain, dances from which most European folk dance derives, it is easy for a dancer to learn the rudimentary steps. Only in the 20th century have folk-dance groups been trained in the kind of virtuosity that makes them exciting theatre for sophisticated audiences. The English folk dances have been artificially preserved by scholars and enthusiasts and are enjoyed by thousands of amateurs, but they are no longer danced by country people. They are, of course, simple in technique and require no special training.

Preservation of folk dances

Court dance. The court dances of the Western world were social accomplishments, polished from material usually found among folk dances. The root dance in the West was the carole of the early Middle Ages. From its two forms—the brarles, which gave the first known rhythms and steps, and the farandole, which gave figures or floor patterns—all court dances developed.

In 12th-century Provence, courtly dancing began to emerge in a dance form called estampies, which may have been processional but is now believed to be the beginning of the *danses à deux*, or couple dances. In 15th-century Italy a distinction is made for the first time between a *danza* and a *ballo*. A *danza* had a uniform rhythm throughout, a *ballo* had varied rhythms.

By the 16th century in the courts of France and England, figure dances with a distinct floor pattern were common. Such dances as the pavane, galliard, and courante were popular and were being described by dancing masters to the court. Intricate steps and rhythms led, during the 17th and 18th centuries, to such dances as the sarabande, gavotte, minuet, and hornpipe.

The French Revolution and the Napoleonic Wars put an end to court dances as spectator dances performed by and for the nobility. Dancing remained a social accomplishment for the upper classes, but the dance forms of Western civilization now divide into those of the ballroom and those of the theatre.

Transitions from court to ballroom styles

Ballroom dance. Ballroom dancing, sometimes called social or popular dance, is nearly always now performed by couples, dancing prescribed steps for their own pleasure either at private functions or in public dance halls. Between 1820 and 1910 the popular dances were the waltz, quadrille, and polka. The Russian and eastern European nobility before the Revolution also danced adaptations of such national dances as mazurkas and czardas. Immediately before World War I the rhythms of Negro and South American dances reached Europe, and the tango and then the rumba became the rage. They were the first of the popular Latin American dances.

The popular teen-age dances of the mid-20th century have no set "steps"; the dancers respond spontaneously to the beat of the musicians. The degree of satisfaction attained by young people "twisting" or "shaking" to the

blare of amplified music in dance halls, further enlivened by psychedelic lighting, is different from the pleasure derived by their elders waltzing to *The Blue Danube*—but it is only a difference of age and time. Fundamentally, both age groups are enjoying the pleasure of dancing in their own way and at their own pace. Once, it was scandalous for a man to put his arm around a woman as they danced. By the 1950s young dancers once again were dancing without touching their partners, apparently oblivious to "coupleness." The end product is doubtless the same—physical pleasure in the activity of dancing and sexual awareness of a partner, whether embraced or half-consciously observed.

Ballroom-dance contests

International contests in ballroom dancing gained favour early in the 20th century. In the 1920s in London, dance teachers codified the basic steps of such dances as the waltz, tango, and fox-trot and encouraged dancers to learn the steps, practice them to professional standards, and enter competitions or championships. The championship type of ballroom dancing now bears little relation to what is danced in society, the steps having become so intricate and artificial that only people who have time for constant practice can accomplish them. Championships continue, nonetheless, to attract thousands of spectators in various countries.

Theatrical dance. When Louis XIV of France in 1671 (the date of the first performance at the Paris Opéra) handed over to the professional dancers the entertainments that had been performed by the court, he laid the foundations of classical ballet as it is known today. His dancing masters, drawing on the earlier work of the Italians, gradually codified the basic principles of the five positions of the feet, the directions of the body, the use of arms, head, and shoulders, and the seven movements in classical dancing: to bend, to stretch, to raise, to slide, to jump, to dart, and to turn. Since the first textbooks were written in French and since France reigned supreme in ballet until the Revolution in 1789, classical technique is still described in French.

Classical ballet is the most highly trained and physically the most eloquent form of theatrical dance. It has been criticized—perhaps inevitably—by many for its seeming artificiality, and many innovators in the 20th century have looked for other forms of expression through dance. In Germany, Mary Wigman became the founder of modern dance in Europe, performing her own dances and teaching generations of others. In the U.S., where the pioneer work of Isadora Duncan was largely ignored during her lifetime, modern dance was brought to its highest achievements by such artists as Hanya Holm (a pupil of Wigman), Doris Humphrey, Martha Graham, and their students. The modern dancer usually performs barefoot but has a demanding technique.

Dance can be used to heighten the effect of dramatic productions, and it has been used with skill in films and on television. The best known of film dancers was the American Fred Astaire, who used tap dancing as the high point of the many film musicals in which he starred. Stage dancing translates badly to film or television. The best work in these media has been created especially for the screen, notably, by the American Gene Kelly, in such films as *Singin' in the Rain* (1952).

Dance has always been part of the musical theatre, sometimes a mere interlude, sometimes an essential part of the production. Agnes deMille used a dance sequence in *Oklahoma!* (1943) to further the plot, and, since that time, dance has occupied an increasingly important role in Broadway musicals. Jerome Robbins achieved an unusual synthesis in his *West Side Story* (1957).

BIBLIOGRAPHY. The earliest surviving manuscript about the art of dancing was written by DOMENICO DA PIACENZA, early in 1400; in it he describes the dances of his day and the dances he himself arranged. This work was written before the invention of printing, but books by Domenico's pupils were printed and survive in the great dance-book collections of the world. Most of them are available on microfilm or in paperback form. The *Livre des basses danses* (*c.* 1450), called *The Golden Manuscript*, is a valuable source of information about medieval dances in western Europe, which has been reissued in facsimile (1912). The Italian dancing master FABRITIO CAROSO in *Il ballarino* (1581) describes and illustrates the dances of his day. The Italian dancing master BALTHAZAR DE BEAUJOYEULX'S grand spectacle staged for Catherine de Médicis in France, the *Ballet comique de la reine* (1581), was described in an elaborately illustrated publication (1582) that was distributed throughout the courts of Europe. THOINOT ARBEAU, *Orchésographie* (1588), was translated by the British ballet critic and historian CYRIL BEAUMONT in 1925. It contains dance descriptions, musical examples, and some small drawings. Written partly in dialogue form, it is subtitled *Whereby all manner of persons may easily acquire and practise the honourable exercise of dancing.* In 1604 the Italian dancing master CESARE NEGRI published his treatise on dancing, *Nuove inventioni di balli*, which is richly illustrated. In 1651 the English musician JOHN PLAYFORD published a collection of dances and their tunes called *The English Dancing-Master*, which is proof that the art of dancing was still popular during the Puritan regime of Cromwell's Commonwealth. In 1682 CLAUDE-FRANCOIS MENESTRIER, a French Jesuit, wrote his *Des Ballets anciens et modernes*, the first printed history of the ballet. In 1701 came the French dancer and choreographer RAOUL FEUILLET, *Chorégraphie ou l'art de décrire la danse;* Feuillet was one of the first to invent a system of dance notation, showing the floor patterns of the dances. He wrote a number of other works on the dance and greatly influenced the English dancing master JOHN WEAVER, who translated Feuillet's book (*Orchesography, or the Art of Dancing*, 1706) and himself wrote in 1712 *An Essay Towards the History of Dancing.* In 1725 came the French dancing master PIERRE RAMEAU, *Le Maître à danser* (1725; Eng. trans., *The Dancing Master*, 1931), a well-illustrated book that sets down the rules of dancing and deportment at that time. NOVERRE'S *Lettres sur la danse, et sur les ballets* (1760) stated the principles of this fiery, reforming ballet master, which are still valid. By 1820 the technique of classical ballet as it is still known had been codified by the Italian dancer-choreographer-teacher CARLO BLASIS in his *Traité élémentaire théorique, et pratique de l'art de la danse* (*An Elementary Treatise upon the Theory and Practice of the Art of Dancing*, 1944), a technical manual. The *Manual of the Theory and Practice of Classical Theatrical Dancing*, by CYRIL BEAUMONT and Polish dancer STANISLAS IDZIKOWSKI (1922), records the method of the great Italian teacher Enrico Cecchetti. *Osnovy klassicheskogo tantsa* (1934; Eng. trans., *Fundamentals of the Classic Dance*, 1948, reprinted 1969), by the Russian ballerina and teacher AGRIPPINA VAGANOVA, describes the school of Soviet ballet; it has been reissued and updated several times.

Among 20th-century scholars who have contributed important books toward an understanding of the art of dancing are the Austrian researcher CURT SACHS, *Eine Weltgeschichte des Tanzes* (1933; Eng. trans., *World History of the Dance*, 1937, reprinted 1963); CYRIL BEAUMONT, author of innumerable, invaluable works; the U.S. musician and ballet writer LOUIS HORST, *Preclassic Dance Forms* (1937), indispensable to the study of early dances; the U.S. writer and ballet patron LINCOLN KIRSTEIN, *Dance* (1935), probably the best single survey of the art, and *Movement and Metaphor: Four Centuries of Ballet* (1970), a brilliant analysis of the component parts of ballet followed by descriptions of 50 seminal works illustrating the development of ballet from 1573 to 1968; and LILLIAN MOORE and IVOR GUEST, historians of the ballet. The art of ballroom dancing is surveyed by the British writer and ballet patron P.J.S. RICHARDSON in his *History of English Ballroom Dancing* (1946) and *Social Dances of the Nineteenth Century in England* (1960); and by the ballet writer ARTHUR H. FRANKS in *Social Dance: A Short History* (1963). In England, CECIL SHARP, founder of the English Folk Dance and Song Society, sought out the national dances and the songs and music that accompanied them and left a legacy that this society not only preserves but constantly seeks to enrich. In North America, TED SHAWN studied the dances of the American Indian. IGOR MOISEYEV rescued Russian national dances and also choreographed them for presentation on theatre stages and in public arenas. On Indian dance, see BERYL DE ZOETE, *The Other Mind: A Study of Dance in South India* (1953); and RINA SINGHA and REGINALD MASSEY, *Indian Dances: Their History and Growth* (1967), a lucid guide to the complexities of Indian dancing, accurate and well illustrated.

Governments throughout the world have begun to realize the export potential of their national dances, and many are financing research into their origins so that they can be shown overseas with some degree of authenticity. The libraries and museums of most of the European opera houses have a wealth of material about the art of theatrical dancing. The museum of the Paris Opéra and the collection of the Bibliothèque Nationale are probably the richest. In London, the Victoria and Albert Museum has many treasures. The Library of the Royal Academy of Dancing contains the priceless collection of rare

dance books assembled by P.J.S. RICHARDSON (of which a bibliographical description was published in 1954 and all of which is on microfilm). The Royal Ballet School also has a valuable archive. The dance collection of the New York Public Library is one of the greatest in the world. Harvard University also has an important dance collection, and many other U.S. university libraries are now buying dance materials. In Stockholm, BENGT HAGER has assembled a superb dance museum that incorporates most of the Rolf de Maré collection, the Archives Internationales de la Danse. The greatest private collection of dance books is undoubtedly that of the Hungarian dancer and teacher DERRA DE MORODA in Salzburg, Austria.

ANATOLE CHUJOY and P.W. MANCHESTER (eds.), *The Dance Encyclopedia*, rev. ed., with an introduction by LINCOLN KIRSTEIN (1967), with nearly 5,000 entries and 274 photographs, is a standard reference work that contains illuminating articles about all forms of dancing.

(M.Cl.)

Dance, Western

The peoples of the West—of Europe and of the countries founded through permanent European settlement elsewhere—have a history of dance characterized by great diversity and rapid change. Whereas most dancers of the East repeated highly refined forms that had remained virtually unchanged for centuries or millenniums, Western dancers showed a constant readiness, even eagerness, to accept new vehicles for their dancing. From the earliest records, it appears that Western dance has always embraced an enormous variety of communal or ritual dances, of socializing dances at different levels of society, and of skilled theatrical dances that followed distinct but often overlapping lines of development.

The West cannot always be clearly distinguished from the non-West, especially in such countries as Russia, where some dances are Asian and others European in origin and character. This article focuses on the dance of Western peoples, noting where appropriate the influence of other cultures.

The articles BALLET; FOLK DANCE; MODERN DANCE; and POPULAR DANCE cover in greater detail the unique nature, techniques, forms, and functions, and the historical developments of each of these kinds of Western dance. In addition, the article DANCE, ART OF, covers the aesthetics and the varieties of dance, both Western and non-Western. Aspects of Eastern dance are detailed in the article DANCE AND THEATRE, EAST ASIAN.

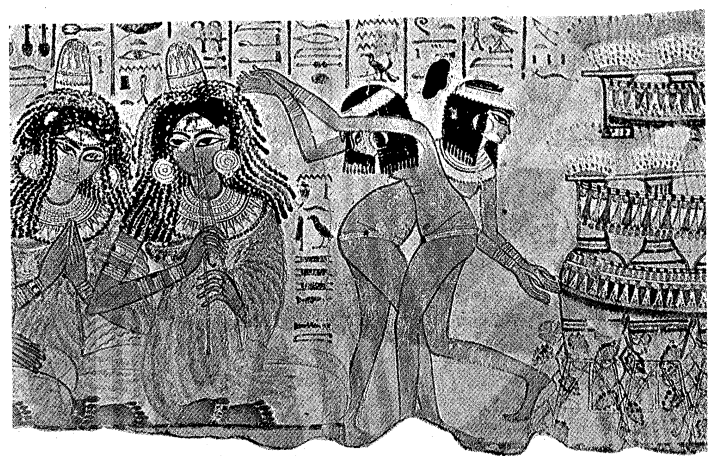
From antiquity through the Renaissance

Before men left written records, a vast span of time elapsed about which scholars can only speculate. Pictorial records in cave paintings in Spain and France showing dancelike formations have led to the conjecture that religious rites and attempts to influence events through magic were central motivations of prehistoric dance. Such speculations have been reinforced by observation of dances of primitive peoples in the contemporary world, though the connection between ancient and modern "primitives" is by no means accepted by many scholars. If the dances recorded in early written records represented a continuity from prehistoric dances, there may have been prehistoric work dances, war dances, and erotic couple and group dances as well. One couple dance surviving in the 20th century, the Bavarian-Austrian *Schuhplattler*, is considered by historians to be of Neolithic origin, from before 3000 BC.

DANCE IN THE ANCIENT WORLD

In the civilizations of Egypt, Greece and its neighbouring islands, and Rome, written records supplement the many pictorial remains. There are still conjectures, but there is far more concrete evidence.

Ancient Egyptian dance. With its highly organized forms of community life, Egypt practiced formalized ritual and ceremonial dances in which the dancing priest-king represented the person of a god or the servant and regenerator of his people. These dances, culminating in ceremonies representing the death and rebirth of the god Osiris, became more and more complex, and ultimately they could be executed only by specially trained dancers.



Egyptian dancing, detail from a tomb painting from Shaykh ʿAbd al-Qurnah, Egyptian, *c.* 1400 BC. In the British Museum.
By courtesy of the trustees of the British Museum

From Egypt also come the earliest written documentations of the dance. These records speak of a class of professional dancers, originally imported from the interior of Africa, to satisfy the wealthy and powerful during hours of leisure. These dancers were considered highly valuable possessions, especially the pygmy dancers who became famous for their artistry. One of the pharaohs prayed to become a "dance dwarf of god" after his death, and King Neferkare (3rd millennium BC) admonished one of his marshals to rush such a "dance dwarf from the Land of Spirits" to his court. African origins

There is considerable agreement that the belly dance, now performed by dancers from the Near East, is of African origin. A report of the 4th century BC from Memphis in Egypt described in detail the performance of an apparently rumba-like couple dance with an unquestionably erotic character. The Egyptians also knew acrobatic exhibition dances akin to the present-day adagio dances. They definitely were aware of the sensual allure of the sparsely clad body in graceful movement. A tomb painting from Shaykh ʿAbd al-Qurnah, now in the British Museum, shows dancers dressed with only rings and belts, apparently designed to heighten the appeal of their nudity. These figures probably were intended to entertain the dead as they had been entertained in life.

Egypt, then, presented a dancing scene that was already varied and sophisticated. In addition to their own danced temple rituals and the pygmy dancers imported from the headwaters of the Nile, there were Hindu dancing girls from conquered countries to the east. This new dance had none of the long masculine strides or the stiff, angular postures seen in so many Egyptian stone reliefs. Lines of movement undulated softly, nowhere bending sharply or breaking. These Asiatic girls brought a true feminine style to Egyptian dance.

Dance in classical Greece. Many Egyptian influences can be found in the Greek dance. Some came by way of Crete, others through the Greek philosophers who went to Egypt to study. The philosopher Plato (*c.* 428–348/47 BC) was among them, and he became an influential dance theoretician. He distinguished dances that enhance the beauty of the body from awkward movements that imitate the convulsions of ugliness. The initiation rites of Egypt had their equivalent in the Cretan bull dance of about 1400 BC. It inspired the labyrinthine dances that, according to legends, Theseus brought to Athens on his return with the liberated youths and maidens.

Another dance form that originated in Crete and flourished in Greece was the *pyrrhichē*, a weapon dance. Practiced in Sparta as part of military training, it was a basis for the claim of the philosopher Socrates that the best dancer is also the best warrior. Other choral dances that came to Athens from Crete include two dedicated to Apollo and one in which naked boys simulated wrestling matches. Female characteristics were stressed in a stately



Kordax dance, Greek vase painting, 5th century BC. In the Museo Nazionale Tarquinise, Italy.
SCALA

and devout round dance in honour of the gods, performed by choruses of virgins.

Numerous vase paintings and sculptural reliefs offer ample proof of a wild and ecstatic dance connected with the cult of Dionysus. It was celebrated with a "sacred madness" at the time of the autumnal grape harvest. In his drama *Bacchae*, Euripides (*c.* 480–406 BC) described the frenzy of Greek women, called bacchantes or maenads. In their dance for generation and regeneration, they frantically stamped the ground and whirled themselves about in rhythmic convulsions. Such dances were manifestations of demoniacal possession characteristic of many primitive dances.

Dionysian dances

It was the Dionysian cult that brought about Greek drama. After the women danced, the men followed in the disguise of lecherous satyrs and sileni. Gradually the priest, singing of the life, death, and return of Dionysus while his acolytes represented his words in dance and mime, became an actor. The scope of the dance slowly widened to incorporate subjects and heroes taken from the Homeric legends. A second actor and a chorus were added. In the lyric interludes between plays, dancers recreated the dramatic themes in movements and steps adopted from the earlier ritual and bacchic dances. In the comedies, they danced the very popular kordax, a mask dance of uninhibited lasciviousness.

These dances and plays were executed by skilled amateurs. At the end of the 5th century BC, however, there came into being a special class of show dancers, acrobats, and jugglers, the female members of which were evidently hetaerae, members of a class of courtesans. No doubt influenced by Egyptian examples, they entertained guests at lavish banquets. The historian Xenophon (*c.* 430–*c.* 355 BC) in his *Symposium* tells of the praise Socrates lavished on a female dancer and a dancing boy at one such occasion, finally himself emulating their beautiful movements. Elsewhere, Xenophon describes a dance representing the union of the legendary heroine Ariadne with Dionysus, an early example of narrative dance.

SCALA



Funeral dance, Etruscan fresco from a tomb cover, 5th century BC. In the Museo e Gallerie Nazionale di Capodimente.

Ancient Roman dance. There was a striking difference between the Etruscan and the Roman peoples in their approach to the dance. Little is known about the Etruscans, who populated the area north of Rome up to Florence and flourished between the 7th and 5th century BC. But it is apparent from their lavish tomb painting that dance played an important part in their enjoyment of life. Women apparently had a prominent role in these couple dances. They were performed without masks in public places and showed a distinct courting character.

In Rome, a basic antagonism to dance seems to reflect the sober rationalism and realism of the Roman people. Nonetheless, Rome did not entirely evade the temptations of dance. Before about 200 BC, dances were evidently in the form of choral processions only. There were seed-sowing processions in spring, headed by priests, and heavy and majestic weapon dances of the Salii, a congregation of the priests of Mars who walked around in a circle while rhythmically beating their shields.

Antagonism of the Romans

Later, Greek and Etruscan influences began to spread, though people who danced were considered suspicious, effeminate, and even dangerous by the Roman nobility. One public official did not believe his eyes when he watched dozens of the daughters and sons of well-respected Roman patricians and citizens enjoying themselves in a dancing school. About 150 BC all dancing schools were ordered closed, but the trend could not be stopped. And though dance may have been alien to the Roman's inner nature, dancers and dancing teachers were increasingly brought from abroad in the following years. The statesman and scholar Cicero (106–43 BC) summed up the general opinion of the Romans when he stated that no man danced unless he was insane.

A form of dance that enjoyed great popularity with the Romans under the emperor Augustus (63 BC–AD 14) was the wordless, spectacular pantomime that rendered dramatic stories by means of stylized gestures. The performers, known as *pantomimi*, were at first considered more or less as interpreters of a foreign language, since they came from Greece. They refined their art until the two dancer-mimes Bathyllus and Pylades became the star performers of Augustan Rome. In contrast to the acrobatic dancers who performed with the jugglers, these *pantomimi* were taken quite seriously by the Romans.

CHRISTIANITY AND THE MIDDLE AGES

Dancing was traditional also to the tribes of barbarians to the north, as attested by the writings of the Christian missionaries. Wherever they went, they found the same fertility-rite dances—if in different guise—the same charm dances to induce good and ward off evil, the same warrior and weapon dances to bolster fighting morale, and the same uncontrolled expressions of the joy of life, which the missionaries attributed to the devil.

Erotic dancing was not the exclusive property of heathen societies. In Byzantium, the Christian emperor Justinian I (483–565) married the notorious Theodora, a dancer who had appeared in the nude in theatrical performances. About 500, St. Caesarius of Arles reported a sacrificial banquet ending in some demoniacal dancing rites performed to the accompaniment of lewd songs. The Anglo-Saxons had little girls performing dances at Easter in which a phallus was carried in front of them.

Ecclesiastical attitudes and practices. The attitude of the Christian Church toward dance was by no means unanimous. On the one side there was the ascetic rejection of all manifestations of lust and ecstasy, and dance was seen as one of the strongest persuasions to sexual permissiveness. On the other side, some early Church Fathers tried to find functions for pagan dances in Christian worship. Thus St. Basil of Caesarea in 350 called dancing the most noble activity of the angels, a theory later endorsed by the Italian poet Dante (1265–1321). St. Augustine (354–430) was strictly against dancing of any kind, but despite his great influence in the medieval church, dancing in churches continued for centuries.

Ambivalence of the church

Charlemagne, the Holy Roman emperor at the beginning of the 9th century, officially prohibited all kinds of dancing, but the ban was not observed. The Teutonic

peoples were accustomed to dancing as part of their religious rites. On Christian feast days, which coincided with their ancient rites of expelling the winter, of celebrating the arrival of spring, and of rejoicing that the days grew longer again, they revived their old dances, though they were camouflaged with new names and executed to different purpose. In this manner the dances became more and more secularized. After such secularization, two lines of development were open: the social dance or the assimilation of dance into theatrical spectacle by the *joculators*, travelling comedians who combined the arts of dancer, juggler, acrobat, singer, actor, mime, and musician in one person.

Dance ecstasies. There were two kinds of dance peculiar to the Middle Ages, the dance of death, or danse macabre, and the dancing mania known as St. Vitus' dance. Both originally were ecstatic mass dances, dating from the 11th and 12th centuries. People congregated at churchyards to sing and dance while the representatives of the church tried in vain to stop them. In the 14th century another form of the dance of death emerged in Germany, the *Totentanz*, a danced drama with the character of Death seizing people one after the other without distinctions of class or privilege. The German painter Hans Holbein the Younger (1497/98–1543) made a famous series of engravings of this dance.

The dance of death

The St. Vitus' dance became a real public menace, seizing hundreds of people, spreading from city to city, mainly in the Low Countries, in Germany, and in Italy during the 14th and 15th centuries. It was a kind of mass hysteria, a wild leaping dance in which the people screamed and foamed with fury, with the appearance of persons possessed. In these convulsive, frantic, and jerky dances, religious, medical, and social influences probably interacted in response to such things as the epilepsy-like seizures of persons suffering from the Black Death. Italy was afflicted with tarantism, an epidemic presumably caused by the bite of venomous spiders. Its effects had to be counteracted by distributing the poison over the whole body and "sweating it out," which was accomplished by dancing to a special kind of music, the tarantella.

Dance and social class. In western Europe by the 12th century, society had developed into three classes, the nobility, the peasantry, and the clergy. This separation contributed to the development of the social dance. The knights created their own worldly and spiritual ideals, exemplified in tournaments and courtly entertainments that were praised in song and poetry by the troubadours and minnesingers. The couple dances of the knights expressed the polished and aristocratic notions of courtly love. The round dances of the peasants were executed by circles or lines of people, often singing and holding each other by their hands. The rustic choral round had strong pantomimic leanings and unpolished expressions of joy and passion. And while the choral rounds almost always were executed to the singing of the participants, the court dances of the knights generally were accompanied by instrumental playing, especially of fiddles, and when there was singing, it emerged from the spectators rather than the performers.

Courtly couples and rustic rounds

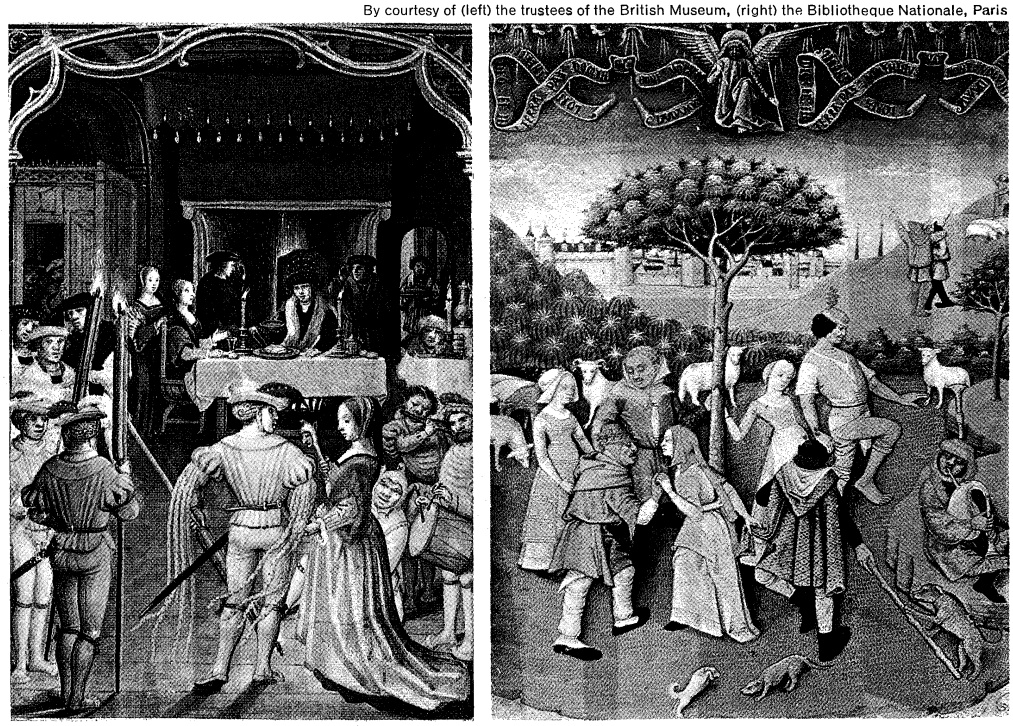
From the late Middle Ages, graphic artists frequently recorded what dancing looked like in all its different manifestations. How dancing adapted to the idealism of knightly love is shown in manuscript illuminations and tapestries. Paintings of the Flemish painter Pieter Bruegel the Elder (*c.* 1525/30–69) leave no doubt that the peasants enjoyed their uninhibited stamping and cavorting.

THE RENAISSANCE WORLD AND THE ART DANCE

France had set the fashion in court dance during the late Middle Ages; with the Renaissance, however, Italy became the centre of the new developments in dance. The Renaissance brought greater mixing of social classes, new fortunes and personal wealth, and greater indulgence in worldly pleasures and in the appreciation of the human body. The period emerged as one of the most dance-conscious ages in history.

Court dances and spectacles. Celebrations and festivities proliferated. The itinerant jugglers of the Middle Ages became highly respected and much sought after as dancing masters. They quickly assumed the function of instructing the nobility not only in the steps but also on posture, bearing, and etiquette. They became responsible for the planning and realization of the spectacular festivities. The social prestige of this newly developing profession grew constantly.

By courtesy of (left) the trustees of the British Museum, (right) the Bibliotheque Nationale, Paris



Late medieval court couple dance and peasant round dance.
(Left) Torch dance from the *Golf Book*, attributed to Simon Bening, school of Bruges, *c.* 1500. In the British Museum (Add. MS. 24,098). (Right) Peasant round dance from the *Hours of Charles d'Angoulême*, French, late 15th century. In the Bibliothèque Nationale, Paris (MS. lat. 1173, fol. 20 v).

The rise of the dancing master

Some of these dancing masters were highly learned men, and their treatises leave no doubt about their scholarly ambitions. Many of them were Jewish, descended from the Klesmorim, a group of medieval Jewish entertainers. The first dancing master known by name was Domenico da Piacenza, who in 1416 published the first European dance manual, *De arte saltandi et choreas ducendi* ("On the Art of Dancing and Directing Choruses"). His disciple, Antonio Cornazano, a nobleman by birth, became an immensely respected minister, educator of princes, court poet, and dancing master to the Sforza family of Milan, where about 1460 he published his *Libro dell'arte del danzare* ("Book of the Art of the Dance"). Such books record little about the actual steps and the melodies to which they were performed, but they are eloquent in the description of the *balli*—works that were invented by the dancing masters themselves. Adapting steps from the various social dances, they used them in a kind of dance pantomime.

In France, numerous forms developed from the branle, a round dance of peasant origin that became fashionable in the courts. One of the most frequently mentioned of all the dances of the 15th century was the *morisca*, or *moresque*, a romanticized version of dances from Moorish Spain. These were first mentioned in 1446 by a Bohemian traveller who visited Burgos, Spain. Later, in Portugal, he encountered similar forms. Sometimes religious motifs of the legendary fight between Charlemagne and the Turkic invader Timur entered the *morisca*, but usually it was performed as a double-file choral dance. It had nothing to do, as was long believed, with the English masked Morris dance, now considered to be a survival from a primitive religious cult.

From such choral dances the ballet emerged. At the court entertainments throughout Savoy and northern Italy, sumptuous spectacles with mythological, symbolical, or allegorical content became increasingly popular. At these early stages, however, pantomime and dance are not easily distinguished. Famous examples of these spectacles are the presentation of the story of Jason and the Golden Fleece at the marriage of Philip the Good of Burgundy in 1430, and the dinner ballet on the same, though widely enlarged, subject staged for the wedding of the Duke of Milan in 1489.

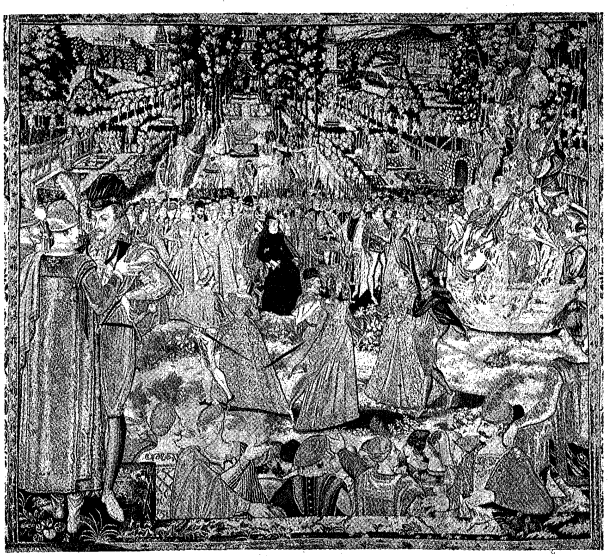
Tudor England of the early 16th century had similar pageants, with the participants disguising "after the manner of Italie." Like the Italian *balli*, the English masque offered an almost unlimited choice of performing variations, from a simple dance in masks to the most elaborate spectacle interspersed with songs, speeches, and pantomimes. As for the actual dances, Robert Copland's *Maner of Dauncynge of Bace Daunces after the Use of Fraunce*, published in 1521 as an appendix to a French grammar, leaves no doubt that the English upper class of that period was thoroughly familiar with continental dance. But whereas the nobility preferred dances of slow, measured, and dignified stature, stylishly performed and modelled upon the standards of the French court, the peasants continued their boisterous dancing, in England as elsewhere, very much as they had for centuries.

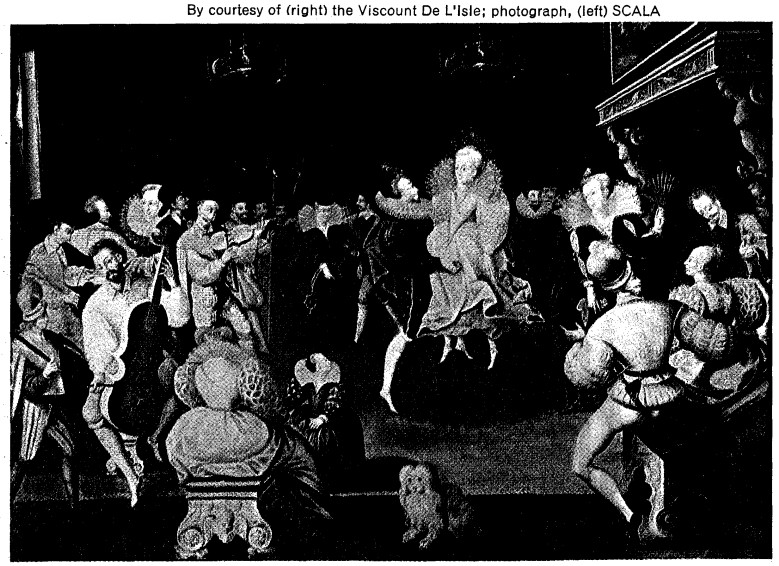
In England in the late 16th century, Queen Elizabeth I gave dancing a further boost. She was a skilled practitioner of the galliard and the volta, with its tight embraces by high-leaping couples. She enjoyed watching the English country dances—the chain, ring, and round dances of ancient origin and constantly new invention. These dances apparently provided a continuous infusion of new vitality into court dances. The nobles vied with one another in the execution of the jig, a sprightly and swift dance of "the folk" accompanied by songs. Dancing schools flourished everywhere in London, giving public displays and contributing considerably to the reputation of "the dancing English." Another extremely important contribution to dance was provided by Spain, which in the late 16th and early 17th centuries enjoyed a cultural renaissance. It was the "golden age" of Cervantes in literature, of Lope de Vega and Pedro Calderón de la Barca in the theatre, of El Greco and Diego Velásquez in painting. With the growth of Spain's empire in the Americas, dances of Afro-American origin found their way back to Europe. The sarabande and the chaconne were brought from Central America before 1600. Both were considered outspokenly obscene in their suggestions of sexual encounters. They became extremely popular in the harbours of Andalusia, where they were polished and their pantomimic literalness somewhat moderated. From there they crossed the Pyrenees and were integrated into the canon of the French court dance.

Spread of Spanish dance

Other dances from abroad played major roles in the shaping of Spain's national dances. The *canarie* of African origin certainly sired the Aragonese jota, while the sarabande brought forth the seguidilla. The Afro-Cuban *chica* lived on in the fandango, and the flamenco dances of the Andalusian Gypsies retained their Moorish heritage into the 20th century. It can be presumed that this exchange of dances was not a one-way traffic, that the European conquerors and colonists similarly influenced the dancing habits of the people in other lands.

By courtesy of (right) the Viscount De L'Isle; photograph, (left) SCALA





Renaissance dances.
(Left) Court dance in early balletic form as seen in "Catherine de' Medici Receiving the Polish Ambassador," tapestry designed by François Quesnel, *c.* 1575. In the Uffizi, Florence. (Right) "Queen Elizabeth Doing a Leaping Turn of the Lavolta, Assisted by the Earl of Leicester," oil painting by an unknown artist, 16th–17th century. In the collection of the Viscount De L'Isle, Penshurst Place, Kent.

The birth of ballet. Meanwhile, dance became the subject of serious studies in France. A group of writers calling themselves La Pléiade aimed for a revival of the theatre of the ancient Greeks with its music, song, and dance. In Catherine de Médicis (1519–89), the Florentine wife of Henry II, the Italian dancing masters found an influential sponsor in Paris. She called to Paris the Italian musician and dancing master Baltazarini di Belgioioso, who changed his name to Balthazar de Beaujoyeulx (early 16th century to 1587). There had been previous fetes in both France and Italy that offered masquerades, pantomimes, and dances with allegorical and symbolical subjects, but none of them compared to the splendours of the *Ballet comique de la reine* that Beaujoyeulx staged in 1581 for Catherine.

Ballet comique de la reine

This "ballet" told the story of the legendary sorceress Circe and her evil deeds. Spoken texts alternated with dances amid magnificently decorative settings. The performers, recruited from the nobility, moved on the floor more like animated costumes than individual dancers. They came together in strikingly designed groups, and they set up geometrical floor patterns that had highly symbolic meanings. (To audiences of the period, for example, three concentric circles represented Perfect Truth, and two equilateral triangles within a circle stood for Supreme Power.) The ballet, which ended in an act of homage to the royal majesties present, had a distinct political moral. Circe had to render her might to the absolutist power of the king of France as the supreme symbol of a peaceful and harmonious world.

The *Ballet comique* launched the species known as *ballet de cour*, in which the monarchs themselves participated. The idealized dances represented the supreme order that France itself, suffering from internal wars, lacked so badly. The steps were those of the social dances of the times, but scholars became aware of how these native materials might be used to propagate the Greek revival. They thoroughly analyzed and systematized the dances, and in 1588 the priest Jehan Tabourot, writing under the pen name Thoinot Arbeau, published his *Orchésographie*, which he subtitled "a treatise in dialogue form by which everyone can easily learn and practice the honest exercise of the dance." This was the first book containing reliable descriptions of how, and to what kind of music, the *basse danse*, pavane, galliard, volta, courante, allemande, gavotte, *canarie*, *bouffon*, *moresque*, and 23 different variations of the branle were performed.

During the 17th, 18th, and 19th centuries

Under kings Louis XIV and Louis XV, France led western Europe into the age of the Rococo in the arts. The Rococo began as a movement toward simplicity and naturalness, a reaction against the stilted mannerisms and preciousness to which the earlier Baroque art was considered to have degenerated. It was a great age of and for dancing, with the minuet the symbol of its emphasis on civilized movement. This formal dance, the perfect execution of which was almost a science in itself, reflected the Rococo idea of naturalness. The statement that "the dance has now come to the highest point of its perfection" by the composer Jean-Philippe Rameau (1683–1764) suggested how conscious the French were of the great strides dance had made. That this was particularly the case in France was confirmed by the English poet and essayist Soame Jenyns (1704–87) in his lines "None will sure presume to rival France, / Whether she forms or executes the dance." None, however, excelled the estimation of his profession by the dancing master in Molière's *Le Bourgeois Gentilhomme* (1670):

> There is nothing so necessary to human beings as the dance . . . Without the dance, a man would not be able to do anything . . . . All the misfortunes of man, all the baleful reverses with which histories are filled, the blunders of politicians and the failures of great leaders, all of this is the result of not knowing how to dance.

THE MATURING OF BALLET

Dance was finally deemed ready for an academy of its own. In 1661, 13 dancing masters who had been members of a professional guild of medieval origin, together with some musicians, composers, and the makers of instruments, were granted a charter by Louis XIV for the Académie Royale de Danse.

Technical codifications and dance scholarship. The academicians were charged with setting up objective standards for perfecting of their arts, with unifying the rules of dance training, and with issuing licenses to dancing instructors. Though the nobility continued for some time to participate in the *ballets de cour*, and Louis himself danced in them until 1669, the dance became more and more the province of highly trained specialists.

After 1700 ballet and social dance took separate paths. But while the ballet continued to absorb new ideas from the folk and social dance, its practitioners and theoreticians looked down on those more common forms. A profusion of books on dance began to appear—treatises, instructions, and analyses as well as the first attempts to record dances by means of written notation. The first history of dance was Claude-François Menestrier's *Des ballets anciens et modernes* ("On Dances Ancient and Modern"; 1682). The second major work of European dance literature, after Arbeau's *Orchésographie*, was Raoul Feuillet's *Chorégraphie, ou l'art de décrire la danse* ("Choreography, or the Art of Describing the Dance"; 1701). It became the standard grammar for the dances practiced at the turn of the century, describing them in minute detail and notating them by a system devised by Feuillet. This gave the position of the feet and directions, combinations, and floor patterns of the steps and leaps. It was unable, however, to register the movements of the upper parts of the body. Feuillet provided as well a complete definition of the principles of the dance first described by the Académie in the 1660s. These included the *en dehors* (*i.e.*, the turnout of the body and its limbs), the five classical positions of the feet, the port de bras (*i.e.*, the positions and movements of the arms), and the leaps to the *grande élévation*, the aerial movements of the dance.

Early works about ballet

In 1706 Feuillet's book was translated into English by John Weaver (1673–1760), a dancer, choreographer, and teacher who worked mainly at the Drury Lane Theatre, London. In 1717 he produced one of the first serious ballets without words, *The Loves of Mars and Venus*. Weaver was the first dance teacher to insist that dance instructors should have a thorough knowledge of anatomy. In 1721 he published his *Anatomical and Mechanical Lectures upon Dancing*, which became a standard work of international importance. Germany also was represented in the field of dance scholarship, most notably by Leipzig Gottfried Tauber in *Der rechtschaffene Tanzlehrer* ("The Correctly Working Dance Teacher"; 1717). These books strongly emphasized the contributions of dance to general education and manners. In this period dance was considered the basis of all education, and well-to-do parents went to great pains to have their children properly instructed.

Varieties of the ballet. As the technical demands of performance became greater and the amateurs gave way to the professionals, the ballet moved from the dance floor onto the stage. There it gradually shed its declamations and its songs and concentrated on telling a story through dance and mime alone. But this purifying process took time. For decades different forms of mixed-media spectacles were seen, from the *comédie ballets* of Molière (1622–73) and the composer Jean-Baptiste Lully (1632–87) to the *opéra-ballets* of André Campra (1660–1744) and Rameau, which were successions of songs and dances on a common theme. The first ballet without the diversions of speech or song was *Le Triomphe de l'amour* (*The Triumph of Love;* 1681), choreographed by Charles-Louis Beauchamp (1636–c. 1719) to Lully's music. Originally a *ballet de cour*, it was revived for the stage with a professional cast. Its star, Mlle Lafontaine, became ballet's first *première danseuse* exactly 100 years after the *Ballet comique* had been produced.

An even more dramatic form known as *ballet d'action* came into being in 1708, when two professional dancers presented an entire scene from the tragedy *Horace* by

Pierre Corneille (1606–84) in dance and mime. Weaver's silent ballets, whose expressive dance much impressed English audiences, also encouraged Marie Sallé, a highly ambitious dramatic dancer. Despairing of the *opéra-ballets* of Paris, she went to London, where she performed in pantomimes and produced a miniature dance-drama of her own, *Pygmalion* (1734). In it she appeared in a

Giraudon



"Marie-Anne Cuppi," called "La Camargo, Dancing," oil painting by Nicolas Lancret, 1730. In the Musée des Beaux-Arts de Nantes, France.

flimsy muslim dress and loose, flowing hair rather than the heavy costumes and elaborate wigs usually worn by ballerinas. Thus lightened, the dancer was able to move with much greater freedom.

Sallé, Camargo, and Campanini

Early virtuosos of the dance. The era of the great dancer was at hand. Marie Sallé (1707–56) was the greatest dancer-mime and an important innovator of her day. Her popularity was rivalled by the Brussels-born Marie Camargo (1710–70), who excelled Sallé in lightness and sparkle. She used the entrechat, a series of rapid crossings of the legs that previously had been used only by male dancers. To show off properly her entrechats and other lithe footwork, she shortened her skirt by several inches, thereby contributing to costume reform. Both ballerinas were depicted by Nicolas Lancret (1690–1743), a painter known for his festive scenes, and both were praised by the writer and philosopher Voltaire (1694–1778), who carefully compared their respective virtues. Both, however, were surpassed by the Italian dancer Barberina Campanini (1721–99), whose fame is less adequately recorded in dance history. By 1739, she had taken Paris by storm, demonstrating jumps and turns executed with a speed and brilliance hitherto unknown. She offered ample proof that the Italian school of dance teaching had by no means died out with the earlier exodus of so many of its best practitioners to the French courts. Despite the great public acclaim that these ballerinas attracted, they were overshadowed by Louis Dupré (1697–1744), known as "The Great Dupré" and "the god of the dance." In grace, majesty, and allure, he was unsurpassed, giving the male dancer a prominence he held for a century. Dupré was also the first of a direct line of great dance teachers that was unbroken in the late 20th century.

THE REIGN OF THE MINUET

In the realm of the social dance, the years between 1650 and 1750 were called "the age of the minuet" by the dance and music historian Curt Sachs.

The French dance suite. At the great balls of the French court at Versailles, the minuet was the high point of the festivities, which culminated in a suite of dances. The opening branle, led by the king and his escort, was a measured circling around, one couple after another. Next came the courante, which had been toned down from its earlier rather capricious figurations. It assumed a continuously greater dignity until it was danced with such gravity and sobriety that it was termed the "doctor dance." It went quickly out of fashion, however, after 1700. Following it in the succession was the gavotte, which opened in the form of a round dance. A couple separated to perform a short solo, then returned to the original circle. Sometimes the suite was extended through an allemande (French: "German"), an old dance form that was introduced into France from the heavily German-speaking province of Alsace in the 1680s. This dance, with its turning couples, the lady on the arm of the gentleman, was a relative of the German ländler and a precursor of the waltz.

The dance suite

Form of the minuet. But the unrivalled king of the social dances was the minuet, named from the *pas menu* ("small step"), a term used at least as early as the 15th century. The earliest surviving specimen was composed by Lully in 1663. Mozart composed a series of twelve minuets as late as 1789. It originated as a folk dance in Poitou, but as a court dance it took its form from the courante. Though today it looks mannered, even artificial, in its time it was looked upon as the most beautiful and harmonious of dances, and to execute it perfectly required prolonged and careful study:

> The *minuet* was performed in open couples; spectators and partners were saluted with ceremonial bows. With dainty little steps and glides, to the right and to the left, forward and backward, in quarter turns, approaching and retreating hand in hand, searching and evading, now side by side, now facing, now gliding past one another, the ancient dance play of courtship appears here in a last and almost unrecognizable stylization and refinement. (Curt Sachs, *World History of the Dance*, trans. Bessie Schönberg, W.W. Norton & Co., Inc., 1937.)

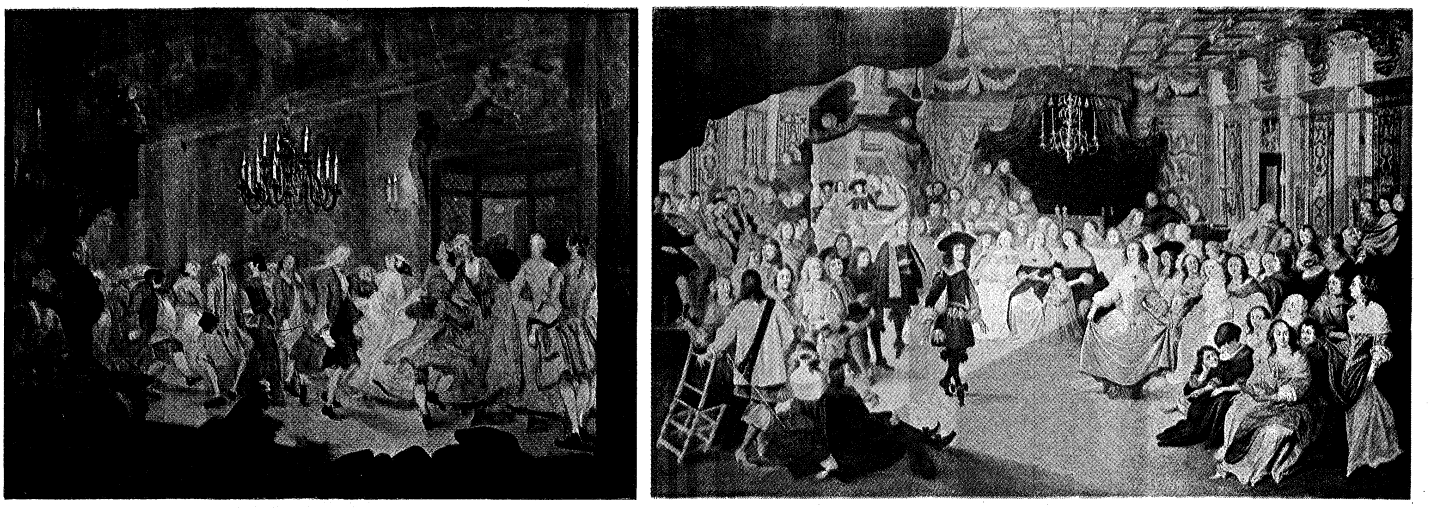
In spite of the great popularity of the minuet before the French Revolution, it was the object of much barbed commentary in the late 18th century. Voltaire compared the metaphysical philosophers of his time with the dancers of the minuet, who, in their elegant attire, bow and mince daintily across the room showing off their charms, move without progressing a single step, and end up at the very spot from which they began.

ENGLISH SOCIAL DANCE

England thoroughly democratized the dance. Though the English Puritanism of the 17th century stigmatized dance as one of man's most sinful occupations, even Oliver Cromwell, lord protector of England under the Puritan rule in the 1650s, could not prevent the appearance of *The English Dancing-Master* (1651), by the bookseller and publisher John Playford (1623–*c.* 1686). This was a collection of English traditional dances and tunes. It had 18 editions in 80 years, each one adding to the repertoire. Its 900 choral dances of rustic origin, which formerly had been danced in the open air but were now usually performed indoors, included an enormous variety of forms and patterns. It was written in straightforward, matter-of-fact language, with no discrimination of dances by social class. Its instructions could be understood and its dances performed by anyone. People could enjoy dancing as a playful, sportive activity rather than as an exercise of courtly etiquette.

These "country dances" could as well be city dances, as is suggested by such names as "Mayden Lane" and "Hide Park" from London locales. Others were named for persons—"Parson's Farewell" and "My Lady Foster's Delight"—and that there were foreign influences can be surmised from "The Spanish Jeepsie" and "A la Mode de France." At the same time, native jigs and hornpipes continued to flourish. The English were particularly fond of the Morris dance. This dance may have received its name from the blackened faces of some of its participants, suggestive of the African Moors, but its origins were in the ancient ritual dances. It was a vigorous male dance, in the form of a dance procession through town streets. Its participants, in the disguises

The Morris dance



17th- and 18th-century dance.
(Left) Contredanse from "Masked Ball at Wanstead Assembly," oil painting by William Hogarth, *c.* 1745. In the South London Art Gallery. (Right) Courante from "Charles II at a Ball at The Hague," oil painting by Hieronymus Janssens (*c.* 1624–93). In Windsor Castle, Berkshire.
By courtesy of (left) the Southwark Borough Council, South London Art Gallery; (right) Copyright reserved

of such popular characters as the fool or the Queen of May, wore jingling bells around their ankles and sometimes galloped about on hobby horses. Other dancers wore antlers, tails, and similar animal masking.

About 1700 the English country dances began to appear on the Continent, where they were somewhat formalized and sometimes substantially altered. In France they were named contredanses. The longways, dances with double lines of dancers facing one another, became *contredanses anglaises;* the rounds became the *contredanses françaises,* which were also known as cotillions and quadrilles. These figure dances, which quickly spread to Spain, Germany, Poland, and other countries, were the dances of the rising middle class. By no means revolutionary in their content, they were nonetheless a distinct declaration of rationality and common sense in dance, a counterbalance to the artificialities and mannerisms of the aristocratic court dances. The orthodox dance teachers might bemoan the decline from the standards that were epitomized in the minuet, but the townspeople and peasants, unconcerned with such niceties, continued in their uncomplicated knowledge that dancing could be fun.

DANCE IN COLONIAL AMERICA

Attitudes. The English colonists in America had mixed opinions about dance. There was the complete disapproval of those who saw only its inherent licentiousness, but from others came at least a tacit toleration of the obviously irrepressible urge to dance. The South, more heavily populated by colonists with aristocratic backgrounds, was generally more inclined to dance than the North, where religious fervour had motivated much of the migration from England. But what was allowed and even encouraged in Connecticut was strictly forbidden in Massachusetts. The general consensus was apparently that dancing in itself was not bad, but that no punishment could be severe enough for what was regarded as lascivious dancing. The Quakers, mainly in Pennsylvania, were very much against dancing, and in 1706 they complained bitterly about a dancing and fencing school being tolerated in Philadelphia, fearing it would tend to corrupt their children.

External and internal influences. Nonetheless, Playford's *The English Dancing-Master* was by no means unknown in America. There were also dancing masters and dancing mistresses to instruct in and lead the dances that had been brought from the Old World. There were society balls in the cities along the coast, and on the inland frontiers the settlers of the widely scattered farmsteads often came together for exuberant feasting and social dancing. Here dancing was considered a socializing virtue expressed in this anonymous observation:

> I really know among us of no custom which is so useful and tends so much to establish the union and the little society which subsists among us. Poor as we are, if we have not the gorgeous balls, the harmonious concerts, the shrill horn of Europe, yet we delight our hearts as well with the simple negro fiddle.

Slave dances

What the colonists saw of American Indian dancing they found very strange and primitive, and there was virtually no exchange of dancing customs between the groups. The situation differed, however, with regard to the Negro slaves, who in the 17th century had brought their own songs and dances from their native lands in Africa.

Religious holidays in New Amsterdam saw dancing in the streets, accompanied by three-stringed fiddles and drums constructed from eel pots and covered with sheepskins. Dutch families joined in the festivities, dancing with the Negroes. When New Amsterdam became New York, however, the English discouraged dancing between whites and blacks; the Negro went on to develop the characteristic dance style that would so deeply affect social dancing in the 19th and 20th centuries.

Early in the 18th century, rather rough theatrical entertainments, acts of acrobatic skill or pantomimes in which dances played an increasing role, began to spread through the American colonies. These often amateurish showings

By courtesy of the Museum of Fine Arts, Boston



"Rustic Dance after a Sleigh Ride," oil painting by William Sidney Mount, 1830. In the M. and M. Karolik Collection of the Museum of Fine Arts, Boston.

got a mighty boost when the first professional companies came from Europe, about the middle of the century, to perform plays and harlequinades with incidental dances.

THE RISE OF THE WALTZ

The age of the minuet was followed by that of the waltz. As the French Revolution approached, the minuet, a form that exuded the essence of earlier decades, died a natural death. The English country dances, expressing the self-satisfaction of the bourgeoisie, fared little better.

The Romantic movement in dance. The young people, whose preferences led the way in creating new forms, had lived through the revolutionary events of the 1780s and '90s. They now looked to dance as a way to unleash deeper emotion, to satisfy the needs of body and soul, and to mobilize more vital and dynamic expression than that permitted by the sober and decorous rules of the dancing masters. The overflow of feeling and the striving for horizons broader than those understood by the traditional canons of French Rationalism were among the factors that generated the Romantic movement in the arts of Europe. This new direction was clearly expressed in the waltz, a dance filled with the Dionysian spirit.

Dionysius revived

Like much of the spirit of the Romantic movement, the waltz was of German origin. It paralleled the Sturm und Drang movement in German literature, which featured the new forms of prose and poetry by Johann von Goethe and Friedrich Schiller. One of the most glowing advocates of the waltz was Goethe, who time and again praised it, nowhere more than in his novel *Die Leiden des Jungen Werthers* (1774; *The Sorrows of Werter*, 1779): "Never have I moved so lightly. I was no longer a human being. To hold the most adorable creature in one's arms and fly around with her like the wind, so that everything around us fades away." Even the aristocrats who formed the Congress of Vienna in 1815, which sought to restore law and order to Europe following the upheavals brought on by Napoleon, delighted in performing this earliest of all nonaristocratic ballroom dances.



Waltzers in "Le Bal à Bougival," oil painting by Pierre Renoir, 1883. In the Museum of Fine Arts, Boston.
By courtesy of the Museum of Fine Arts, Boston

Spread of the waltz. The waltz started as a turning dance of couples. It was especially popular in south Germany and Austria, where it was known under such different names as *Dreher*, ländler, and *Deutscher*. More than any other dance it appeared to represent some of the abstract values of the new era, the ideals of freedom, character, passion, and expressiveness. This may explain somewhat its eruption into the limelight of international popularity. This popularity was scaled in 1787 when it was brought to operatic stage. Vienna became the city of the waltz, for there it surpassed everything in wild fury. It swept over national frontiers, and in 1804 the French were reported to be passionately in love with this light, gliding dance. "A waltz, another waltz" was the common cry from the ballroom floor, for the French could not get enough of the dance.

Some guardians of the public morality disapproved of the "mad whirling" of the waltz and it did not arrive in England until 1812. At the Prussian court in Berlin it was forbidden until 1818, though Queen Luise had danced it while still a princess in 1794. The guardians could do no more than delay its total victory, and it conquered the world without sanction of courts, of dancing masters, or of other powers. After many centuries of leadership, France no longer set the fashions. In 1819 Carl Maria von Weber's *Invitation to the Dance* represented the declaration of love of classical music to the waltz. Shortly thereafter began the age of the Viennese waltz kings, most notably expressed by the Strauss family.

Offspring and rivals. The waltz sired a great variety of offspring throughout Europe. Germany developed such variations of the waltz as the schottisch, with turns like those of the waltz. France had its airy *balance valse*, and the Americans later on had their Boston waltz, a slower, gliding variant. About 1840 a serious rival to the waltz emerged in the polka, a Bohemian dance that took its name from the Czech word *půlka*, "half step." It was full of fiery vigour. Another Bohemian folk dance finding favour in the dance halls was the *rejdovák* or redowa, while Poland's mazurka and krakowiak enjoyed great popularity. No ball could be concluded without a galop, in which couples galloped through the hall with accelerated polka steps, an exhausting exercise that required considerable reserves of stamina.

FOUNDATIONS OF MODERN BALLET

The ideals of naturalness, character, soul, passion, and expressiveness came to govern the ballet.

Noverre and his contemporaries. The French dancer-choreographer-teacher Jean-Georges Noverre (1727–1810), who became known as the "Shakespeare of the dance," was the first major reformer of ballet. He defined his artistic positions in *Lettres sur la danse, et sur le ballet* (*Letters on Dancing and Ballets*), published in 1760 and continuously reprinted ever since. He worked in Paris, London, Stuttgart, and Vienna, and his influence spread as far as St. Petersburg. He preached the dignity of the ballet and tried to purge it of its excessive artificialities and conventions. He choreographed subjects of mythology and history in highly dramatic narrative forms. He collaborated with some of the major composers of the period, including Mozart, on his ballets.

The influence of Noverre

Noverre was not alone, and the others around him were full of the same zest to give a new meaning to ballet. In Vienna he had a feud with the Italian choreographer Gasparo Angiolini (1731–1803) over Noverre's reforms of the *ballet d'action*. Angiolini claimed these for his teacher, the Austrian choreographer Franz Hilverding (1710–68). In Bordeaux, Noverre's pupil Jean Dauberval premiered in 1789 *La Fille mal gardée* (*The Ill-Guarded Maiden*), usually called *Vain Precautions* in English, which became the first durable ballet comedy. It introduced the *demi-caractère* dance, which featured what were considered to be "true-to-life" characters. In London, still another pupil, Charles Didelot, created in 1796 *Flore et Zéphyre*. This was the first attempt to bestow on the individual dances within the ballet a certain

period and local coloration, and to break the uniformity of step and movement of the corps de ballet by assigning individual tasks to its various members. Later, Didelot thoroughly reformed the ballet school in St. Petersburg, which had existed since 1738. There he also created the first ballets of the Russian national repertory. Among these were the first ballets to be based on the works of the Russian writer Alexandr Pushkin (1799–1837), whose stories continued to be used as ballet libretti for many decades.

In Milan, Salvatore Viganò, who had worked under Dauberval and Didelot and who had choreographed in 1801 the first performance of Beethoven's *Creatures of Prometheus*, was praised by the French writer Stendhal for his thrilling ballets based, among other subjects, on Shakespeare's *Othello* and *Coriolanus*. He was followed by Carlo Blasis, who was more noted as a teacher and theoretician. His *Traité élémentaire, théorique, et pratique de l'art de la danse* (1820; *Elementary Treatise upon the Theory and Practice of the Art of Dancing*) became the standard work of ballet teaching for the 19th century. In 1837 he founded the Imperial Ballet Academy, through which Milan became, with Paris and St. Petersburg, a third ballet centre of world renown.

The Romantic ballet. During the 1830s and '40s the Romantic movement flooded ballet stages with nature spirits, fairies, and sylphids. The cult of the ballerina replaced that of the male dancer, whose last and greatest representative had been the Italian dancer Gaetano Vestris (1729–1808). The techniques of female dancing were greatly improved. Skirts were shortened further, and blocked shoes permitted toe dancing. Choreographers strove for a more expressive vocabulary and to highlight the individual qualities of their dancers.

La Sylphide (1836) stated a main subject of the Romantic ballet, the fight between the real world and the spiritual world. This theme was enhanced and expanded in *Giselle* (1841) and *Ondine* (1843). Paris and London were the taste setters, and it was London that in 1845 witnessed the *Pas de quatre*, for which the French choreographer Jules Perrot brought together, for four performances, four of the greatest ballerinas of the day, the Italians Marie Taglioni (1804–84), Carlotta Grisi (1819–99), and Fanny Cerrito (1817–1909), and Lucile Grahn (1819–1907). After this the decline of Romantic ballet was rapid, at least in these cities. It continued to flourish into the early 1860s, however, in Copenhagen under the choreographer Auguste Bournonville, whose repertoire was kept alive by the Royal Danish Ballet into the second half of the 20th century. Russia, under the French-born Marius Petipa (1819–1910) and his Russian aide Lev Ivanov (1834–1901), built a world-famous ballet culture of its own. Linked at first with Paris, it gradually developed its own balletic idiom from native as well as imported sources. The high point of the classical ballet under the tsars was reached with the St. Petersburg productions of *The Sleeping Beauty* (1890), *The Nutcracker* (1892), and *Swan Lake* (1895), all with music composed by Peter Tchaikovsky, and *Raymonda* (1898), composed by Aleksandr Glazunov (1865–1936). While the ballet prospered in St. Petersburg and Moscow, it waned in Paris. Its ballerinas even appeared in male roles, as in *Coppélia* (1870). In Milan the extravaganzas of Luigi Manzotti (1838–1905) offered anything but dancing while glorifying the progress of mankind through material discoveries and inventions. The 19th century also saw an unprecedented increase in travel and in cross-cultural influences. Many seemingly exotic dance styles arrived on the Western scene. Troupes from as far as India and Japan appeared at expositions in Paris and London, starting a lively interest in folk and ethnic dancing. Ballerinas of the Romantic ballet travelled from one European city to another, from Milan to London to Moscow. The Austrian dancer Fanny Elssler toured the Americas in the early 1840s for two years, visiting Havana twice. The great choreographers, too, went from city to city. The language of dance became a medium of international communication without regard for difference in geography or spoken language.

Russian rise, Western decline

THEATRE AND BALLROOM DANCE

Other dance entertainments of a lighter kind gained immense popularity during the 19th century. In Paris the all-female cancan became the rage. Its electrifying high kicks were shockingly exhibitionistic and titillating. After

Giraudon



Cancan with "Jane Avril Dancing," oil on cardboard by H. Toulouse-Lautrec, 1892. In the Louvre, Paris.

1844 it became a feature of the music halls, of revues, and of operetta. It was raised to musical prominence by operetta composer Jacques Offenbach (1819–80) and vividly depicted by the painter Henri de Toulouse-Lautrec (1864–1901). London enjoyed the Alhambra and Empire ballets, which were mostly classical ballets with spectacular productions. But it was America that provided the widest variety. There were patriotic spectacles, popular after the Revolutionary War, such as *The Patriot, or Liberty Asserted*, in which dance figured prominently.

More important and of longer range results were the minstrel shows, extravaganzas, burlesques, and vaudevilles. These represented a confluence of a wide assortment of dance and theatrical influences, especially from the Negro culture. White men affected black faces and black dances, and black men affected the faces and dances of the white. Dances were tap and soft-shoe, the buck-and-wing, and similar routines. Theatrical productions offered all kinds of dance, from European-imported ballets through entirely native exhibitions of female beauty verging on the striptease. American dancers began to establish reputations both in America and Europe. The ballerina Augusta Maywood (1825–76?) was the first American dancer to perform at the Paris Opéra.

During the 19th century there was also an enormous increase in the number of public ballrooms and other dancing establishments in the fast-growing cities of the West. Here were first encountered American imports such as the barn dance, then called the military schottische; the Washington Post, a very rapid two-step in march formation; and the cakewalk, which contorted the body to degrees previously unknown. For the first time Europe found in the New World a new infusion of blood

for its dancing veins. The tempo of the dances quickened, reflecting perhaps the quickening pace of life and the great social changes of the century.

The 20th century

Two trends were evident during the first years of the 20th century, before World War I. As if aware of some impending catastrophe, the wealthy society of Europe and the Americas indulged itself to the full in quicker waltzes and faster galops. At the same time, it tried to revive the minuet, gavotte, and pavane, producing only pale and lifeless evocations. There had hardly ever been such a frantic search for new forms, such radical questioning of values previously taken for granted, such a craze among the youth of all nations for individual expression and a more dynamic way of life. All the arts were deeply influenced by the rapid accumulation of discoveries in the physical and social sciences and an increasing awareness of social problems.

New directions in the dance

Overall, it was an incredibly lively time for the dance, which never before had generated so many new ideas or attracted so many people. The ballet was completely rejuvenated under the leadership of Russian impresario Sergey Diaghilev (1872–1929). It inspired some of the foremost composers and painters of the day, becoming the primary theatre platform for the most up-to-date work in the arts. Proponents of another reform movement, "modern dance," took their cue from the American dancer Isadora Duncan to strike in another way at the artificialities that Romantic ballet had generated. It took vigorous roots in Germany, where its expressionistic forms earned it the name *Ausdruckstanz* ("expressionistic dance"). The ballroom dances were thoroughly revolutionized through infusions of new vitality from South American, Creole, and Negro sources. With the overwhelming popularity of Afro-American jazz, the entire spirit and style of social dancing altered radically, becoming vastly more free, relaxed, and intimate through the following decades.

There was also a renewal of interest in the folk dances that had been the expressions of the common people in past centuries. This was fostered partly through special folk-dance societies, partly through various youth movements that saw that these dances might assist in shaping new community feelings. Theatrical dance of all kinds, from the highly stylized, centuries-old dances of the Orient to exhibitions of naked female flesh, reached new heights of popularity.

DIAGHILEV AND HIS ACHIEVEMENTS

The artistic consequences of Diaghilev's Ballets Russes were enormous. Diaghilev's interest in dance began while he was a member of a small circle of intellectuals in St. Petersburg who fought to bring Russia's arts onto the wider European scene. The painters Alexandre Benois and Léon Bakst were his earliest collaborators.

The Ballets Russes. The Russian ballet troupe that Diaghilev brought to Paris in 1909 boasted some of the best dancers from the imperial theatres in St. Petersburg and Moscow. They set all Paris ablaze. No living person could remember ballets of such quality. For the next 20 years the Ballets Russes, which never appeared in Russia, became the foremost ballet company in the West. Diaghilev, who never choreographed a ballet himself, possessed a singular flair for bringing the right people together. He became the focus of the ballet world, striving for the integration of dance, music, visual design, and libretto into a "total work of art" in which no one element dominated the others.

Between 1909 and 1929, the contributions of many of the finest dancers and choreographers and of some of the most avant-garde, style-setting painters and composers made the Diaghilev company the centre of creative artistic activity. The group became a haven for Russian dancers who emigrated after the 1917 Revolution. It was the first large, permanently travelling company that operated on a private basis and catered to a cosmopolitan Western clientele.

Michel Fokine (1880–1942) was the first choreographer to put Diaghilev's ideas into practice. He worked with contemporary composers, notably the Russian Igor Stravinsky (1882–1971) and the Frenchman Maurice Ravel (1875–1937). He drew also upon many eminent composers of the past, such as the Russians Aleksandr Borodin (1833–87) and Nicolay Rimsky-Korsakov (1844–1908), and the Pole Frédéric Chopin (1810–49). His major scenic artists were Benois and Bakst, whose contributions to theatrical design had influences beyond the sphere of ballet. Among his dancers were the Russians Anna Pavlova (1881–1931), who left after the 1909 season to dance with her own company throughout the West as well as the Orient, and Vaslav Nijinsky (1890–1950), who succeeded Fokine as the company's choreographer. A classic dancer, Nijinsky was an anticlassic choreographer, specializing in turned-in body movements and in unusual footwork. His composers included Stravinsky and the French Impressionist Claude Debussy (1862–1918).

The work of Fokine

After Nijinsky's career was cut short by his insanity, the dancer Léonide Massine (1896–) assumed the role of choreographer. He quickly became noted for his wit and the precisely characterizing gestures of his dancers. His musical collaborators included Stravinsky; Manuel de Falla (1876–1946), whose work was full of the flavour of his native Spain; Ottorino Respighi (1879–1936), noted for his musical evocations of Italian landscapes; and Erik Satie (1866–1925), a Frenchman known for his originality and eccentricity. Massine's designers included leading painters of the School of Paris such as André Derain (1880–1954) and Pablo Picasso (1881–1973). Following Diaghilev's death, Massine created a furor in the 1930s with his ballets based on symphonies by Tchaikovsky and Johannes Brahms.

Another of Diaghilev's choreographers was Nijinsky's sister, Bronisława Nijinska (1891–1972), who became famous for her massive ensemble groupings and her talent for depicting the follies of contemporary society. Diaghilev's last choreographic discovery was the Russian-trained George Balanchine (1904–). Balanchine's *Apollon Musagète* (1928) began his long association with Stravinsky and led the way to the final enthronement of neoclassicism as the dominant choreographic style of the following decades.

The continuing tradition. When Diaghilev died his was no longer the only ballet company touring the world. Anna Pavlova's company visited places in Europe, the Americas, Australia, and the Orient that had never heard of, let alone seen, ballet. A troupe assembled by Ida Rubinstein (1885–1960) had Nijinska as a choreographer and Stravinsky and Ravel as composers. The Ballets Suédois featured, from 1920 to 1925, another group of avant-garde, largely French and Italian composers, painters, and writers. New dancers came from the schools in Paris, London, and Berlin that were directed by self-exiled Russian teachers. Important developments took place in London, where Dame Marie Rambert (1888–), a Diaghilev dancer, founded the Ballet Rambert, and Ninette de Valois (1898–) founded the company that became in 1956 the Royal Ballet. In New York, Balanchine set up the School of American Ballet in 1934. From it he drew the dancers for the several companies that led ultimately the founding of the New York City Ballet in 1948.

The Soviet ballet. Although Diaghilev's achievements were ignored there, the Soviet Union in the 1920s abounded with the daring choreographic experiments of Fyodor Lopukhov (1886–1973) and others. Despite the official imposition of "socialist realism" as the criterion of artistic acceptability in 1932, ballet gained enormous popularity with the Soviet people. They loved their dancers, who were superbly trained by generations of teachers under the leadership of Agrippina Vaganova (1879–1951).

MODERN DANCE

Despite the recovery of ballet from its sterility in the late 19th century, other dancers questioned the validity of an art form so inescapably bound to tradition by its relative-

ly limited vocabulary. They wished to change radically the culture concept of expressive stage dancing. In a period of women's emancipation, women stepped to the front as propagandists for the new dance and toppled the conventions of the academic dance. They advocated a dance that arose from the dancer's innermost impulses to express oneself in movement. They took their cues from music or such other sources as ancient Greek vase paintings and the dances of Oriental and American Indian cultures.

American and German schools

The pioneers of this new dance were Isadora Duncan (1878–1927), who stormed across European stages in her loosely flying tunic, inspiring a host of disciples and imitators, and Ruth St. Denis (1877–1968), who surprised American and European audiences with her Oriental-style dances. With her partner Ted Shawn (1891–1972) she founded Denishawn in 1915, which, as a school and performing company, became the cradle of America's early protagonists of modern dance; notable among them were Martha Graham (189?–), Doris Humphrey (1895–1958), and Charles Weidman (1901–).

In the German *Ausdruckstanz* the central figure was Rudolf Laban (1879–1958), who was more a theoretician and teacher than a choreographer. His researches into the physiological impulses to movement and rhythm crystallized in a formidable system of physical expression. His system of dance notation, known most widely as Labanotation, provided the first means for writing down and copyrighting choreographies. His most prolific disciples were Mary Wigman (1886–) and Kurt Jooss (1901–). Jooss became known for his dances containing strong elements of social commentary. When Wigman toured America in the 1930s, Americans became aware that they were not alone in their search for new forms of expressive dance. Wigman left behind one of her closest collaborators, Hanya Holm (1898–), who became another major figure on the American scene.

Across America schools opened, spawning small groups of dancers who performed on college campuses and on small stages in the cities. Each choreographer and company brought different materials, artistic points of view, and performing styles to the dance. Perhaps the single common denominator of the many facets of modern dance was the search for new and valid forms of artistic expression.

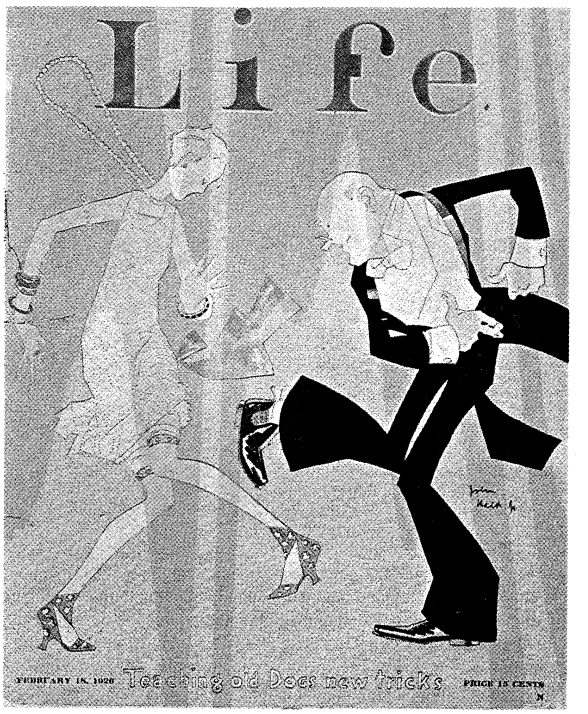
NEW RHYTHMS, NEW FORMS

The changes in the social climate that marked the new century had a notable impact on the ballrooms.

Latin-American and jazz dances. The younger generation in Europe eagerly took up the more vivacious, dynamic, and passionate social dances from the New World. The turning dances of the 19th century gave way to such walking dances as the two-step, the one-step or turkey trot, the fox-trot, and the quickstep, performed to the new jagged rhythms. These rhythms were heavily African in origin, whether from the Latin-American tangos and rumbas or from the American Negro jazz. It is impossible to say how watered down this music was from its original forms, but its impact was enormous in shaping the ragtime popular before 1918, the syncopated rhythms and mellower swing that followed it, the acrobatic jitterbug of the 1930s and 1940s, and the rock and roll of the next decades.

Dance contests and codes. After 1912, when the tango became the rave of the dancing world, even elegant hotels invited their clientele to their "tango teas." In fashionable restaurants professional couple dancers demonstrated the new styles. In 1892 New York saw one of the first cakewalk competitions, and in 1907 Nice advertised the first tango contest. After the first world dance competition in 1909, in Paris, it became an annual event, which in 1913 lasted for two weeks. But it was England that acted as arbiter of taste for the new movements in social dance. There the first dance clubs, like the Keen Dancers' Society (later the Boston Club), were founded in 1903. In 1904 the Imperial Society of Teachers of Dancing was established, and in 1910 the periodical *Dancing Times* made its bow. After World War I the English version of

The English style



Charleston from the cover of *Life*, designed by John Held, Jr., 1926.
Culver Pictures

the fox-trot was acknowledged as the essence of the internationally acclaimed "English style." Victor Silvester's *Modern Ballroom Dancing* (1928) became the handbook of the dancing world until it was succeeded by Alex Moore's *Ballroom Dancing* (1936). The English style involved strict definitions for the five standard dances—quickstep, waltz, fox-trot, tango, and blues—to which was added after 1945 the Latin American rumba, samba, calypso, and cha-cha-cha. What was left of the social barriers existing in 1900 between the exclusive and the popular dancing establishments was swept away.

Many observers were indignant about the changes taking place. Even so liberal a historian as Curt Sachs could not refrain from stating:

> Since the Brazilian *maxixe* of 1890 and the *cakewalk* of 1903 broke up the pattern of turns and glides that dominated the European round dances, our generation has adopted with disquieting rapidity a succession of Central American dances, in an effort to replace what has been lost to modern Europe: multiplicity, power, and expressiveness of movement to the point of grotesque distortion of the entire body. . . . All [of these dances are] compressed into even movement, all emphasizing strongly the erotic element, and all in that glittering rhythm of syncopated four-four measures classified as *ragtime*. (From Curt Sachs, *op. cit.*, pp. 444–5.)

Sachs went on to note the rapid rise and fall in popularity of individual dance and suggested an impermanence to the entire movement.

Impact on folk dancing. As social dancing spread with the advent of the radio and the phonograph, the regions where genuine folk dancing was practiced became fewer. It continued least corrupted by the new forms in those countries outside the mainstream of Western urbanization and industrialization. Spain maintained its vigorous tradition of flamenco dancing, and in Hungary the composers Béla Bartók (1881–1945) and Zoltán Kodály (1882–1967) collected the remnants of a wealth of folk song and dance folklore. Minority groups such as the Basques in Spain did likewise to maintain their identity against the overpowering influences of their neighbours. Folk dancing remained a vital reality in the Soviet Union, especially in those European and Asiatic provinces that had distinctive ethnic populations and were far removed from Moscow, Leningrad, and other centres with Western contacts. In the industrial nations of Europe and the Americas, special nationwide councils and



"El Jaleo," oil painting by John Singer Sargent (1856–1925). In the Isabella Stewart Gardner Museum, Boston.
By courtesy of the Isabella Stewart Gardner Museum, Boston

societies were founded to preserve the traditional folk dance that was under threat of extinction.

Experiments. Technological progress itself became the subject of dance and dancing. In the Soviet Union, there were experiments during the 1920s with dances created to express the whirl of the urban sidewalk and of rushing autos, the accuracy of machine work, and the grandeur of skyscrapers. In Germany, the painter Oskar Schlemmer (1888–1943) realized his vision of a dance of pure, geometrically shaped form in the *Triadisches Ballet* performed in Stuttgart in 1922. In 1926 a sound vision of the technological ages, *Ballet mécanique* (*Mechanical Ballet*), by the American composer George Antheil (1900–59), was scored for mechanical pianos, automobile horns, electric bells, and airplane propellers. It was written not for the live dancer but for an animated film.

THE DANCE SINCE 1945

Dance of all kinds emerged from World War II, more vital and more expansive than before.

Social dance. In the realm of social dancing, motion pictures and television helped to spread such rock and roll dances as the "twist" more rapidly and widely than dances had travelled before. A characteristic of this new generation of jazz-based dances was the lack of bodily contact between the participants, who vibrated their legs, gesticulated with their hands, swung their shoulders, and twitched their heads. Many observers attempted to draw social implications of all kinds from these dances, which began to spread also among the youth of the Communist nations of Eastern Europe and Asia. Among the more interesting interpretations was that of Frances Rust:

> . . . this type of dancing can be thought of as "progression" rather than "regression." Historically speaking, country-dancing of a communal or group nature gives way, with the break up of communities, to partnered-up ballroom dancing with a concentration on couples rather than groups. This, in turn, is now replaced amongst young people by partner-less dancing, which, although individualistic, seems none-the-less, to be rooted in a striving for community feeling and group solidarity (from *Dance in Society;* Routledge and Kegon Paul, 1969).

Dance in the theatre. On the postwar ballet scene there were no revolutionary developments such as those of Diaghilev earlier in the century. The classical ballet style reigned supreme throughout the West and in the Soviet Union. The popularity of ballet and the establishment of many apparently permanent companies made inevitable, however, wide variations in style and content. International tours were resumed on a large scale. There was also considerable interaction in terms of style and personnel between ballet and modern dance. The musical theatre, an inheritor of the techniques of these theatrical forms, became more artistically conscious and mature.

Companies presenting dances from India, Ceylon, Bali, and Thailand were no longer considered exotic on Western stages, and their influences contributed to both ballet and modern dance. Numerous ensembles sprang up, their repertoires based on traditional national dances adapted for the stage. Many were modelled on the Soviet Moiseyev folk-dance company, which attracted large audiences during its frequent European and American tours. Similar companies existed in several eastern European countries, in Israel, and in some new African nations, as well as in Brazil, Mexico, and the Philippines.

Dance and the film. Both the cinema and television played important roles in making some of the best and most interesting dance performances accessible to millions of persons. Film served also as a means to preserve the work of major dancers and choreographers for future audiences. Aside from these distributive and preservative functions, however, the dance probably made more contributions to the film than the film made to the dance. In the 1930s and 1940s, Hollywood developed a kind of spectacular choreography featuring glittering chorines or stylish couples and soloists, a mode adapted by television. Movies centred on ballet and ballets commissioned for television had little critical approval except in Sweden, where Birgit Cullberg (1908–) choreographed some notable productions for television.

From the beginning of the 20th century the dance scene became extremely multifaceted and colourful. If some of its manifestations appeared contradictory, that could be regarded as proof of its vitality. No other century granted dance so prominent a role among its social activities. Indications of this prominence included a vast increase in dance research and writing, the opening of colleges and universities in America to special dance faculties, and establishment in the Soviet Union of institutes for the study of choreography. And dance notation promised great advances in recording specific choreographies and as a basic linguistic tool in dance education.

BIBLIOGRAPHY. CURT SACHS, *Eine Weltgeschichte des Tanzes* (1933; Eng. trans., *A World History of the Dance*, 1937, reprinted 1963), the most comprehensive, systematic, and factual history of dance in all its epochs and forms, with special emphasis on its earliest beginnings and close attention to dance accompaniment; W.F. RAFFE, *Dictionary of the Dance* (1965), detailed descriptions of the particular dances, their background, and history; ANATOLE CHUJOY and P.W. MANCHESTER (eds.), *The Dance Encyclopedia*, rev. ed. (1967), the most comprehensive ballet dictionary available, including entries on specific productions, artistic biographies, and histories of ballet in various countries; LINCOLN KIRSTEIN, *Dance: A Short History of Classic Theatrical Dancing* (1942, reprinted 1969), a very thorough book on the pre-balletic forms of dance as well as classic theatrical dance; WALTER SORELL, *The Dance Through the Ages* (1967), a general, readable survey of the worldwide dance scene from prehistoric times through today, with superb pictures of ancient and modern dance; A.H. FRANKS, *Social Dance: A Short History* (1963), the first attempt at relating the origins and developments of the most important social dance forms to their social environment; FRANCES RUST, *Dance in Society* (1969), a study giving even more documentary evidence of the social dances and their relationships to the changing structures of society, with emphasis on the English scene and the teen-age explosion in dance during the 1960s.

(H.Ko.)

Dance and Theatre, East Asian

From ancient times the performing arts have played a vital role in the civilizations of East Asia—in China, Korea, and Japan. Their origin in Japan is explained by a myth that relates dance and theatre to the gods who created earth and life. Amaterasu, the Great Heavenly Shining Goddess, retired to a cave in anger, thus depriving the world of light. In order to lure her out, another goddess mounted an overturned tub, bared her body, and danced vigorously as the audience of gods sang and beat time. Intrigued by the laughter and shouting, Amaterasu came out and joined them, thus ending her self-imposed exile. The gods, having discovered the pleasure of performing and watching dance, passed their accomplishment on to man.

In early times the performances of plays and of dances were closely tied to religious beliefs and customs. In China, records from around 1000 BC describe magnifi-

Origin of dance and theatre in religious customs

cently costumed male and female shamans who sang and danced to musical accompaniment, drawing the heavenly spirits down to earth through their performance. Impersonation of other characters through makeup and costume was occurring at least by the 6th century BC. Many masked dances in Korea have a religious function. Performances invoking Buddha's protection are especially popular and numerous in Japan and Korea. Throughout East Asia the descendants of magico-religious performances can be seen in a variety of guises. Whether designed to pray for longevity or for a rich harvest or to ward off disease and evil, the rituals of impersonation of supernatural beings through masks and costumes and the repetition of rhythmic music and patterns of movement perform the function of linking man to the spiritual world beyond. Hence, from the earliest times in East Asia, dance, music, and dramatic mimesis have been naturally fused by their religious function.

In East Asia the easy intermingling of dance and theatre, with music as a necessary and inseparable accompanying art, also derives from aesthetic and philosophic principles. In the West, by contrast, concert music, spoken drama, and ballet have evolved as separate performing arts. Confucian philosophy holds that a harmonious condition in society can be produced by the proper actions of man, including the playing of music and the performance of dances that are appropriate and conducive to moderation. Throughout China's history, poems were written to be sung; songs were danced. Dancing, while it might be pure dance without meaning, could just as easily be used to enact a story in the theatre. Zeami Motokiyo (died 1443), the most influential performer and theoretician of Nō drama in Japan, described his art as a totality, encompassing mimesis, dance, dialogue, narration, music, staging, and the reactions of the audience as well. Without arbitrary divisions separating the arts, there has developed in East Asia exceptionally complex artistic forms that produce on their audience an impact of extraordinary richness and subtlety.

Dance may be dramatic or nondramatic; in all traditional theatre forms, some elements of dance will be found. Puppets, masks, highly stylized makeup, and costuming are common adjuncts of both dance and theatre. Dialogue drama (without music) is rare but does exist.

Current dance and theatre forms

The major dance and theatre forms performed today in East Asia can be described as follows: unmasked dances (folk and art dances in each country), masked dances (Korean masked dances and *bugaku* and folk dances in Japan), masked dance theatre (Nō in Japan and *sandae* in Korea), danced processionals (*gyōdō* in Japan), dance opera (Peking and other forms of Chinese opera), puppet theatre (*khoktugaksi* in Korea and *bunraku* in Japan), shadow theatre (in China only), dialogue plays with traditional music and dance (Kabuki in Japan), dialogue plays with dance (*kyōgen* in Japan), and straight dialogue plays introduced into China, Korea, and Japan from the West in the 19th century.

Characteristics of East Asian dance and theatre, and their social milieu

COMMON TRADITIONS

China, Korea, and Japan have been historically close for centuries, thus accounting for many common artistic traditions. From pre-Christian times, until the 8th and 9th century AD, the great trade routes crossed from the Middle East through Central Asia into China. Hinduism, Buddhism, some knowledge of ancient Greek, and much knowledge of Indian arts entered into China, and thence in time into Korea and Japan. Perhaps before Christ, the Central Asian art of manipulating hand puppets was carried to China. For more than 700 years, until 668, in the kingdom of Koguryŏ, embracing northern Korea and Manchuria, court music and dances from Central Asia, from Han China, from Manchuria, and from Korea, called *chisŏ* and *kajisŏ*, were performed. Many of the dances were masked; all were stately as befit serious court art. They were taken to the Japanese court in Nara around the 7th century. Called *bugaku* in Japan, they have been preserved for 12 centuries and can still be seen performed at the Imperial Palace in Tokyo, though they have long since died out in China and Korea. In Koguryŏ's neighbouring kingdom of Paekche, a form of Buddhist masked dance play was performed at court and, in the 7th century, it too was taken to the Japanese court at Nara by a Korean performer, Mimaji, who had learned the dances while staying at the southern Chinese court of Wu. Called *kiak* in Korea and *gigaku* in Japan, the Aryan features of some of its masks clearly indicate Indian (or Central Asian) influence. Such complicated genealogies are common in East Asian performing arts.

Very likely by the 7th century, gypsy-like puppeteers, who originally had been nomads from Central Asia and had taken up abode in northern Korea, migrated to Japan. The men continued to be both herdsmen and puppet manipulators in their new homeland, while the women performed dances and sang as popular entertainers. (There may have been a native puppet tradition in Japan as well.) In time the art of puppet manipulation joined with that of epic storytelling to produce the famous *bunraku* puppet theatre. Musical accompaniment for *bunraku* and for other popular plays in Japan, such as Kabuki, was provided primarily by the samisen, a three-stringed lute, borrowed from China by way of Okinawa. The lion dance, originally from China, is performed in a score of versions in Korea and Japan as well as in China, India, Ceylon, and Bali. Certain myths are dramatized in common as well. The story of the angel or nymph who flies down to earth and arouses the love of a mortal man is known in many parts of the world. It is dramatized in Southeast Asia (especially in Burma and Thailand, as the play *Manora*), in Chinese opera, and in both Nō and Kabuki theatre in Japan (as the "Feather Robe"). The legend of the one-horned wizard who traps the dragon gods of rain and causes a searing drought originated in India and was later transmitted to Japan by the Chinese, where it is dramatized in Nō ("The One-Horned Wizard") and in Kabuki ("Saint Narukami").

Western influences

The direction of artistic exchange was reversed in the last century. As part of Japanese national policy following the Meiji Restoration (1868), artists studied Western performing arts. Chinese and Korean actors, dancers, and playwrights studying in Japan took back to their countries Western theory and practice in ballet, modern dance, and theatre. In time, Western theatre and dance directly influenced Chinese and Korean performing arts. Perhaps most influential in all three countries was the Western dramatic theory of realism. It diametrically opposed the traditional intermingling of music and dance with drama, and it eschewed the stylization and symbolism that lay at the heart of East Asian performing arts for more than 2,000 years. A conflict between traditional and Western performing arts came into being that continues to the present.

As has been noted, dance and theatre are accompanied by music in all except the most unusual cases. Music may be instrumental or vocal. The music is especially composed for each *bunraku* puppet play in Japan and for most dance plays and court dances. Fixed melodies accompany most folk performances. In Chinese opera and in Japanese Kabuki, melodies appropriate to scene, action, character, or mood being portrayed are selected from a standard musical repertory of several hundred tunes. The knowledgeable spectator easily identifies scenes by the music that accompanies them (a similar system is found in Southeast Asian theatre). The close linking of music with dance and theatre can be seen in the Korean drum dance, in which the dancer also is a musician who plays the drum, and in a number of Japanese Kabuki and puppet plays that show characters expressing their deepest feelings by playing a musical instrument. Equally important, the performer demonstrates to the audience his skill in yet another refined accomplishment.

The performing arts of India are closely linked to sculpture and painting by the unusual phenomenon that bodily positions in all these arts are regulated by similar, indeed almost identical, codes. The code of hand gestures, for example, for the dancer and the actor set forth in the *Nāṭya-śāstra* ("Principles of Dramatic Art"), a 6th-cen-

tury-BC treatise on dramaturgy, is identical with that for Buddhist temple sculpture, or painting. Although these hand positions (*mudrā*) from India are also seen in Chinese, Korean, and Japanese statues of the Buddha, they have never been adopted by performing artists (as, by way of contrast, they were by dancers and actors in Cambodia, Thailand, Burma, Java, and Bali). In three notable instances, however, the performing arts in China and in Japan can be seen to be closely related to the visual arts. During the Sung dynasty (960–1279) in China, Northern and Southern schools of painting evolved that were totally different in style; the former used bold outline and brilliantly contrasting colors of deep green, blue, and gold, while the latter emphasized delicate, monochrome ink painting of misty landscapes. Northern and Southern schools of opera at the time reflected the same contrasting characteristics: the former dynamic, vigorous, and filled with action, the latter emphasizing wistful emotions and soft, gentle singing. Zen Buddhism was a common source of inspiration in the 15th and 16th century in Japan for Nō dance drama, for the tea ceremony, for ink painting, and for the art of rock-and-sand gardens. Spareness of form, discipline, and suggestion rather than explicit statement are Zen attributes found in these and other arts cultivated by the military ruling class (samurai) of the time. In 18th-century Japan, a lively and faddish urban culture produced both Ukiyo-e woodblock prints and Kabuki. In fact, Ukiyo-e artists, such as Sharaku, established their fame by portraying famous Kabuki actors as their subjects. Eroticism, verve, brilliant colouring, and an intense interest in the passing moment characterize equally both Kabuki theatre and Ukiyo-e visual art.

Aesthetic parallels of visual arts and performing arts

Although dances were often performed to sung poems and plays were either written in verse form or contained references to classic poems, the performing arts traditionally are seen as distinct from literature in East Asia. A century and a half passed in Kabuki before the first complete play script was preserved, and in China, where a tradition of written literature goes back to 1400 BC, no play text was considered worth committing to paper until late Sung, around the 13th century. With few exceptions, playwrights have rarely been accorded the same status as writers of poetry, novels, or criticism. As a result, the performing arts in East Asia succeeded by and large in escaping the stultifying grip that literature came to hold on Indian Sanskrit drama and some would say, still, at least in part, holds on drama in the West.

Artistic borrowings among East Asian countries

The general outlines of artistic borrowings among East Asian countries can be traced from historical records. But borrowing tells only half of the story. No matter how strong the initial outside influence, in time, assimilation of the foreign art took place. Older native performing traditions reasserted themselves, and new creativity altered the borrowed elements. This can be seen even in *bugaku* dances in Japan; although they are believed to preserve ancient Chinese and Korean forms to a very remarkable extent, native Japanese qualities are also present. Local styles predominate even more in the popular arts. Japanese *bunraku* puppet plays and Kabuki theatre show almost no observable signs of foreign influence. In spite of certain general cultural similarities, then, the dance and theatre of China, Korea, or Japan exhibit definite local characteristics not shared by the arts of their neighbouring countries.

In China singing became highly developed, and the most important theatre performances are built around song (hence the term Chinese opera). Dance as a separate art has a weak tradition there, and, at least in the 20th century, is tied very closely to the theatre. The shadow theatre, known from Morocco through Egypt and Greece, in India, Indonesia, Malaysia, Thailand, and Cambodia, is found in East Asia only in China. In Korea, dance predominates over drama. There are scores of Korean court and folk dances and danced plays, but no sophisticated dramatic forms evolved until this century. Masked dances especially are characteristic of Korea. In Japan, from an early tradition of imported pure dance and from folk dance, complex theatrical forms evolved that include dance drama, epic narrative performed as a puppet play, and dialogue dramas either accompanied or unaccompanied by music.

Aesthetic principles

The aesthetic principles that govern dance and theatre in East Asia are radically different from those of the West. The dancer in the West attempts to free himself from the pull of the earth; he tries to leap and soar in the air. The dancer in China, Korea, or Japan stands firmly on the dance floor, often scarcely raising his feet in the air, and moves in relatively slow and often geometric patterns. Arms and hand movements are important and varied while in Western dance the hands are little used. Whether movement is dance or not, it is always stylized. Speech is stylized as well, whether it is dialogue or narration, chanted or sung. The intent may be to portray archetypes, human or mythological, especially in shadow and puppet theatre and in masked dances and plays. There is great emphasis on form, both for its ritual value and because audiences are trained to recognize the beauty implicit in form. There may be a purposeful contradiction between artistic ends and means: children were moved as human puppets in China, and adults acted before lighted screens to make a living shadow play; in Japan, puppets execute realistic daily actions, while live Kabuki actors refrain from duplicating daily life. The East Asian audience is prepared to respond in quick succession to a sequence of different stimuli—physical characterization, human speech, song, narrative commentary, visual composition, formal movement patterns—over long periods of time; for eight hours in Nō or up to twelve hours in Kabuki. This differs from the West, where the spectator expects to be exposed to a clearly focussed theatrical image for only two or three hours. The East Asian experience is more diverse, more extended, more conventional than the Western experience in the theatre.

A further important characteristic of dance and theatre in China, Korea, and Japan is that performing arts developed very largely within an oral tradition. By and large the performers themselves created the forms; only gradually did specialists in choreography, musical composition, or playwrighting take their places in performing groups. Even after forms reached maturity, traditions of dance, acting, and music were passed on orally to the next generation. A major exception is that play scripts in China and Japan came to be written in full.

SOCIAL CONDITIONS

Social position of the performer

It is notable that although some dance and theatre forms were highly regarded in China, Korea, and Japan, performers were usually looked down upon. Wandering performers especially were despised in the agrarian societies, where attachment to the land was valued and Confucian teaching, strong throughout East Asia, stressed veneration of one's parents, which included tending their graves and making offerings for their welfare in the spirit world. The place of drama or of dance in these societies depended in part upon their audiences, whether they were court nobles, villagers, or town merchants.

Chinese emperors, Korean kings, and Japanese emperors and military rulers all supported performers at their courts. During the T'ang dynasty, the 8th-century Chinese emperor Ming Huang established schools in the palace city of Ch'ang-an for music, dancing, and acting. The latter school was set up in a pear garden; ever since, actors in China have been called "children of the pear garden" (*li yüan tzu ti*). Over a thousand young people from all ranks of society drew government salaries while studying and performing at lavish state banquets and for official ceremonies. Acting or dancing might be a permanent job (at least until old age made one less attractive) at the Chinese court, but, in Korea, performers at the court held other positions in the government and were mobilized from around the country only for rehearsal and performance. In Japan, dancers and musicians have been attached to the imperial household from the 7th century until the present time. First *gigaku* and then *bugaku* dances were official performing arts, while shrine dances (*kagura*) were also partly under imperial patronage. The military rulers of Japan (shogun) incorporated into their retinues Nō actors and musicians beginning in the 15th

century, and, in time, provincial lords also began to follow this practice.

Court patronage

Court support resulted in high artistic levels in all countries. Performers were relieved of financial problems and could devote themselves, often full time through their entire lives, to their art. Audiences were educated and for the most part discerning. The importance attached to official performances undoubtedly spurred the artists to extend themselves to their utmost. In time, however, such forms as Japanese Nō and *bugaku* and Chinese *k'un-ch'ü* opera became so rarefied that they could be appreciated only by a small elite group.

At the Chinese and Korean courts young female dancers were part of the ruler's personal retinue (often his concubines); they were not allowed to mix with men of the court, so that some courts' arts were performed solely by men and others solely by women. This custom and the consequent artistic practice of male and female impersonation is also found in court theatre of Cambodia and Thailand. In Japan, women seldom performed at court and the major dance and theatre forms have been the province of male performers. Since it was unusual for rulers or courtiers themselves to take part in performance (they often did in Java, Bali, and Thailand), the court artist was usually a middle-level civil servant.

Folk and popular theatre and dance

Folk performers, on the other hand, are local villagers who, like the *sandae* masked dancers of Korea or the girls who perform festive *ayakomai* dances in Japan, are amateurs who do not live by their art. The midsummer Bon dance for spirits of the dead or early spring rice-planting dances in many areas of Japan or various auspicious dances held at the New Year in Korea and China were performed only once a year, and hence a high level of artistry was not usually achieved. Because many folk performances were held as part of religiously sanctioned rituals (Korean mask plays ensuring harvest, dances and dance plays of many varieties in Japan dedicated to local Shintō deities), performers achieved considerable status in the local community by their participation in these essential communal rites.

Performers of popular dance and theatre in East Asia live as do commercial artists everywhere—by their ability to draw audiences who are willing to pay money for a seat at a show. The shadow and puppet performers of China, Peking opera actors and musicians, and Kabuki and *bunraku* puppet performers in Japan are popular artists. Neither a part of village culture nor patronized by the court, they have always been held suspect by their rulers. Kabuki, in particular, was faced with repressive government action throughout its history. Popular theatre grew in importance in China and in Japan concurrently with the growth of large urban centres and a moneyed, mercantile economy, especially since the 16th century. (An important urban popular theatre, however, has not developed in Korea.) Troupes perform nightly through the year, when it is possible, and consequently, in popular theatre, large repertories of standard plays are created (some 350 in Kabuki and more than 200 in Chinese opera). Popular theatre forms in China and Japan are intensely theatrical, though they lack literary qualities which would recommend them to the intelligentsia. Indeed, Kabuki in Japan and Peking opera in China have had little official status until recent years in spite of their immense audience popularity and their obvious excellence as performing arts. Travelling troupes that perform shadow or puppet plays, do acrobatics and juggling, dance and sing, and perform versions of court or popular entertainments have long been a feature of Chinese and Korean village and provincial town life. Artistically the forms are related to folk performing arts; socially the performers are considered outcasts, wandering entertainers of no status who belong to the popular tradition of performing arts.

The development of dance and theatre in the East Asian nations

CHINA

Formative period. Singing and dancing were performed at the Chinese court as early as the Chou dynasty (*c.* 1122–221 BC). An anecdote describes a case of realistic acting in 402 BC, when the chief jester of the court impersonated mannerisms of a recently deceased prime minister so faithfully that the emperor was convinced the minister had been restored to life. Drama was not yet developed, but large-scale masques (a short allegorical performance with masked players) in which dancing maidens and young boys dressed as gods and as various animals were popular. Sword-swallowing, fire-eating, juggling, acrobatics, ropewalking, tumbling and similar stage tricks had come from the nomads of Central Asia by the 2nd century BC and were called the "hundred entertainments." During the Han dynasty (206 BC–AD 220) palace singers acted out warriors' stories, the forerunners of military plays in later Chinese opera, and by the time of the Three Kingdoms (AD 220–264) clay puppets were used to enact plays. These evolved into glove and stick puppets in later years.

T'ang period. The emperor Ming Huang showed interest in the performing arts, stimulating many advances in stage arts during the T'ang dynasty (618–907). More than a thousand pupils were enrolled in music, dance, and acting schools. Spectacular masked court dances and masked Buddhist dance processions that soon were learned by Korean and Japanese performers were part of court life. Three types of play are recorded as having been popular. "False Face" was about Prince Lan Ling, who covered his gentle face with a horrifying mask to frighten his enemies when he went into battle. Some suggest the colourful painted faces of warriors in today's Chinese opera derive from this play. "The Swinging Wife" was a farcical domestic play, in which a weaving and sobbing wife bitterly complained about her brutal husband, who then appeared and, singing and dancing, abused his wife even more. The embezzling rascal hero of "The Military Counselor" became a stock character in later plays. Thus, by T'ang times, three basic types of drama were known: military play, domestic play, and satire of officialdom.

Popular types of plays

Sung period. A form of play called the variety play (*tsa-ch'ü*) was created by writers and performers in North China during the Northern Sung dynasty (960–1126). None of the scripts has survived, but something of their nature can be deduced from the 280 titles which remain and from court records. A play consisted of three parts: a low-comedy prologue, the main play in one of two scenes (consisting of extended sequences of songs, dancing, and perhaps dialogue), and a musical epilogue. Two, three, or four variety plays would be included in a program along with a sampling from the "hundred entertainments." In the following Southern Sung dynasty (1126–1279), northern writers continued composing plays of this general type under the name professional scripts (*yüan-pen*). None of the 690 professional scripts of which the titles are known has survived. Concurrently a new form of drama, southern drama (*nan-hsi*), emerged in the area around Hangchow in southern China. Originally the creation of folk authors, it soon became an appealing and polished dramatic form. A southern drama tells a sustained story in colloquial language; flexible verses (*ch'ü*) were set to popular music, making both music and poetry accessible to the ordinary spectator. Professional playwrights belonging to Hangchow's book guilds (*shu-hui*) wrote large numbers of southern dramas for local troupes. Of these, 113 titles and three play texts remain, preserved in an imperial collection of the 15th century. "Chang Hsieh, the Doctor of Letters" (probably the oldest of the three) dramatizes the story of a young student who aspires to success, earns a degree and position, but callously turns his back on the girl who faithfully loves him.

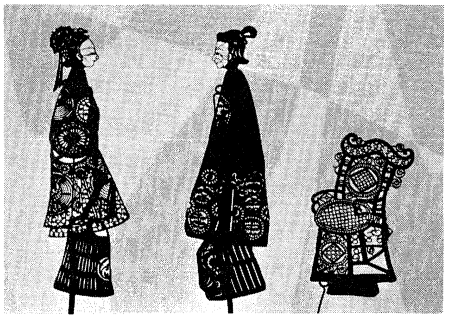
Professional theatre districts became established during the Sung dynasty. Major cities contained several districts (17 in Hangchow), with as many as 50 playhouses in a district. Plays performed by puppets and mechanical dolls were extremely popular.

Shadow theatre

A legend attributes the origin of shadow theatre in China to an incident said to have occurred around 100 BC: a priest, claiming to have brought to life the Emper-

or's deceased wife, cast a woman's shadow on a white screen with a lamp. Others suggest the shadow play dates only from the Sung period. In any case it was widely performed in Sung times in the theatre districts. Puppets were made of translucent leather and coloured with transparent dye so they cast (like some Indian puppets) coloured shadows on the screen. In this respect they were unlike Javanese shadow puppets, which, though brilliantly coloured, are opaque and cast a colourless shadow. Shadow plays are still occasionally seen in China. Singers, dancers, actors, acrobats, and other performers were all employed at the professional theatres of the districts. Troupes were as small as possible for economic reasons, containing as few as five or six performers. They would tour the countryside if they had no work in the large cities, thus spreading throughout the vast region of China urban styles of performing arts.

By courtesy of the Government Information Office, Republic of China



Chinese shadow play.

Yüan period. Scholars turned to writing drama in the Yüan period (1279–1368) when they were removed from their positions in the government by China's new Mongol rulers, descendants of Genghis Khan. They developed the earlier northern style of *tsa-ch'ü* into a four-act dramatic form, in which songs (in the same mode in one act) alternated with dialogue. Singing was restricted to a single character in each play. Melodies were those of the Peking region. The beauty of poetic lyrics was highly valued, while plot incidents were of lesser importance. About 200 plays survive, from thousands of romances, religious plays, histories, domestic, bandit, and lawsuit plays that were composed. *Hsi hsiang chi* ("Romance of the Western Chamber"), by Wang Shih-fu, is a 13th-century adaptation of an epic romance of the 12th century. The student Chang and his beautiful sweetheart Ying Ying are models of the tender and melancholy young lovers who figure prominently in Chinese drama. Loyalty is the theme of the history play "The Orphan of Chao," written in the second half of the 13th century. In it the hero sacrifices his son to save the life of young Chao so that Chao can later avenge the death of his family (a situation developed into a major dramatic type in 18th-century popular Japanese drama). "The Chalk Circle," demonstrating the cleverness of a famous judge, Pao, is well known in the West, having been translated into English and adapted (1948) by the German playwright Bertolt Brecht in *The Caucasian Chalk Circle*. The class of bandit dramas are mostly based on the novel "The Water Margin" and its 108 bandit heroes, who live by their wits doing constant battle against corrupt and avaricious officials. The life of the common man is portrayed with considerable reality in Yüan drama, though within a highly formalized artistic frame. The lasting worth of Yüan plays is attested to by the fact that they have been adapted constantly to new musical styles over the years so that Yüan masterpieces make up a large part of the traditional opera repertory.

Ming period. Plays of the Yüan period were widely popular with the people. When under the native Chinese Ming rulers (1368–1644) Mongol influence was eradicated, drama was, for a time, forbidden. Revived in the south, it increasingly became a literary form for a scholarly elite. The best known Ming play is "Lute Song," written in 42 scenes, by the scholar Kao Ming in the 14th century. Its heroine, Ts'ai Yung, sets a perfect example of Confucian filial piety and marital fidelity, caring for her husband's parents until their tragic death and then playing the lute to eke out a living as she patiently searches for her husband.

In the early 16th century, a musician, Wei Liang-fu, of Soochow, devoted ten years to creating a new style of music called *k'un ch'ü*, based on southern folk and popular melodies. At first it was used in short plays. Liang Ch'en-yü, poet of the 16th century, adapted it to full-length opera in time, and it quickly spread to all parts of China, where it held the stage until the advent of Peking opera, two centuries later. Important *k'un ch'ü* dramatists were T'ang Hsien-tsu (died 1616), famed for the delicate sensitivity of his poetry, Shen Ching (died 1610), who excelled in versification, and the creator of effective theatrical pieces, Li Yü (born 1611). A large-scale performance of *k'un ch'ü* for the emperor Ch'ien-lung in 1784 marked its high point in Chinese culture. *K'un ch'ü* had begun as a genuinely popular opera form; it was welcomed by audiences in Peking in the 1600s, but within decades it had become a theatre of the literatti, its poetic forms too esoteric and its music too refined for the common audience. In 1853 Soochow was captured by the Taiping rebels, and thereafter *k'un ch'ü* was without a strong base of support and declined rapidly.

Origin of Peking opera

Ch'ing and Manchu periods. Peking opera, or *ching hsi*, came into being over a period of several decades at the end of the 18th century, during the Ch'ing dynasty (1644–1911). In the wake of the Taiping Rebellion, *k'un ch'ü* troupes resident in Peking returned to their homes in the south. Their places in Peking's theatres were quickly taken by opera troupes from the surrounding provinces, especially Anhwei, Hupeh, Kansu, and Shansi. Anhwei opera had been performed on the occasion of the emperor Ch'ien-lung's birthday in 1790. Peking opera was born of an amalgamation of elements from several sources: rhythmic beating of clappers to mark time for movements (from Shansi and Kansu), singing in the two modes of *hsi p'i* and *erh huang* (from Anhwei), and increased use of acrobatics in fighting scenes. Undoubtedly, court support for Peking opera from Tz'u-hsi (1835–1908), the Empress Dowager, contributed to its rise, but it was also very widely patronized by local audiences. It became the custom to rehearse in public teahouses, and in time these became regular performances providing troupes with much of their financial support.

Differences between *ching hsi* and *k'un ch'ü* drama

Essentially, *ching hsi* was a continuation of northern-style drama, while *k'un ch'ü* marked the culmination of southern-style drama. Musically they are very different: the former uses loud clappers and cymbals for scenes of action and the penetrating sound of fiddles accompanies singing; in the latter, the flute is the major instrument, and strings and clanging cymbals are absent. A limited number of melodies are repeated many times in Peking opera (set to different lyrics), while in *k'un ch'ü* the melodic range is much wider. Peking opera lyrics are in colloquial language (they are often criticized as lacking in literary merit). Overall, the newer opera form is highly theatrical and vigorous, while the older form is restrained, gentle, and elegant. Some Peking operas are Yüan plays or *k'un ch'ü* operas adapted to the new northern musical system. Many plays first staged as Peking opera are dramatizations of the war novel *Romance of the Three Kingdoms*, written in the 14th century by Lo Kuan-chung. Mei Lan-fang, the most famous performer of *ching hsi* female roles in the 20th century, introduced a number of these highly active military plays into the repertory. *K'un ch'ü* dramas told a long and involved story in great detail, often in 40 or 50 consecutive scenes. It became the custom in Peking opera to perform a bill of a number of acts or scenes from several plays, like a Western concert program.

Concurrent with the national forms of drama mentioned before, local opera is found in every area of China (the different forms have been estimated at 300). They are performed according to local musical styles and in regional languages. General characteristics of most forms of Chinese opera are similar, however. Action occurs on

a stage bare of scenery except for a backdrop and sidepieces. A table and several chairs indicate throne, wall, mountain, or other location. (More elaborate scenery is used in Canton and Shanghai, influenced by Western drama and motion pictures.) Actors enter through a door right and exit through a door left. Costumes, headgear, and makeup identify standard character types. Actors play a single role type as a rule: male (*sheng*), female (*tan*), painted-face warrior (*ching*), or clown (*ch'ou*). Each role type can be subdivided into several role subtypes. Actors undergo seven years of training as children, during which time their appropriate role type is determined. Singing is essential for *sheng* and *tan* roles; minor actors and actors of clown roles must be skilled in acrobatics that enliven battle scenes. Singing is accompanied by a large number of conventionalized movements and gestures. For example, the long "flowing water" sleeves that are attached to the costumes of dignified characters can be manipulated in 107 movements. Pantomime is highly developed and several scenes have become famous for being enacted without dialogue: in "The White Snake" a boatman rows his lovely daughter across a swirling river; in "The Crossroads" two men duel in the dark; in "The Flower" a maiden threads an imaginary needle and sews. Symbolism is highly developed. Walking in a circle indicates a journey. Circling the stage while holding a horizontal whip suggests riding a horse. Riding in a carriage is represented by a stage assistant holding flags painted with a wheel design on either side of the actor. Four banners indicate an army. A black flag whisked across the stage means a storm, a light blue one a breeze or the ocean. Chinese opera is one of the most conventionalized forms of theatre in the world. It has been suggested that the poverty of troupes and the need to travel with few properties and little scenery led to the development of many of these conventions.

Eastfoto



A scene from *The White-Haired Girl*, a propaganda opera-ballet performed in the People's Republic of China.

Role of Confucian morality in drama

Confucian morality underlies traditional Chinese drama. Duty to parents and husband, loyalty to one's master and elder brother or sister, were virtues inculcated in play after play. Spiritualism and magical powers, derived from Taoism, are themes of some dramas, but by and large, Chinese drama is ethical rather than religious in direction. Plays were intended to uphold virtuous conduct and to point out the dire consequences of evil. The Western tragic view, which holds that man cannot understand or control the unseen forces of the universe, has no place in Chinese drama; the typical play concludes on a note of poetic justice with virtue rewarded and evil punished, thus showing the proper way of human conduct in a social world.

20th century. With the establishment of the Republic of China in 1911, court support for Peking opera by the Manchu dynasty ended. Troupes, however, continued to perform for private patrons and in public at teahouses and in theatres. Following the liberal ideals of the time, attempts were made to write in colloquial language (rather than in classical Chinese, as previously) and old plays considered undemocratic were dropped from the repertory. A school for Peking opera acting, modelled on Western pedagogical methods, was established in 1930, actresses being admitted for the first time in three centuries. The basic style of opera remained unchanged, however.

Influence of Western drama

Western drama was first introduced by Chinese students who had studied in Japan and there learned of Western plays. In 1907 a Chinese adaptation of *Uncle Tom's Cabin* was successfully staged in Shanghai by students, marking the beginning of a proliferation of amateur study groups devoted to reading and staging Western plays. Originally aimed only at a small group of Western-educated intelligentsia, spoken drama's appeal was broadened to the middle class by the China Traveling Dramatic Troupe, which toured many cities from its home in Shanghai. In 1936 it performed "Thunderstorm," a four-act tragedy by Ts'ao Yü. An extremely successful playwright in the Western style, by 1941 Ts'ao Yü wrote six important plays, including "Peking Man"; heavily influenced by O'Neill and Ibsen, he portrayed dissolute members of the old gentry class and new rising entrepreneur class.

Nationalism, the upheaval of World War II, and changes of government in China, Korea, and Japan between 1945 and 1949 are reflected in contemporary theatre and dance in East Asia. In China an estimated 60,000 performers were mobilized into some 2,500 propaganda troupes during the Sino-Japanese War beginning in 1937 under the direction of the well-known playwright T'ien Han. Hundreds of thousands of ordinary Chinese in the army were exposed to modern forms of drama for the first time, and, equally significant, artists discovered regional folk legends, songs, and dances, which they then incorporated into their work. The creation of new flexible forms blending local and Western theatrical elements that is characteristic of contemporary Chinese theatre owes much to this wartime experience. *Yang-ko* folk dances of Hopeh province were developed into a propaganda dance play by the Communist Eighth Route Army and performed in areas under their control during the war, culminating in the creation of "The White-Haired Girl," a play originally done in *yang-ko* style and subsequently adapted to opera and into ballet using Western orchestra. The heroine, an escaped concubine of a cruel landlord, symbolized all the people victimized by oppressive governments and social systems.

Dance and theatre in the People's Republic of China

In 1942 Mao Tse-tung stressed that art was a means of "uniting and educating the people." This philosophy has been carried out in the People's Republic of China since its establishment in 1949, with two chief results: dance and theatre are highly valued and strongly supported by the government; and content and form of dance and of drama are decided by appropriate government organs. Since 1949, folk dances of the many regions of China have been encouraged. Regional language opera has flourished. Amateur groups as well as professional are extremely numerous.

Under Communism there have been far-reaching reforms in the performing arts. Traditional opera was supported by placing performers on government salaries and by establishing opera schools in most provinces. But because traditional opera had been created by a feudal class, the new government strove to reshape it into an art form more accessible to a mass audience. Some reforms were technical (abolishing the stage assistant and using act curtains) and were designed to modernize staging. Also, antidemocratic, superstitious, erotic, or feudal sentiments were proscribed. Through 1953 most changes in the repertory consisted of rewriting offending passages or scenes. Beginning in 1964, operas with contemporary themes, with soldiers and workers as heroes costumed in ordinary dress, using realistic scenery and lighting, and acted in mixed real and conventional style were en-

couraged. Traditional music was largely retained. During the period of the Great Proletarian Cultural Revolution (1966–1969) most traditional operas ceased to be performed in favour of new works, realistic in style and content, and using, paradoxically, Western musical instrumentation.

The performing arts in Nationalist China

The Nationalist government has supported Peking opera on Taiwan since establishing the headquarters of the Republic of China on that island. Troupes of the Air Force and the Army are active, and a private school, Foo Hsing, trains children in Peking opera performance. Local opera, sung in the Taiwanese dialect, is extremely popular in commercial theatres, and approximately 200 itinerant Taiwanese troupes tour glove-puppet plays to towns and villages.

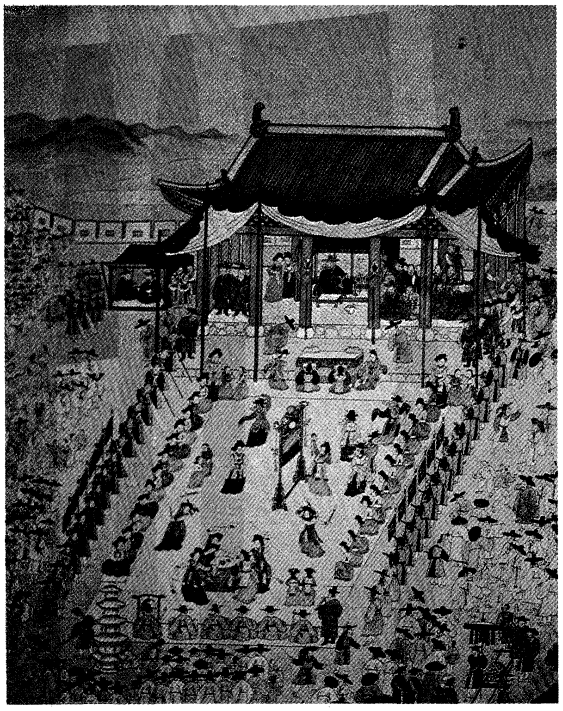
KOREA

In addition to folk dances, the main theatrical forms that developed in Korea are ritual court dances, masked dances, and puppet plays. Of these, masked dances and masked-dance plays have perhaps the oldest and richest traditions. Archaeological evidence suggests that masks were used at least by the 3rd century AD to impersonate animal spirits and thereby placate them. Various kinds of masks—demon masks, medicine masks, spirit masks—were worn by shamans as they danced to draw into themselves the spirit being addressed, in order to cure an illness or otherwise affect daily life. Magical properties continued to be associated with masks even after performances ceased to have religious or magical functions and became merely entertainment.

Three Kingdoms period. Lack of records makes it impossible to describe accurately dances and dance plays of Korea prior to the period of the Three Kingdoms (*c.* 57 BC–AD 668). Chinese, Japanese, and Korean accounts beginning in the 7th century give some indication of court arts in the Three Kingdoms of Koguryŏ, Paekche, and Silla. In Koguryŏ, encompassing what is now Manchuria and northern Korea, Central Asian music and dances were combined with local styles of music and dance. Twelve of 24 pieces in the repertory were mask dances. So highly regarded were the arts of Koguryŏ that they made up a separate Korean component of the Nine Departments of Musical Art and Dance at the T'ang court in China (25 musical and dance items were identified as Korean), and from the 7th century they were introduced into Japan, where they became the basis of *bugaku* (court masked dance). The strongly Buddhist state of Paekche in the southwest had been in contact with both China and Japan from early in the Christian Era. Typical of Paekche was a Buddhist masked-dance processional (*kiak*), originating in southern China and taken to Japan in 612 by a resident of Paekche, Mimaji. No Korean account of *kiak* survives, but Japanese accounts make clear that it was performed as a Buddhist ceremonial for evangelical purposes.

Great Silla period. The third kingdom, Silla, absorbed Koguryŏ and Paekche in the 7th century, and during the Great or Unified Silla period (668–935) the folk and court performing arts of all parts of Korea intermingled. Several major types of masked dance are mentioned in Silla records. The spirit of a noble youth who died to save his father's throne was memorialized in a masked sword dance (before this time, palace dancing girls had performed sword dances, but always unmasked). Masked dances called "The Five Displays" are mentioned in a Silla poetic composition of the 9th century. They included acrobatics, ball juggling, farcical pantomime (some of it unmasked), shamanistic masked dances, and the lion dance. The similarity of several to Japanese *bugaku* dances has been noted. Others believe "The Five Displays" derive from the "hundred entertainments" of China. Finally, an important dance play honouring Ozoyong, the son of the Dragon God of the Eastern Sea, dates from this period. Ozoyong showed such generosity toward the spirit of plagues that henceforth the spirit promised never to enter a household where a portrait of Ozoyong was hung. Originally derived from animistic beliefs, the dance was modified by Buddhism and was developed in the Yi dynasty (1392–1910) into a spectacular dance play, performed by a cast of five masked dancers, 16 unmasked dancing girls, and accompanied by an ensemble of 37 musicians.

The dance play honouring Ozoyong



Korean female drum dance, sword dance, and masked dance, detail of a watercolour on paper scroll depicting a reception for the governor of P'yongyang, Yi dynasty (1392–1910), by Kim Hong-do (18th century). In the National Museum, Seoul.
By courtesy of the National Museum of Korea, Seoul

Koryo period. The two major court festivals at which performances were held during the Koryŏ period (935–1392) were Buddha's birthday, or the Feast of Lanterns, in the second lunar month, and the midwinter ceremony honouring spirits of local gods. Dances and masked plays from Silla times were carefully preserved and performed on these occasions in a specially decorated and candle-lit ceremonial room. New masked plays memorializing loyal warriors who had died in battle were added from the 10th century. Buddha was offered gifts of wine and food, and performance was dedicated to maintaining a reign of peace and harmony. From the time of King Mun Jong (1047–83), T'ang style dances and sung dramas were performed on other occasions; modified by Korean forms, they became part of Korean court dance in centuries following. Folk dances and plays undoubtedly go back many centuries before this; in the Koryŏ period, professional troupes also became part of urban life. The practice of court performers holding civil-service jobs in the major cities and in provincial towns probably accounts for the fact that knowledge of court performing arts began to reach beyond the confines of the court during this time. Popular troupes began the process of secularizing religious masked dances (such as the *narye*, which formerly was performed to exorcise evil). They performed acrobatics and shows of skill and at least by the 12th century were staging satirical dialogue plays (masked and unmasked), which held officialdom up to ridicule. (The development of social satire is found in many Asian drama forms: the Vidusaka jester in Sanskrit drama, the god-clown-servants of Indonesian *wayang* shadow plays, and the servants of *kyōgen* comedies in Japan are major roles in these forms.)

Satirical theatre

Yi and modern periods. Buddhism was rejected as a state religion by the Yi dynasty (1392–1910), with the result that court entertainments were no longer scheduled according to Buddhist days of worship but at any time court entertainment was required. A Chinese envoy to the Yi court in 1488 described court performances that included: the *ozoyong* dragon-god dance play, children's

dancing, acrobats, ropewalking, and displays of animal puppets. Following invasions by the Japanese (1592) and by the Manchus (1636), court support declined. Former palace performers formed professional troupes, in the process adapting court forms to popular tastes. These performers included all the miscellaneous stage arts in their repertory and created from the various court dances and masked plays a type of folk masked play usually termed *sandae togam kŭk*. A prominent feature was the satirical treatment of depraved Buddhist monks and of grasping officials (naturally, favourite themes for a popular audience). Satirical plays were occasionally performed at court as well, but the banishment in 1504 of an actor for ridiculing the institution of kingship in a court play suggests satire was not welcomed.

In addition to professional groups, villagers in different areas of the country formed folk groups to perform their own local versions of the *sandae* masked play and dances. Today, the *sandae* masked play is performed by villagers in Yangju, in Pongsan in northern Korea, and in South

Song Kee-yep



The scene of *Aesadang-nori* from Yangju *sandae* mask-dance drama of Korea.

Kyŏngsang. Performers are males. Masks cover either the whole head or the face and are made from paper or gourds or, occasionally, are carved from wood. They are simply, even crudely, painted to represent the stock characters of the play: monks, shaman, noblemen, young dancing girl, and others. There may be 20 or 30 masks used; often they are burned and made anew each year to insure their ritual purity. Performance encompasses singing, dancing, pantomime, and dialogue. The stories enacted vary with the village, but common scenes include offerings to the gods, criticism of venial Buddhist priests, exposure of corruption by gentry and officials, flirtation, and a funeral service that brings absolution. Performances may be given as a rainmaking rite.

Puppet plays

The origin of puppet plays in Korea has not been determined; however, in the Koryŏ period puppet plays were widely performed and very popular among the people. Several types of puppet play developed in Korea. The folk puppet play "Khoktu Khaksi," named after the wife of the main character, is still performed in the summer months in southern Korea by farmers in troupes of six or seven players and musicians. Twelve or 15 puppets make a set (compared to more than a hundred in Indonesian or Japanese puppet theatre); they are simply made glove and stick figures that can be manipulated by a single puppeteer. One play, with variations, is performed. It consists of eight relatively independent scenes that satirize a figure of the gentry who is the major character. Scenes satirizing depraved monks and insulting the gentry, a domestic triangle, and Buddhist prayers for the dead appear to be adapted from masked plays.

Gu gug (literally "old plays") became popular around the middle of the 19th century. They were dramatic songs, danced to gestures and simple group movements. Troupes played throughout the countryside and in the new National Theatre, built in Seoul by the government in 1902. Until the 1930s, variety programs of *gu gug* and female court dances were popular entertainments at commercial theatres in the city. Sentimental melodramas, called "new school," or *shimpa*, plays (the same name as in Japan), were performed by a dozen troupes that formed and disbanded between 1908 and around 1930. The new school movement was begun by the novelist Inzig Yi. Other major figures had learned the style while studying in Japan. In 1931 the actor He-seng Hong and others organized the first drama and cinema exhibition in Korea; the following year its organizers formed the Society for Research in Dramatic Art, which studied and staged translations of modern European plays. Its members were primarily teachers and students of foreign languages and literature concerned with bringing Western *sin gug* ("new drama") into Korea. By 1940 about 100 amateur new drama groups had come into being.

Since World War II. In Korea, after 1940, all dramatic groups were obliged to belong to the Japanese-organized Dramatic Association of Korea. Many groups survived the war with Japan by touring small towns and villages. Performances lagged immediately after World War II because of unsettled conditions. A National Theatre was established in Seoul just before the Korean War began; national support included subsidies for performances. The destruction of the war has never been wholly repaired in the performing arts, for although the National Theatre was reinstituted in 1953, its operation has not grown as originally planned.

In addition to the National Theatre established by the South Korean government, a Drama Center devoted to training in modern drama was founded in Seoul in 1962 with private support. During the 1960s historical plays derived from "old plays" and girls' operetta, using traditional music and dance styles, were performed to large audiences in Seoul. Folk dances and puppet plays are declining in number in Korean villages, as a result of rapid urbanization and the spread of motion pictures. In Seoul, motion pictures and television had largely replaced dance plays in the theatres by 1970.

Like the Chinese Communists, the North Koreans have made an effort to encourage and support ideologically desirable dance and theatre forms, especially folk forms.

JAPAN

Among the most varied and technically complex theatre arts in Asia are those of Japan. While dance remained predominant over drama in Korea and singing is perhaps the most important single element of Chinese performance, in Japan music and dance gradually evolved into highly developed dramatic and theatrical forms, the most important of which are Nō dance drama, popular Kabuki theatre, and *bunraku* puppet drama.

Formative period. From prehistoric times, dances have served as an intermediary between man and the gods in Japan. *Kagura* dances dedicated to native deities and performed at the imperial court or in villages before local Shintō shrines are in essence a symbolic re-enactment of the propitiatory dance that lured the Sun Goddess from the cave in ancient myth. Though *kagura* dance has been influenced by later more sophisticated dance forms, it is still performed much as it was 1,500 years ago, to religious chants accompanied by drums, brass gongs, and flutes. At the same time, villagers had their rice-planting dances, performed either at New Year's as a prayer for good planting or during the planting season in early summer. These lively dances were later, in the 14th century, brought to the cities and performed as court entertainment and called *dengaku* ("field music").

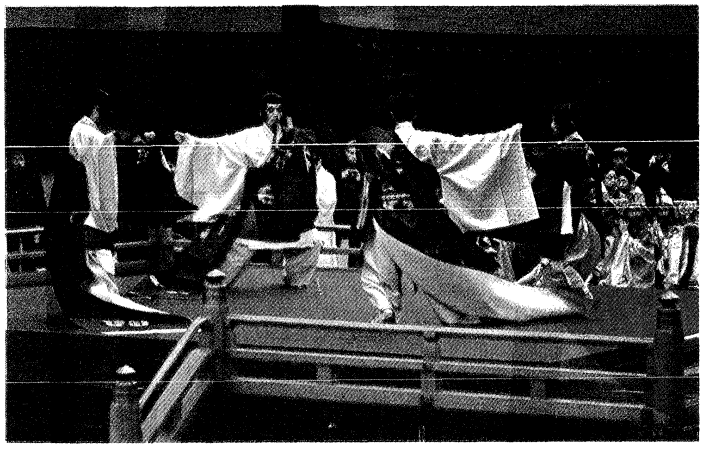
7th to 16th centuries. A massive influx into Japan of Chinese and Korean arts and culture occurred between the 6th and the 10th centuries. It has been noted that a Korean performer, Mimaji (Mimashi in Japanese),

The *gigaku* processional dance play

brought the Buddhist *gigaku* processional dance play to the Japanese court in 612. The current regent to the throne, Prince Shōtoku, was a devout Buddhist who gave Mimashi a place at the court so that he could establish an official school to train Japanese dancers and musicians in *gigaku*. Other Korean and Chinese performers from Paekche and Koguryŏ were invited in following years. *Gigaku* masks cover the entire head (as do Korean folk masks today). Carved of wood and painted with lacquer, the 223 masks that remain (most in the Shōsō-in, or imperial storehouse, in Nara) date back as early as the 7th century. They are superb examples of the art of mask making, strong featured and beautifully conceived. From a description of a 13th-century performance, *gigaku* apparently consisted of a succession of scenes enacted as characters passed by. Masks characterized an Aryan-featured dignitary called Baramon (or Brahman, indicating Indian origin), a fierce wrestler, a Buddhist monk, a princess of the state of Wu in China, a bully, a wistful old man, and others. Some scenes were serious, others were earthy slapstick.

Bugaku court dances introduced from Korea were also patronized by Prince Shōtoku. They supplanted *gigaku* as official court entertainment, and *gigaku* disappeared as a performing art by the 12th century. It was the custom to have performers of *bugaku* enter from dressing rooms to the right and the left of the raised platform stage: "right" dances, costumed in orange or red, were those from India, Central Asia, or China proper; "left" dances, costumed in blue-green, were those from Korea and Manchuria. *Bugaku* is performed by groups of four, six, or eight male dancers who move in deliberate, stately steps, repeating movements in the four cardinal direc-

ZEFA—G. Haasch



Bugaku, a court dance adapted to Japanese tastes from the dance and music of 8th-century China and Korea.

Performance of *bugaku* court dances

tions. Musical accompaniment is by drums, bells, flute, and *shō* (panpipe). A composition consists of three sections, introduction, development, and conclusion or "scattering" (*jo-ha-kyū*). Japanese performers and courtiers created new compositions within the old style in the 10th and 11th centuries. Still, *bugaku* represents a remarkable preservation of ancient Chinese, Indian, and Korean music and dance that has long since disappeared in their country of origin. *Bugaku* has been performed by musicians attached to the imperial court and to major Shintō shrines from the 7th century without break to the present day.

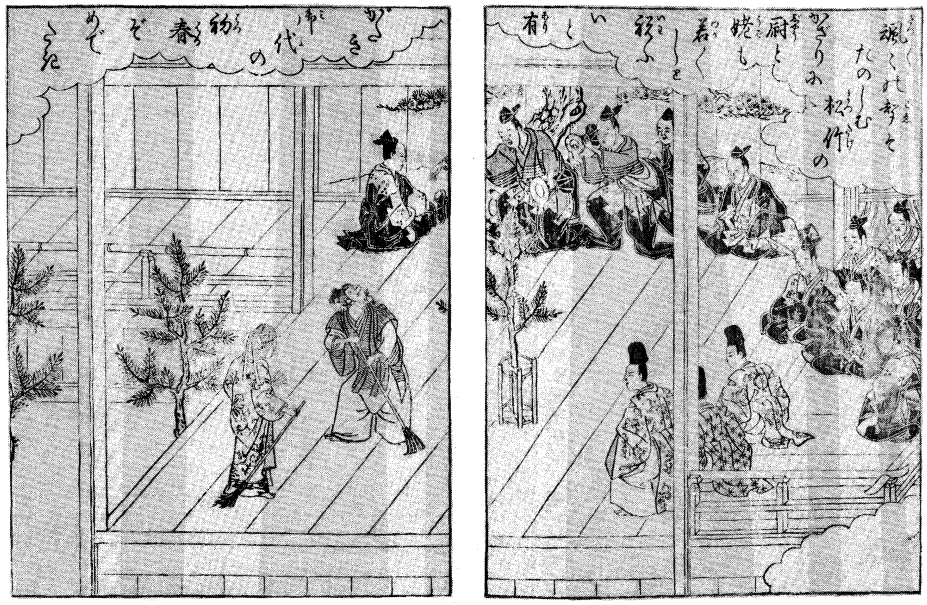
Juggling, acrobatics, ropewalking, buffoonery, and puppetry—the "hundred entertainments" of China and called *sangaku*, "variety arts," in Japan—became widely popular as well. During the Heian period (794–1185) professional troupes, ostensibly attached to temples and shrines to draw crowds for festival days, combined these lively stage arts, now called *sarugaku*, with dancing to drums from *dengaku*, and began to perform short plays consisting of alternate sections of dialogue, mimicry, singing, and dancing. Sometime in the 14th century a *sarugaku* actor from Nara named Kan-ami Kiyotsugu incorporated in his plays a chanted dance (*kuse-mai* or *kōwaka-mai*), for the first time creating the possibility of dramatic dance that could carry forward a story. This fusion of dance, drama, and song, which soon came to be known as *sarugaku* Nō, or simply Nō, marked a revolutionary advance in Japanese theatrical art. Kan-ami's son, Zeami Motokiyo, refined the style of performance, composed 50 or more of the finest Nō plays in the repertory, and wrote treatises on the art of acting and dramaturgy that have not been surpassed in his country.

When Zeami was 11, the military ruler of Japan, the shogun Yoshimitsu, saw him perform, became enamoured of the boy's beauty, and took him into his residence in Kyōto as a companion. For most of his life, Zeami benefitted from the patronage and the refined audiences that stemmed from this circumstance. The *sarugaku* troupe that Kan-ami and later, Zeami, headed was one of four in Nara; the others soon adopted the changes in performing style and the plays created by father and son. Through Zeami's lifetime, *dengaku* troupes continued to be potent rivals for patronage of the warrior class (indeed for a time, Zeami lost favour at court to the *dengaku* actor Zoami). The artistic superiority of Nō was such, however, that by Zeami's death *dengaku* had ceased to be important and gradually died out.

Nō theatre

The borrowings of Nō from other arts are many. The exquisite masks for which Nō theatre is famous have a quality of serenity, a neutrality of expression that places them in a rank perhaps unmatched in the world. Yet, historically there is no doubt that they are derived from earlier *bugaku* and *gigaku* masks and hence are related, if distantly, to the masks of Korea, China, and India. One evidence of the special development of Nō masks is that they are smaller than previous masks; they cover only the face proper. From *bugaku* music, Zeami took the three-part structure of the Nō drama. Each individual play, therefore, is divided into three sections. A normal Nō program consists of five plays, which are grouped into three dramatic units: the introduction, the development, and the conclusion. The first play, a "god" play, comprises the introduction; the development section includes three plays—the second drama or "warrior" play is the introduction of the development; the third, or "woman," play is the development of the development; the fourth, or "living person," play is the conclusion of the development; and the fifth, or "demon," play is the conclusion. Drums and flute were taken over from earlier musical forms, and Nō chanting is based on the style of Buddhist prayer chants. The songs' poetic meter of alternating phrases of seven and five syllables had come from China six centuries earlier and was the standard Japanese poetic form. On the other hand the Nō stage represents an advance on the simple square platform of *bugaku*. A sharply peaked roof over the stage is supported by four pillars—to help the performer orient himself as he looks out through tiny eye holes in the mask—and a long ramp, *hashigakari*, emphasizes the appearance and the departure of major characters.

Typical of a number of Nō plays that are dramatizations of Chinese history and legends is the 15th-century *Yōkihi*, by Komparu Zenchiku, based on an ancient narrative poem ("The Song of Everlasting Sorrow") by Po Chü-i. The original describes Emperor Ming Huang's love for Yang Kuei-fei (Yōkihi in Japanese), which led to the eventual destruction of his realm; the Nō play emphasizes the Buddhist sentiment of the evanescence of mortal life and the inevitability of pain and sadness. In the major scene of the play the ghost of Yang Kuei-fei returns to mourn her parting from the Emperor. Every Nō play contains Chinese poems, quoted verbatim or paraphrased so as to appeal to the educated spectator. It was a first principle of dramatic writing, said Zeami, to base a play on a well-known incident the central character of which the audience would find readily familiar. Zeami's plays emphasized the quality of restrained beauty (*yūgen*), a concept derived in part from Zen Buddhism. Later plays, especially those by Kanze Nobumitsu (died 1516), such as "The Maple Viewing" and "The Ataka Barrier," emphasize action and spectacle (*fūryū*).



Nō play on a Nō stage, woodcut by Suzuki Harunobu from "E-hon shogei nishiki," 1763. In the British Museum, London.
By courtesy of the trustees of the British Museum

The *kyōgen* farce comedy

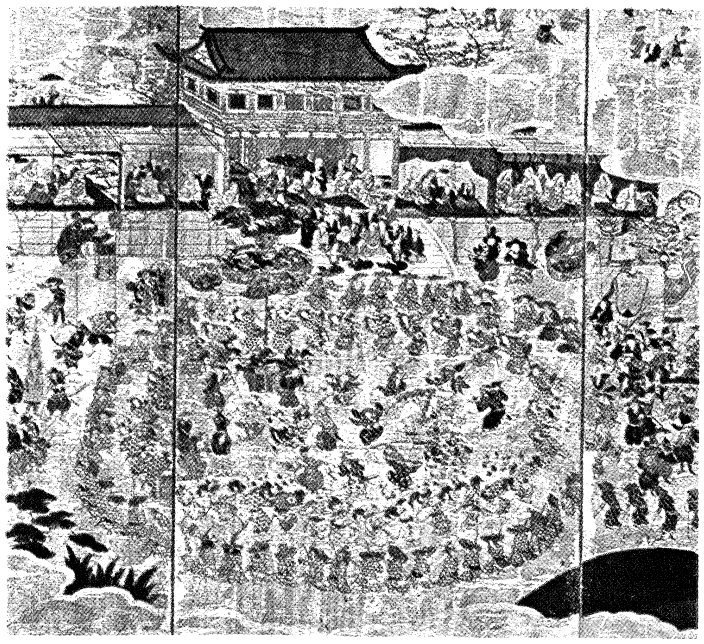
On the usual Nō program, each play was followed by a *kyōgen* farce comedy, performed not by the chief (*shite*) or supporting (*waki*) actors of Nō but by *kyōgen* actors who also acted the roles of villagers or fishermen in Nō plays. The antecedents of *kyōgen* cannot be described with certainty, but it is probable that *kyōgen*'s short sketches of master–servant quarrels, husband and wife arguments, animal fables, and scenes of rustic life derive from early *sangaku* entertainments. A few *kyōgen* plays are accompanied by the drums and flute of Nō. The ritual play *Ōkina*, performed as an auspicious prayer for longevity at the beginning of a Nō-*kyōgen* program, is in both repertories, and some suggest that the *kyōgen* version is the older. The style of *kyōgen* music (*komai*) is distinct from that of Nō music; it is derived directly from popular songs. *Kyōgen* plays with music are, however, a rarity. The usual play is a straight dialogue drama, making it perhaps the oldest developed form of nonmusical play in East Asia. Dialogue is composed in colloquial language of the 15th century, in short phrases suitable for comedy. Movement is highly stylized, again for comic effect. Masks may be worn for the roles of animals and demons, but most roles are played unmasked. *Kyōgen* texts do not seem to have been committed to writing until the 16th century, suggesting that actors traditionally adlibbed their parts. Today, *kyōgen* actors commit lines to memory.

Momoyama period. Nō and *kyōgen* were dance and theatre forms that had come to express the gravity and decorum of a rigidly formal samurai ruling class by the end of the Momoyama period (1568–1600). Artistically severe and highly disciplined, Nō was imbued with the sternly pessimistic philosophy of Buddhism. In content, Nō plays taught the folly of worldly power and position, that time destroys all living things. The heroes of play after play pray for Amida Buddha's divine intercession in order that they, tormented ghosts of dead warriors and court ladies, may break free of earthly attachments to achieve salvation. In contrast to this, commoners of the cities in the late 16th century began to perform their own radically different dances and plays that were up-to-date, lively, exciting, and at times morally licentious. They were intended to appeal to literate townsmen, well-to-do wives of merchants, workers, and the fops and wits and dandies of the burgeoning cities.

Edo period. During the Edo period (1603–1867) Nō was assiduously cultivated by warriors (samurai) as a refined accomplishment. Following the shogun's example, Nō troupes were patronized by provincial lords throughout the country. Actors received minor samurai status. *Sarugaku* had been a popular art in its beginning, an entertainment for townsmen and villagers at temple and shrine festivals. Even after Zeami had become a court favourite, Kan-ami continued to tour small temples in the countryside with his troupe. But, during the Edo period, commoners were forbidden by law to study Nō and were excluded from performances except on special "subscription" occasions, when any person, high or low in rank, could see Nō performed outdoors in a large enclosure. Nō became the exclusive theatre art of the warrior class, while *bugaku* continued as the chief performing art of the imperial court.

The earliest important urban entertainments of the commoner in Japan were secularized forms of Buddhist dance plays (*ennen*) and folk dances (*yayako odori* and *kaka odori*) that came to be called *fūryū* ("drifting on the wind") dances. They were enormously popular. In 1604 in the capital city of Kyōto, five teams of 100 dancers each participated in a mass *fūryū* dance commemorating the death of the ruler. The frenzied *fūryū* dancing of

Harumi Konishi



Fūryū dance, detail of "Fukoku sairei zu," a painted *byōbu* screen, 17th century. In the Fukoku Shrine, Kyōto.

100 men and 150 women earlier in 1567 had broken the boards of the dance platform.

Origin of Kabuki

In 1603 several kinds of urban dances were arranged into a new dance, called Kabuki, by a young woman named Okuni. Other troupes of female performers, actually prostitute-performers, adopted the sensuous and popular style of Okuni's Kabuki dance. A scroll of the period shows Okuni as a young, fashionably dressed samurai, indolently leaning on a sword, dallying with a teahouse girl. Around her neck hangs a Christian crucifix, not as a religious article but as an exotic decoration recently introduced by "southern barbarian" Portuguese merchants. Other pictures of the time show young women playing the three-stringed samisen as they recline sensuously on tiger skins, dancing girls circling about them. Audiences of monks, warriors, young lovers, and townsmen gaze raptly at this appealing and even bizarre sight. The original sensuous appeal of women's Kabuki continued long after women were banned from the stage in 1629. From around Okuni's time until 1652 troupes of young boys, actually professional sodomites, performed graceful Kabuki dances to samisen music to attract customers. The appearance of professional women and boy performers in Kabuki was a phenomenon of urban society. In the past, men alone had performed *gigaku*, *bugaku*, Nō, and *kyōgen*. Court nobles and samurai lords had always been able to take mistresses in any number they wished; now the commoner and townsman could, with his new wealth, purchase the favour of a newly risen class of women whose role was to cater to their desires. In the same way, taking young boys as sexual companions, which was common practice among the Buddhist clergy and the warrior class (Zeami, for example, was the shogun's catamite), becomes a feature of Japanese urban life now as well.

Only in 1653, when the authorities required Kabuki to be performed by adult males, did Kabuki begin to develop as a serious art. During the Genroku era (1688–*c.* 1703), most of Kabuki's essential characteristics became established. Large, commercial theatre buildings holding several thousand spectators were constructed in the three major cities—Edo (Tokyo), Kyōto, and Osaka. The stage, which previously had been simply a copy of the Nō stage, became wider, deeper, and was equipped with a draw curtain to separate acts; and in the early 1700s a ramp (*hanamichi*) was constructed from the rear of the auditorium to the stage for actors' entrances and exits. The idea of the rampway came from the Nō *hashigakari*, but in typical Kabuki fashion, it was transformed into an infinitely more theatrical device. From the puppet theatre Kabuki borrowed the use of fairly elaborate scenery, the revolving stage (100 years before its use in Europe), traps, and lifts. To the old Nō drums and flute were added the new *samisen*, a large drum, a dozen bells, cymbals, gongs, and two types of wooden clappers, making the resulting music flexible and varied.

Nō and *kyōgen* plays were often performed as Kabuki in the early decades. A print of approximately 1670 shows Kabuki actors performing "The Sumida River," with costumes and properties modelled closely on the Nō original. But it was not considered proper for "beggars of the riverbed," as Kabuki actors were called, to stage the art which had become the exclusive privilege of the warrior class. By Genroku times, new Kabuki dramatic styles had emerged. The actor Sakata Tōjūrō (died 1709) developed a relatively realistic, gentle style of acting (*wagoto*) for erotic love stories in Kyōto, while in Edo, a stylized, bravura style of acting (*aragoto*) was created at almost the same time by the actor Ichikawa Danjūrō I (died 1704) for bombastic fighting plays. In the play "Sukeroku: Flower of Edo" written by Tsuuchi Jihē in 1713, the two styles are blended most successfully. The hero, Sukeroku, is a swaggering young dandy and lover acted largely in the Edo style, while Sukeroku's brother, Shinbei, is a meek, gently comic foil in the Kyōto style. Genroku period Kabuki plays are lusty and active and contain much verbal and physical humour.

Kabuki theatres were required to be built in special entertainment quarters (similar to and often near licensed quarters for prostitution ordered established by the government at about the same time). Side by side in these entertainment quarters were puppet theatres. Puppets, imported from Korea centuries earlier, were fused with epic storytelling and the resulting narrated play accompanied by *samisen* music sometime before 1600. The earliest tales were about the Princess Jōruri (hence, *jōruri* as another name for puppet plays). This and other legends were in the nature of Buddhist miracle stories, the obligatory scene being one in which Buddha sacrifices himself or otherwise brings to life one of the main characters. Simple doll puppets, held overhead by one man, animated these blood and thunder *kojōruri* ("old *jōruri*") puppet plays.

The role of Chikamatsu in the puppet theatre

A new style of puppet play was created in 1686 by the writer Chikamatsu Monzaemon (died 1724) and the chanter Takemoto Gidayū at the Takemoto Puppet Theatre in Osaka, the city which was and continued to be the home of puppet theatre in Japan. Chikamatsu went on to become Japan's most famous playwright. Although he is best known for his puppet plays, he wrote Kabuki plays as well, most of them for Sakata Tōjūrō. From Tōjūrō, Chikamatsu learned the soft style of Kabuki performance and the situation that is so unique to early Kabuki, in which a comic lover visits a courtesan in the licensed quarter and quarrels with her. Between "Love Suicides at Sonezaki," written in 1703, and "Love Suicides at Amijima," written in 1721 three years before his death, Chikamatsu composed a dozen domestic tragedies handling the theme of lovers' suicide. As early as 1678 Kabuki plays were dramatizing current city scandals, lovers' suicides, murders, and tragic deaths. One of the most characteristic features of Kabuki was its contemporaneous dramatic subject matter; puppet drama was much changed when Chikamatsu brought this quality from Kabuki into his puppet plays.

In the 1720s and 1730s puppet plays gradually became more dramatic and less narrative under the influence of Kabuki. A revolutionary three-man puppet was created in which mouth, eyes, eyebrows, and fingers could move, encouraging writers to compose dramatic plays calling for complex emotional expression. A theatre manager and writer, Takeda Izumo (died 1756), collaborated with several other authors on all-day history plays, the so-called "Three Great Masterpieces" of puppet drama: "The House of Sugawara," "Yoshitsune and the Thousand Cherry Trees," and "Chushingara: The Treasury of Loyal Retainers." "Chushingara," the best loved and most often performed drama ever written in Japan, typifies mature puppet drama. It is based on an actual event that occurred in 1703: 47 retainers killed the enemy of their lord and were then sentenced to commit suicide by disembowelment (*seppuku*, or less properly, *harakiri*). The first four acts of the 11-act play dramatize the events leading up to the death of their lord; in the remaining acts the unswerving loyalty of a number of retainers is shown, which culminates in the 11th-act nighttime vendetta. The major scenes of Hangan's and Kampei's suicides are intensely emotional scenes of self-sacrifice. These scenes normally occur as the final section of the third act in a five-act history play and are called *sewa* ("family") scenes because, although the figures are samurai, tearful family separation is the emphasis of the scene. "Chronicle of the Battle of Ichinotani" contains a *migawari* ("child substitution") scene, typical of puppet history plays, which is, if anything, even more tear provoking: in response to the wishes of his lord Yoshitsune, General Kumagai slays his own son, so that the son's head may be substituted for that of a prince who has been condemned to die. Although the emotionalism and lack of humour of these puppet plays were foreign to Kabuki, they were so popular with audiences that they were adopted with little change to Kabuki. Today, the best puppet plays are equally a part of the Kabuki and puppet theatre repertories.

19th-century Kabuki plays

Few new puppet plays were added to the repertory after 1800 and in Osaka and Kyōto few Kabuki plays were being written. During the 19th century the most important Kabuki dramas were written in Edo, by Tsuruya



Bunraku, a scene from the comedy "Fishing for a Wife," a puppet performance of a Kabuki dance version of the *kyōgen* original play. Chief puppeteers wear conventional dress, minor puppeteers wear black. Pine-and-plank scenery indicates the Nō-*kyōgen* origin of the play.
Hiroshi Kaneko

Namboku and Kawatake Mokuami. They wrote all the standard types of Kabuki play—*sewamono* (domestic), *jidaimono* (history), and *shosagoto* (dance plays)—in large numbers; each wrote between 150 and 200 plays in his professional career. They were wholly products of Edo's urban culture. They spent their lives in the Kabuki theatre as writers. Although neither was formally educated, their plays reflect with great discernment the desperate social conditions that prevailed as the feudal system in Japan neared its collapse. Thieves, whores, murderers, pimps, and ruthless masterless samurai are major figures in a new type of play, *kizewamono*, or the raw domestic play, which Namboku created and Mokuami developed. They wrote for the talents of star actors: Namboku wrote for the finest *onnagata* (female impersonator) of his time, Iwai Hanshirō V, and Mokuami wrote for Ichikawa Danjūrō VII and a remarkable actor of gangster roles, Ichikawa Kodanji IV. Each was a master of Kabuki art, and between them they added new dimensions to Kabuki's stylized form. Namboku created rhythmic dialogue composed in phrases of seven and five syllables; Mokuami used puppet-style music to heighten the pathos of certain scenes and wrote elaborately conceived major speeches which required exceptional elocutionary skill on the part of the actor.

Meiji period. Nō, puppet theatre, and Kabuki were affected in differing degrees by the abolition of feudalism in 1867. At a stroke, the samurai class was eliminated and Nō lost its base of economic support. Important actors retired to the country to eke out a living as menial workers. For several years Nō was not performed at all, except that Umewaka Minoru, a minor actor, gave public performances in his home and elsewhere between 1868 and 1876. In 1881 a public stage was built in Shiba Park, Tokyo, for performances sponsored by the newly formed Nō Society and by its successor, the Nō Association. The most influential supporter of Nō during the Meiji period (1868–1912) was the aristocrat Iwakura Tomomi. The study of Nō came to be a highly regarded activity among the middle classes, and in time each of the five Nō schools (Kanze, Hōshō, Komparu, Kongō, and Kita) became financially stable, sponsoring their own performances and building their own theatres in the major cities.

The end of feudal society forced Nō to seek and cultivate a new audience; the popular audience of Kabuki and the puppet theatre, however, continued with little change during the Meiji period. Kabuki audiences remained large and loyal, but audiences for puppet plays continued to decline as they had for the previous hundred years. There was a brief revival of interest in Osaka puppet drama in the 1870s under the impetus of the theatre manager Bunrakuken (the popular term for puppet drama, *bunraku*, dates from this time). Learning to chant puppet texts became a vogue during the late Meiji period. Still, commercial theatres did not prosper. In 1909 the Shōchiku theatrical combine supported performances at the Bunraku Puppet Theatre in Osaka, but by 1914 this was the only commercial puppet house remaining.

As they always had, Kabuki writers and actors of the Meiji period tried to place current events on the stage. Thus, the actor Onoe Kikugorō V began acting in a series of contemporary plays, dressed in daily kimono or Western clothes and with his hair cut Western fashion (the origin of *zangirimono*, or the so-called "cropped-hair plays"), in the last decades of the 19th century. Western influence was also seen in theatre construction, with the first European style theatre built for Kabuki in Tokyo in 1878. Released from previous government restrictions, Kabuki artists created dance dramas from the Nō plays "The Ataka Barrier," "The Maple Viewing," and others, in which the elevated tone of the Nō original was purposely retained. Kabuki attendance was more than a million spectators yearly. But in spite of prosperity and seeming adaptation to new conditions, by the early decades of the 20th century, new artistic creation in Kabuki reached an end, and henceforth Kabuki would be restricted almost as much as *bunraku* and Nō to a classic repertory of plays.

Scholars and artists, learning of Western drama, organized successive groups designed to reform Kabuki—that is, to eliminate excessive stylization and to press for a more realistic manner of performance. The actor Ichikawa Danjūrō IX acted in historically accurate (and reportedly dull) *katsureki geki* ("living history" plays) written by the journalist Fukuchi Ōchi. Three *shin* Kabuki ("new Kabuki" plays) written by the scholar Tsubouchi Shōyō were influenced by Shakespeare, whose plays Tsubouchi was then translating. In 1908 a young actor, Ichikawa Sadanji II, returned from a year's study and observation in Europe. These and other influences produced few long-lasting changes in Kabuki, but they did set the stage for the creation of new kinds of drama that would depart radically from traditional forms.

Contributions of Kawakami

The first plays in Japan consciously based on Western models were those arranged and acted in by Kawakami Otojirō. Kawakami's first plays were political and nationalistic in intent. After performing in Europe and America (1899 and 1902), he introduced to Japan productions of Shakespeare, Maurice Maeterlinck, and Victorien Sardou. Most of these "new school," or *shimpa* plays, however, were little more than crude melodramas. Actresses performing in *shimpa* marked the first time women had appeared on the professional stage since Okuni's time. One *shimpa* troupe continues to perform today, in a style which retains turn-of-the-century sentiment and mannerisms.

In 1906 the Literary Society was established by Tsubouchi Shōyō to train young actors in Western realistic acting, thus beginning the serious study of Western drama.

Among the "new drama," or *shingeki*, groups devoted to studying and performing the works of such 19th-century European playwrights as Henrik Ibsen, George Bernard Shaw, Anton Chekhov, Leo Tolstoy, and Gerhart Hauptmann, were: the Art Theatre, founded by a Tsubouchi disciple, Shimamura Hōgetsu; The Free Theatre, modelled by Osanai Kaoru on the various Free Theatres of Europe; and The Stage Association. By the 1920s these groups had dissolved and others had taken their places. The major *shingeki* group after World War I was the Tsukiji Little Theatre. The members of *shingeki* troupes were earnest amateurs working for the day their efforts might raise their troupe to a level of professional excellence and financial independence.

Since World War II. In Japan during World War II Nō continued to be performed occasionally, Kabuki was limited to short daytime programs, while the majority of modern drama groups, pacifist or leftist, were forced to cease performing or to abandon politics. During the first three years of the American occupation (1945–1948) traditional theatre was censored; important Kabuki and *bunraku* classics, such as "The Subscription List" and "Chushingura," were banned. Modern drama, on the other hand, was encouraged as being democratic in spirit. The disillusionment of Japanese youth with wartime policies that had led to their country's defeat and occupation found fervent expression in postwar *shingeki* drama. Old groups were revitalized and scores of new groups sprang up. The Min-gei-za, or People's Art Theatre, staged plays strongly attacking militarism, including American action during the Korean War. The Bungaku-za, or Art Theatre, continued its prewar policy of producing translations of Western classics and new Japanese plays. The Haiyu-za, or Actors' Theatre, built its own theatre and established a dramatic school that became a training ground for a new generation of actors, directors, and writers. The resurgence of modern spoken drama peaked in the early 1960s.

Postwar *shingeki* drama

Avant-garde theatre groups have sprung up since the 1960s that perform happenings, multimedia presentations, and extemporaneous living theatre in any type of situation, from tiny cabaret rooms to travelling tent shows. The plays of established *shingeki* playwrights (Tanaka Chikao, Kinoshita Junji, Morimoto Kaoru, Iizawa Tadasu) are criticized by these young groups as being too Western-influenced and excessively literary. In another direction, highly acclaimed novelists Mishima Yukio and Ariyoshi Sawako have written plays consciously designed to appeal to an audience wider than the small group of intelligentsia traditionally associated with *shingeki*. Western ballet, modern dance, and opera may be seen in the same theatres that stage, at other times, popular sword-fighting plays (*shinkokugeki*), domestic comedies (*shinkigeki*), or all-girl operettas and revues (best known as *takarazuka*). Theatre plants are among the best equipped in the world.

The effects of modernization on performing arts in 20th-century Japan have been great. In the 1950s the country's motion-picture industry became the second largest in the world. Yet, in a decade it was being displaced by television, which soon reached the most isolated hamlets. The largest audiences today are those of the mass media; live dance and theatre, though still widespread, have seen their audiences decline.

In the traditional performing arts, professional Kabuki troupes perform under commercial Shōchiku sponsorship, and at the new National Theatre of Japan in Tokyo (since 1966); a single *bunraku* puppet troupe subsidized by the government performs to small audiences in major cities; Nō is avidly studied by thousands of amateurs, and frequent performances (in the form of recitals) can be seen in major cities; numerous schools teach Kabuki dance (*nihon buyō*) as an artistic accomplishment; musicians and dancers of the imperial *bugaku* group now give public performances each year at the National Theatre of Japan.

BIBLIOGRAPHY

China: L.C. ARLINGTON and HAROLD M. ACTON (eds. and trans.), *Famous Chinese Plays* (1937), partial translations of 33 plays, colour plates, and an authoritative introduction makes this early work still valuable; CYRIL BIRCH (ed.), *Anthology of Chinese Literature* (1965), contains recent translations of two Yüan plays: *Li K'uei Carries Thorns* and *Autumn in the Palace of Han;* D. KALVODOVA, V. SIS, and J. VANIS, *Chinese Theatre* (1959), impressions of a Peking opera performance, valuable for its many colour plates of costume and makeup; LIU WU-CHI, *An Introduction to Chinese Literature* (1966), contains the best analysis in English of individual playwrights and their works; A.C. SCOTT, *The Classical Theatre of China* (1957), a standard work; *Literature and the Arts in Twentieth Century China* (1963), a recent book containing information on theatre in contemporary China; *Traditional Chinese Plays*, 2 vol. (1967–69), translations of four Peking operas with detailed stage directions.

Korea: DUHYON LEE, *Korean Mask-Dance Drama* (1969), a useful publication of the Seoul Ministry of Culture and Information including a 20-page English summary and many photographs; SANG-SU CHOE, *A Study of the Korean Puppet Play* (1961), a detailed study with illustrations and translations of two play texts; WON-KYUNG CHO, *Dances of Korea* (1962), a short account by a professional dancer.

Japan: JAMES R. BRANDON and TAMAKO NIWA, *Kabuki Plays* (1966), contains translations with stage directions of two Kabuki dance plays: *The Subscription List* and *The Zen Substitute;* EARLE ERNST, *The Kabuki Theatre* (1956), the best description and analysis in English; and (ed.), *Three Japanese Plays* (1959), contains translations with stage directions of *The Maple Viewing* (Nō), *The House of Sugawara* (*bunraku*), and *Benten the Thief* (Kabuki); MASAKATSU GUNJI, *Buyo: The Classical Dance* (1970), a good introduction to Japanese dance; *Kabuki* (1969), a superbly illustrated introduction; DONALD KEENE and HIROSHI KANEKO, *Bunraku: The Art of the Japanese Puppet Theatre* (1965), a lavishly illustrated appreciation of the art; *Nō: The Classical Theatre of Japan* (1966), a superbly illustrated introduction; DONALD KEENE (trans.), *Major Plays of Chikamatsu* (1961), translations of 11 important puppet plays; RICHARD N. MCKINNON (comp.), *Selected Plays of Kyōgen* (1968), nine *kyōgen* comedies translated into sprightly English; YUKIO MISHIMA, *Five Modern Nō Plays* (1957), a modern rewriting of five Nō classics; NIPPON GAKUJUTSU SHINKOKAI, *The Noh Drama* (1960), the most complete and accurate single volume of Nō translations.

(J.R.B.)

Dandolo, Enrico

Enrico Dandolo, who was doge of the Republic of Venice from 1192 to 1205 and whose long and active life covered almost the whole of the 12th century, is remembered chiefly for his promotion of the Fourth Crusade, which led to the overthrow of the Greek Byzantine Empire and the aggrandizement of Venice.

Dandolo was born in Venice, probably in 1107. His father, Vitale, had held important public positions. During Enrico Dandolo's public life he was sent on many important missions for the Venetian government. He accompanied the doge Vitale II Michiel on an expedition to Constantinople in 1171. The following year, with the Byzantine ambassador, he went again to Constantinople, where, according to one account, he was so assiduous in defending the interests of the Venetians that the Emperor had him blinded. Later he was to be known as the blind doge. But the chronicler Geoffroi de Villehardouin, who wrote the history of the Fourth Crusade and knew Enrico Dandolo personally, stated merely that he did not see well because of an injury to his head. After his diplomatic mission to Constantinople, Dandolo went as ambassador to the King of Sicily (1174) and then to Ferrara (1191). When the doge Orio Mastropiero retired to a monastery, Dandolo was elected doge on June 1, 1192, at the age of 85.

His constitutional and legal reforms

In one of his first actions as doge, he swore the "ducal promise," spelling out the rights and duties of the office of doge. Dandolo also revised the penal code and published the first collection of civil statutes, setting the customary law of Venice on a firm juridical basis. He also revised the coinage, issuing a silver coin called the *grosso*, or *matapan*. This began a wide-ranging economic policy intended to promote trade with the East. Dandolo's image appears on the *grosso* coin; he is wearing a cloak and holding the "ducal promise" in his left hand and the gonfalon (banner) of St. Mark's in his right.

He also concluded treaties with Verona and Treviso (1192), with the Patriarch of Aquileia (1200), with the King of Armenia (1201), and with the Byzantine Empire (1199) and the Holy Roman Emperor (1201). He fought a victorious war against the Pisans in 1199.

But the prominent place Enrico Dandolo occupies in history must be attributed to the part he played in the Fourth Crusade: the arrangements made with the French barons for the transportation of their army; his provision of funds in exchange for their assistance in conquering Zara (Zadar), a Christian town on the Dalmatian coast then held by the King of Hungary; and his success in persuading the crusaders to help the Venetians conquer Constantinople. The personality of the doge stands out vividly in the accounts of the chroniclers. Although quite old, he was always found in the front line. At the assault of Constantinople he stood in the bow of his galley, completely armed and with the gonfalon of St. Mark's in front of him, encouraging his men as they made their landing.

Expansion of the Venetian Empire

After the capture of Constantinople, Dandolo took for himself and the doges of Venice the title "lord of the fourth part and a half of the whole empire of Romania." The title corresponded exactly to that part of the territories of the Byzantine Empire apportioned to the Venetians in the division of spoils among the crusaders. Since he had been one of the most powerful leaders of the expedition, Dandolo remained in Constantinople to direct all the operations there and also to look out for the interests of Venice. It is said that he had some valuable marble shipped to his son Renier for the construction of the great palace of the Dandolos on the Grand Canal. Ruins of a building in Moorish style and an ancient column of green marble were discovered in an excavation performed during the 19th century in the San Luca section of Venice, where the Dandolo palace had been located.

Dandolo died in Constantinople in 1205, at the end of May or the beginning of June. He was buried in the vestibule of the church of Santa Sophia in a marble tomb, on top of which was sculptured the doge's cap and the coat of arms of St. Mark's. The tomb was probably destroyed when Santa Sophia was converted to a mosque after the conquest by the Turks in 1453.

When Dandolo became doge, the Venetian republic faced considerable problems both internally and abroad. He resolved the internal problems by giving Venice an advanced civil code and constitutional system. In his pursuit of Venetian interests in the Adriatic and in the East, he was able, through shrewd commercial transactions, to acquire large territorial possessions. His burial at Constantinople was symbolic of that city's importance in the rise of Venice to wealth and power.

(An.L.)

Dante

The greatest poet of Italy, generally acclaimed with Shakespeare and Goethe as one of the three universal geniuses of western European literature, Dante Alighieri was also a prose writer, rhetorician, theorist of his own Italian vernacular literature, moral philosopher, and political thinker, with an immense variety of literary output. The loftiness of his art and the breadth and depth of his interests make him one of the most important figures of medieval and Romance literature. By writing his masterpiece *La divina commedia*, or *The Divine Comedy*, in Italian rather than in Latin he influenced decisively the evolution of European literature away from its origins in Latin culture and toward the expression of a new civilization. The poem itself is the greatest Christian poem. It is a profound vision of the medieval Christian world in terms of the principal problems with which it was most interested: man's moral "obligation to be"; the relationship between reason and faith; the value of learning and poetry as the means of understanding the supernatural; the reaching of the metaphysical through analysis of reality; and the understanding of the Christian revelation through theological study. But the choice of the vernacular Italian language, as opposed to Latin, as the means of communication and the wide range of styles employed mark a precise turning point in Italian literature in its earliest development. In addition to *The Divine Comedy*, the remaining works of Dante hold an important place in the history of Italian literature and make their essential contribution to the formation of a literary awareness and tradition, establishing new literary forms and new aims of thought and influencing his successors the poet Petrarch and the prose writer Boccaccio in the 14th century.

Alinari—Mansell



"Dante and His Work" by Domenico di Michelino, 1465. In Florence Cathedral.

Family and early years. Along with his early biographers and a few surviving documents, Dante, or Durante, Alighieri himself is the source of much biographical data about his noble ancestry and its origins and about the circumstances of his own birth and its date, which fell under the constellation of Gemini (between May 15 and June 15), 1265, in Florence; this biographical fact and many of those which follow are given, in particular, in passages in *The Divine Comedy*, in *La vita nuova*, and in *Il convivo*. He was the great-great-grandson of Cacciaguida, who was born at the end of the 11th century, within the "ancient boundaries" of Florence, was knighted by Conrad III, and died in the Second Crusade of 1147. In the *Paradiso* Dante traced his ancestry as far back as the legendary Roman origins of his city. His was a family of ancient urban nobility. His kin were not wealthy country landowners, although they did possess a few modest properties in the immediate outskirts of Florence. They were active instead in the economic life of the Florentine commune, or community—which was founded on trade and industry. Bellincione, Dante's grandfather, was a moneylender in Florence and in nearby Prato. Alighiero di Bellincione d'Alighiero, Dante's father, continued Bellincione's practice until the time of his death, which occurred before 1283. Dante seldom mentions these close relatives in his works; he may not have been proud of their activities. During the long struggle between the Guelf (papal) and the Ghibelline (imperial) parties for supremacy in the city, Dante's family's participation in the political life of Guelf Florence was of only minor importance. Bellincione and Alighiero do not appear among the names of those suffering hardship (as Guelf supporters) under the Ghibellines in their period of power between 1260 and 1266; moreover, Dante's father, after the Guelf defeat of Montaperti (1260), seems not to have been sent into exile by the Ghibellines. "The great town on the fair river of the Arno" was, by the time of the poet's birth, already devoting its energies to territorial and economic expansion, and he deemed this to be the principal cause of the internal discord afflicting the commune from 1216 onward. He described, in an episode of the *Paradiso* relating to Cacciaguida, how the steady infiltration of the feudal nobility into the city's economic

His family's economic activity

and political life had been a root cause of the great divisions and dislocations that shook medieval Florence. It had caused the magnates themselves to split into Guelf and Ghibelline factions, provoked the rich Guelf bourgeoisie's struggle against the magnates (with the former emerging powerful between 1250 and 1260); and had finally resulted in Guelf dominance after 1266, when Charles I of Anjou defeated and slew Manfred the imperial claimant at the Battle of Benevento. From then on Florence remained within the sphere of Angevin and papal influence but not without fierce social strife, resulting from the antimagnate politics of the Guelf commune. Unrest was further aggravated by external conflicts caused by campaigns of territoral conquest. Such was the background of events known to the young Dante that in their later development were to provoke active participation by the mature poet. Dante grew up—after the death of his mother, Bella (probably of the Abati family), and the remarriage of his father to Lapa di Chiarissimo Cialuffi—in the company of an older sister and a half brother, Francesco, and half sister, Gaetana, from his father's second marriage.

Early studies and first writings. Dante's precocious studies of grammar and rhetoric brought him into contact with the Latin authors; but it was also the stimulation of Florence's cultural environment that fostered a natural inclination toward poetry while he was still a youth. Florence in those years was the meeting place of many literary currents. For Dante, in the years of his literary apprenticeship, two masters stood out above all the others: Brunetto Latini and Guido Cavalcanti. Dante himself confesses that it was Brunetto who taught him "how man becomes eternal," that is, how he leaves a lasting trace of himself through his literary works. But beyond the poet's own important admission, it is easy to grasp the extent of the influence; Latini, who had returned to Florence from exile in France in 1266 and, after having regained important public appointments in the commune, died in about 1294, almost certainly furnished Alighieri with the first documents of *Ars dictandi* —that is to say, of the art of letter writing—and so of writing ornate Latin prose; also, active as a rhymer in vernacular Italian and as a philosophical prose writer, he undoubtedly represented a model which Dante kept before him from the beginning of his apprenticeship as poet. Brunetto's rhetorical and literary instruction is evident in the numerous adaptations from his texts, present in Dante's major and minor works, both in Latin and in Italian, and particularly in two early poetic exercises, forcefully executed and not far apart from each other—the *Detto d'amore* and the *Fiore* (adaptations in Italian verse of the widely known French poem, the *Roman de la rose*). These works show the rhetorical technique and popularizing style dear to Latini and yet give evidence of a spirited freshness.

Influence of Brunetto

Besides the "paternal image" of Brunetto, Dante was influenced by the circle of Florentine poets who worked in the wake of the Sicilian school and of Guittone d'Arezzo. It was the poetry and friendship of Guido Cavalcanti, however, that most influenced the concrete development of Dante's art. Having reached his majority (at the age of 18, being fatherless) and being about to take Gemma Donati as his wife (he had been betrothed to her since 1277) in about 1285, Dante sent Cavalcanti a sonnet, "To Every Captive Soul and Gentle Heart." It was the sonnet that reappeared nearly ten years later in the opening of the *Vita nuova* ("The New Life"), which was dedicated to Cavalcanti. Dante's first poetic experience thus grew out of schemes set forth by the Sicilian school and by Guittone. And then, under Cavalcanti's influence, accents of acute distress and tormented love appeared beside the light, graceful tones of certain ballate (ballades), in new, more dramatic forms. Dante's poetry later acquired a marked individuality, and in the so-called rhymes of praise for Beatrice, his ideal lady, the poet distinguishes his own style from the models of traditional love poetry, fully developing the lesson of Guido Guinizzelli (the master of his own masters) and transcending it with the canzone, or lyric ("Ladies Who Have Understanding of Love"). This canzone was a true poetic manifesto of the *stil nuovo*, the "new style." Under the stimulus of new ideas and that of the rhymes of praise, Dante became the champion of a kind of love poetry connected with his discovery of the value of the beauty of Beatrice as a mystic analogue and revelation of the divine. (This cultural attitude toward woman as a means of metaphysical knowledge of the divine sustains the masterly poetry of the third book of the *Divine Comedy*, the *Paradise*.) And he persuaded himself of the need to renounce every hope of and desire for concrete amorous recompense. These achievements foreshadowed the beginnings of Dante's ideological clash (later expressed in literary and political terms) with Guido Cavalcanti and resulted in their separation.

Influence of Cavalcanti

The "Vita nuova," as literary form and autobiography. The death of Beatrice Portinari, on June 8, 1290, was succeeded, for Dante, by a period of intense study. The texts of such authors as Boethius, Cicero, St. Augustine, Aristotle, and other philosophers read during this period were a certain source for many changes in his works. The *Vita nuova* (*c.* 1293) represented a crystallization of these innovations. This work is a collection of 31 lyric poems, composed between 1283 and 1291, arranged within a prose framework with a conceptual plot. Ideally a "book of the memory," the work is based on the experiences of Dante in relation to his adolescent love for Beatrice. The lyric poems, upon their insertion into the prose framework, underwent stylistic changes. The prose that orders these verses and explains their literary and spiritual meaning takes on the form of an autobiographical novel and commentary and describes what has been well defined as a *"legenda sanctae Beatricis"* ("legend of Beatrice sanctified"). The autobiographical material covers about 18 years, beginning with Dante's first meeting with Beatrice, at the age of nine. After their second meeting, nine years later, the beatific experience of her gentle salutation invokes, for the poet, in a "dream-vision," a presentiment of her imminent death, seen as a culmination of the ultimate perfection of nature. The work narrates the events of about ten years more. Dante, wishing to conceal his intense love for Beatrice, which had totally engrossed him, pretended amorous sentiments for two other women, writing for them "certain trifles in verse." These were interpreted by many as "beyond the limits of courtesy," and Beatrice herself refused to greet him. So Dante, deprived of his only bliss, addressed himself to Beatrice, declaring in verse the nobility and loftiness of his love, asking nothing and hoping for nothing but content unrecompensed to praise her spiritual beauty. The praise itself becomes joy and reward. The verses of praise and the canzone "Ladies Who Have Understanding of Love" mark the beginnings of a new style in Dante's poetry, that of the New Verse, which later became the exemplar of the *stil nuovo;* and these poems in the *Vita nuova* mark the beginnings also of the theme (which later underlies the poetry of the *Paradise*) of the parallel relationship between Beatrice the lovely creature and God the Creator, leading to the notion of the identity of love and a gentle heart—*i.e.*, of the metaphysical cause and its effect. During an illness, Dante had a second vision of the imminent death of Beatrice, which provoked a period of sorrowful meditation. Upon her actual death, in 1290, he was comforted by the certainty that Beatrice had soared "to heaven on high, to the kingdom where the angels have peace." After the first anniversary of her departure, Dante was moved by the compassion of a "fair lady, young and quite lovely," so that for a short while he withdrew from his sorrow. But if the heart gives in to a natural need for comfort, reason keeps watch and imposes victory over "evil desire"; the image of Beatrice returned with greater force, provoking shame and remorse. Thus Dante reached the final, supreme "wonderful vision" of Beatrice in the glory of the empyrean, probably foreshadowing her glorification in the *Paradise* "upon the throne her merits have assigned her." Following this vision, the poet proposed "to say no more about his blessed lady until he could do so in a way more

Experiences underlying the *Vita nuova*

Dante's supreme vision of Beatrice in glory

worthy of her" and say "what hath not before been written of any woman." This was his promise of a full, new, and eminently poetic glorification from which, about 15 years later, the *Divine Comedy* derived its origin and force. But already, in the *Vita nuova*, Beatrice appears, in life and death, as the first of the great inspiring muses of modern poetry, comparable in importance to Laura, the inspiration of Petrarch, half a century later.

Political life and exile in 1302. The death of Beatrice marked a precise turning point in the poet's life. For about 30 months (as he said in the *Convivio*) he attended "the schools of the religious and the disputations of the philosophers" and in deep study prepared the ground for the great works of his maturity. He became acquainted with such classical authors as Virgil, Ovid, Lucan, and Statius. Through his reading of St. Thomas Aquinas' commentary he came to know the works of Aristotle, especially *Ethics* and *Politics;* and he read medieval authors such as Albertus Magnus, St. Bonaventure, and Averroës. These studies gave a new dimension to his poetry, extending its limits beyond the themes of love, as is reflected in the allegorical canzoni in praise of philosophy and his doctrinal verses. To the same period belong four *rime petrose* (December 1296) written for Pietra, a lady who had rashly rejected his love; they too give evidence of mature experiment in metre and style, modelled on that of a Provençal troubadour, Arnaut Daniel, and look forward to the style of the *Divine Comedy*. The philosophical canzoni, with their judgment of current ideas and modes of life, are the fruit of meditation on daily experience of the class prejudices that underlay the violence of the magnates, a theme that was to recur in the eighth canto of the *Inferno* with the episode of Filippo Argenti. Dante already adhered firmly to the democratic ideals of the Guelf commune. In the meantime he had taken part as a cavalryman in the Battle of Campaldino (June 11, 1289) against Ghibelline Arezzo and in the military operations two months later against the Pisan castle of Caprona. Following the overthrow of Giano della Bella in 1295, and the subsequent reform of his *Ordinamenti di Giustizia* (Ordinances of Justice) of 1293, which had severely limited the power of the nobility, the poet enrolled himself in the Guild of Doctors and Pharmacists (for his philosophical studies) and became politically active. Between November 1295 and April 1296, he was a member of the special Council to the Captain of the People (upholder of popular authority in parallel with the podesta, or supreme authority), although he never took the floor and in fact five times had to justify his absence from meetings. Between May and September 1296 he belonged to the general Council of the Hundred (citizens' parliament) and supported legislation against the magnates. The moral and democratic leanings displayed in these stands was also expressed by Dante in the above mentioned canzoni celebrating the moral virtues of nobility and loveliness.

Political activity

Little is known about the poet's activity in the immediately succeeding years, and there is no documentation of Dante's official position regarding the events of 1297, when Pope Boniface VIII proclaimed the crusade against the powerful, anti-papal Roman Colonna family and sent Cardinal Matteo d'Acquasparta to Florence to solicit help. But an episode about Guido da Montefeltro in the *Inferno* clearly expressed the poet's opinion about the "arrogant fever" of Boniface's thirst for power, which led to the stirring up of open conflict between the two Guelf factions of the city, the Black Guelfs (headed by the Donati family, of magnate ancestry) and the more moderate White Guelfs (headed by the Cerchi, a family of bankers and merchants). Donati support of the Pope transformed what had until then been a struggle between municipal parties into a conflict between the commune and the papacy. The *signoria* of the White Guelf party struck hard against the Blacks, sending them into exile. Boniface sent Cardinal Matteo d'Acquasparta to Florence a second time, in 1300, ostensibly to reconcile the parties but secretly to favour the Blacks. On May 7 Dante was nominated ambassador to San Gimignano to consolidate the Guelf League, then supporting the Pope in a war against the Aldobrandeschi family of Santa Fiora. He was elected one of the six priors, or presidents, of the guilds for the period June 15–August 14, 1300, during which time he and his colleagues gave proof of their impartiality, exiling the heads of both parties; among the latter was Guido Cavalcanti, his "first friend," exiled to Sarzana. But events followed close one upon another: during the succeeding priorate (to which Dante no longer belonged) the *signoria* government under the Whites recalled their followers from exile, while the Blacks assembled in the spring of 1301 in the church of Sta. Trinità to make another bid to recover power. These were the last months of Dante's political and civic activity: surviving documents present him as the leader of a large group of Whites; and the role that he played in Florentine events between 1300 and 1301, especially after his priorship, cannot be minimized. On September 13,1301, and then again on the 20th and 28th of the same month, he urged that full powers be given to the priorate, in view of the dangers threatening the city. In fact, on October 4, Charles of Valois, brother of the French king, Philip IV the Fair, in concert with the Pope, arrived at Castel della Pieve, the Black Guelfs' headquarters on the boundary of Florence. The White party controlling the *signoria* attempted once more to reach a compromise, sending an embassy of three citizens with four Bolognese doctors of law to the Pope. Dante Alighieri, one of the three ambassadors, left Florence at the end of October, never again to return. On November 1, 1301, Charles entered the city and soon won the support of the more extreme Blacks. As the struggle between the parties degenerated into fierce political persecution with the victory of the Blacks, trials of Whites, accused of Ghibellinism (pro-imperialism) and public embezzlement, were instituted. Dante was among the first and most severely stricken, being condemned for barratry (poor administration of the public funds) on January 27, 1302, and sentenced to pay 5,000 small florins, an enormous sum in those days, within three days and to remain outside Tuscany for two years. With his colleagues, Dante was also accused of opposing the Pope and Charles of Valois, as well as of having favoured the scission of parties in Pistoia that had damaged the Blacks. The poet, journeying from Rome, was not present to pay the fine, and so on March 10, he and his 14 co-defendants were by default condemned to death.

Election to the priorate

Political persecution of the White party

Life in exile. The poet reacted strongly and united with the other White and Ghibelline (pro-imperial) exiles to seek military aid from other pro-Ghibelline families—the Ubaldini in Mugello, Scarpetta Ordelaffi in Forlì, and Bartolomeo della Scala in Verona—in an attempt to re-enter Florence by force. Dante participated as an ambassador at those negotiations, and on June 8, 1302, in the church of S. Godenzo, with 16 other Florentines he undertook to compensate the Ubaldini in the event of damages resulting from military operations against Florence. When Pope Benedict XI succeeded to the papacy (October 1303), the exiles saw better prospects for peace. In March 1304, Benedict sent Cardinal Nicolò da Prato to Florence to mediate between the parties. On this occasion Dante composed Epistle I, on behalf of the exiles gathered in Arezzo; the letter accepting his mediation was sent to the Cardinal, but the Blacks succeeded in wrecking the negotiations and forced the Cardinal to leave the city after he had issued an interdict. On the death of Benedict in July 1304, the Whites and Ghibellines took up arms again. Their defeat at La Lastra above Florence (July 20, 1304) ended all hope for the exiles. Dante had already expressed his dissent regarding the decision taken: supporting a policy of reconciliation, he had long since refused to take up arms against his native city and made "a party by himself." Evidence of this attitude can be found in his other works of these first years of exile—in the envoy of the canzone "Three Women Have Come Round My Heart" (1304), in a moving reference to Florence in *De vulgari eloquentia* ("Of Eloquence in the Vulgar Tongue") and in the *Convivio* ("The Banquet"); later, this attitude developed into an expressive pathos that

animates the episode of Farinata degli Uberti in *Inferno*, X, representing the exaltation of magnanimous patriotic love.

The De vulgari eloquentia and the Convivio. Forced to wander about Italy alone in poverty (he went to Forlì and Verona in 1303), seeking protection and friendship, Dante's only comforts were study and poetry.

Dante at Bologna

Between 1304 and 1306 he was taken in by Bologna, a favourable environment for studies in philosophy, law, and rhetoric. Two of his major works were begun here: *Il convivio* and the *De vulgari eloquentia*, which were written to console himself for his painful solitude and also to show himself to Florence and the world as a man of culture, a philosopher and a poet, whose perspectives of thought and art were enlarging far beyond the "municipal" dimensions of his early years. These two works, both of them unfinished, were permeated by a deep nostalgia for his distant native city. They are both written with a concise, almost effortless rhythm and with the enthusiasm of mature insight. They differ, however, in outlook and in purpose.

Originally planned in four books but interrupted at chapter xiv of Book II, the *De vulgari eloquentia* (1304–05) was written in Latin, and it reveals in its interior biography of Dante his linguistic, philosophical, and rhetorico-artistic preoccupations. It is a clear literary manifesto directing poets to establish norms that can be widely understood. In Book I, Dante set up an opposition between *locutio vulgaris* (language as means of communication, variable in time and space) and *grammatica* (literary language, stable in being fixed by rules). The basic and yet paradoxical aim of the work was to give rules (thus artificial stability) to the Italian vernacular. The poet delineated the historic evolution of language from the first human tongue (that of Adam, preserved in the Hebrew language). He went on to examine in particular the three vernaculars (of Provence, France, and Italy) belonging to the southwestern nucleus, and then concentrated his attention on the Italian, selecting one of its many dialects for elevation to the status of a literary language. Having examined them all, Dante found evidence of this language (defined as the "illustrious vulgar tongue") only in examples of the work of those poets (the Sicilians, the Bolognese school, and the Florentines of the *stil nuovo*, including Dante himself) who could detach themselves from their local manner of speech. The highest of the styles recognized by the rhetoricians—the tragic, with its elevated themes of arms, love, and virtue —was appropriate to this vulgar tongue.

The *De vulgari eloquentia* remained unfinished. On October 2, 1306, the commune of Bologna expelled the Florentine exiles. The *Convivio* was continued, in spite of Dante's renewed wanderings, until the end of 1307 or the beginning of 1308. It was interrupted, however, by the election in November 1308 of Henry of Luxembourg as Henry VII, king of the Romans, an event that the poet had earlier anticipated in the work. A no less important cause of the interruption was the burning conception and vast design of *The Divine Comedy*.

First planned as a work of 15 treatises, the *Convivio* gives its readers a portrait of Dante as a vigorous and mature though sometimes eclectic thinker and as a man passionately interested in science. His encyclopaedic and didactic intention, as well as choice of material, link him with his old teacher Brunetto Latini. Latini, in his encyclopaedic *Trésor*, had written in French in order to break away from the traditional use of Latin in scientific works. Dante, instead, employed his own native vernacular and so set down the foundations for Italian literary prose. In the first treatise he stated his intention to set out material for 14 moral and philosophical canzoni. In the second book, he referred again to the episode of the "fair lady" in the *Vita nuova* and interpreted it as a conflict between his love of Beatrice and his love of philosophy. The third book was dedicated to the glorification of philosophy, and the fourth expounded the nobility of man, which is not tied to hereditary privilege or to wealth but to the goodness of individual human nature and the possession of moral and intellectual virtues. The *Convivio* contains themes and conclusions which were later developed more fully: for example, the digression on the four senses of the Scriptures (sacred and profane); on the two kinds of allegory; and on the authority of the emperor and the necessity of the empire for the well-being of humanity.

On October 6, 1306, Dante, expelled from Bologna, had concluded peace at Sarzana between Marchese Franceschino Malaspina and the Bishop of Luni. In 1308 he was probably in Lucca, where the presence in that year of his eldest son, Giovanni, is confirmed by a reference to Gentucca, a girl of Lucca, in *Purgatory*, XXIV, the second part of the *Divine Comedy*. Having returned to the Casentino (the country north of Arezzo), he sent shortly afterward to Moroello Malaspina, a Guelf leader, the canzone "Love, Since After All I Am Forced to Grieve" with an explanatory Epistle IV. He was probably in the Casentino when he learned of the election of Henry of Luxembourg as German king (1308). In exile he had long meditated the events that had overtaken him and had become convinced that those events and the ensuing disorder had occurred only because there was no Holy Roman emperor. The heart of the exile exulted, therefore, while the deliverer prepared to come to Italy. And when Clement V agreed to crown the new emperor in Rome, the poet made himself the sounding board of the general expectation and delight, in Epistle V ("*Universis et singulis*," 1310), urging the princes and peoples of Italy to rejoice at the coming of the *Rex pacificus*. Only Florence, allied to Robert of Naples, opposed the emperor; and the poet expressed his indignation in Epistle VI (March 31, 1311) against the "most wicked Florentines." He then turned to Henry, to whom homage had already been paid in Milan, and urged him to crush the head of the viper that attempts to bite its mother (Epistle VII, April 17, 1311). For this plea, Dante was excluded from an amnesty granted by Florence to other exiles on September 2, 1311, as Henry was approaching the city to besiege it. Out of reverence for his native city, which he described in *Paradise* as an "evil stepmother," he did not appear among the exiles who encamped with the imperial troops on the plain of San Salvi, and his name is missing from a list of the condemned published by the commune in March 1313.

Dante's delight in the prospective imperial coronation

The Monarchia and later minor works. The treatise of political philosophy, composed in Latin and arranged in three books, is definitely later than the *Convivio*, the arguments and themes of the fourth book of which it develops. It was composed in the years immediately following the arrival of Henry VII in Italy. The first two books of the *Monarchia* treated again the *Convivio*'s theme that the empire was necessary to the well-being of the world, while the third and last book considered whether the emperor depended directly upon God or derived his power through the pope, vicar of God. For Dante, imperial authority, as that of the pope, issued directly from God, so that direct power in temporal things was denied to the church, the spiritual organ. The treatise adumbrated the two ends assigned by God to all humanity, one of which is attainable in time, the other in eternity: earthly happiness is attainable with the guidance of the emperor; and the supreme happiness of celestial paradise is attained only with the guidance of the pope. The emperor, as devoted son of the father, must thus give reverence to the pope.

Dante's view of imperial and papal authority

A rift that soon opened between Henry and Clement V and the sudden death of the Emperor dashed the poet's hopes; after staying for some time in Tuscany he returned about 1316 to Verona, where Cangrande della Scala (imperial vicar nominated by Henry VII in 1312) was founding a powerful Ghibelline state in northern Italy. To these years belong the last three letters now known: Epistle XI (June 1314; urging the Italian cardinals in conclave after the death of Clement V to elect an Italian pope who would bring the seat of the papacy back to Rome from Avignon); Epistle XII (May 1315; to a Florentine friend, refusing an amnesty because its conditions are regarded as humiliating); and

Epistle XIII (*c.* 1316; dedicating *Paradise*, which he had just begun, to Cangrande, and setting down the beginning of a commentary on it together with the general framework of the *Commedia* and an explanation of its literal and allegorical meanings). This began the final stage in the life of Dante: he left Verona about 1318 and went to Ravenna as guest of the poet Guido da Polenta. In the tranquility of Ravenna, his surviving children gathered round him: Pietro, Iacopo, and Antonia (who after the death of her father was to become a nun, taking the name of Sister Beatrice). He composed two Latin eclogues there in reply to Giovanni del Virgilio (a Bolognese grammarian), who had urged him to write a poem in Latin verse on a historical theme and had invited him to Bologna, promising him the poetic laurel. He may have stayed briefly at Verona. On his return to Ravenna from Venice, where he had been sent on a difficult embassy by Guido da Polenta, Dante, who had only very recently finished *Paradise*, was struck down by malaria and died during the night of September 13–14, 1321. But he left to Italy and to the world the *Commedia*, which posterity judged to be divine.

"The Divine Comedy." Dante began working on the *Commedia* in about 1308, after leaving off the *Convivio* and *De vulgari eloquentia*. On a vast fresco depicting his sorrows and hopes, his fierce hatreds and his cherished beliefs, he reaffirmed in poetic terms his ethical and political conception of the world and of the ends and duties of man within the twofold order of nature and grace. A poet above all, he felt that only in poetry would he be able to express fully his dream of a spiritual and civilized renewal of the whole of humanity. The poem, though unique, (1) is closely related to the tradition of medieval allegorical poetry, at the same time, (2) is modelled on Virgil's *Aeneid* (considered in the Middle Ages to be an allegorical work), and (3) is inspired by the poetry of the Bible and by the Christian wisdom of the Holy Scriptures. Divided into three books, or *cantiche* (treating of Hell, Purgatory, and Paradise—the first composed by 1312, the second by 1315, the third between 1316 and 1321), it is written in Italian vernacular and is composed of 14,233 hendecasyllables in terza rima (a rhyme scheme aba, bcb, cdc, and so on), arranged in 100 cantos (one being a prologue to the entire work, and each *cantica* having 33 cantos). Thus the number 3, a symbol of the Trinity, is always present in every part of the work, with its multiples and in its unity. The title, with its medieval meaning, is *Commedia* because the subject, horrible in the *Inferno*, becomes desirable and pleasing in the other two *cantiche*, *Purgatory* and *Paradise*. The literal subject of the work is the journey Dante makes through the world beyond the grave, repeating by divine grace the experience of Aeneas in his descent to Avernus. At the age of 35, on the evening of Good Friday, 1300, the poet finds himself wandering astray in a dark wood. After a night of anguish, he sets out toward a hill illuminated by the sun, but three wild beasts—a female ounce (a species of leopard), a lion, and a wolf: symbols of lust, pride, and avarice—bar his path and force him back toward the darkness of the wood. Virgil, however, sent by the Virgin Mary, St. Lucy, and Beatrice, appears to help him. He guides Dante through the infernal realm and the mountain of Purgatory, at the summit of which the Roman poet is replaced by Beatrice, who then conducts Dante (raising him from heaven to heaven by the brilliant and loving power of her glance, which is that of a blessed soul contemplating God) as far as the empyrean, where the poet enjoys for a brief moment the supreme vision of the divinity.

Modes for the *Commedia*

Time has confirmed the *Commedia* to be a sublime work of poetry, for its greatness of conception and construction, for the concreteness of its imagery, for its characters—prismatically reflected within the framework of the imaginary journey—and for its artistic mastery and impassioned moral power. For many readers it is equalled only by the great poems of classical antiquity and the sublime poetry of the Bible. Dante intended the poem to have, in addition to its purely literary sense, an allegorical meaning that is not superimposed from outside but is the conceptual nucleus around which the poetic images flower; happiness on this earth (represented by the terrestrial paradise) and blessedness not on earth but in eternity (found in the vision of God in paradise) are the ends assigned by Providence to all humanity and are attainable with the help of the two guides—emperor and pope—assigned by God to man. Earthly happiness, Dante held, is acquired through the practice of the moral and intellectual virtues; celestial happiness by living this life according to the Christian virtues of faith, hope, and charity. Dante's progressive liberation from the slavery of passion corresponds—in the progressive spiritual probing that guides the poet—with his every judgment of men and events, of the passions and inclinations of troubled humanity, and corresponds, too—at the poem's structural level—with his overcoming, in the course of his journey, a particular sin or reaching a particular level of blessedness. The poet proceeds in continual ascent from the "dark wood" (symbol of the uncultivated life, without reason or virtue) to the possession of God as infinite good, which becomes the reward for the soul that contemplates it in the luminous darkness of faith. The poetic experience of the *Commedia* is allegorically the history of Dante's own soul, the journey of his mind to God, serving as a prime example for every reader and helping him to rediscover the "straight way" of a moral life that leads to perfection. Dante wanted readers of the *Commedia* to find in it, as in a great, sacred allegory, an exemplum, a picture, that would lead them to meditate on the evil of the tragic contemporary scene and that would also indicate the solutions: the empire as the sole remedy and bridle for human avarice and the necessity for a return on the part of the church to purity and evangelical poverty. In this sense, taken with Epistle XIII to Cangrande, the purpose of the *Commedia* can be said to be to "remove those living in this life from the state of misery and lead them to the state of felicity," because Dante's exceptional poetic experience could be repeated at the level of ordinary life by anyone who wished to move with equal resolve toward the same spiritual goal.

MAJOR WORKS

POETRY: *La vita nuova* (*c.* 1293; trans. by B. Reynolds, 1969); *Commedia*, consisting of *cantiche* I, *Inferno; cantiche* II, *Purgatorio; cantiche* III, *Paradiso* (composed *c.* 1310–21; *The Divine Comedy*, trans. by. W.F. Ennis, 1965); *The Comedy of Dante Alighieri*, verse trans. by D.L. Sayers and B. Reynolds, 3 pt. (1949–62).

OTHER WORKS: *De vulgari eloquentia* (1304–05; Eng. trans., *Dante's Treatise "De Vulgari Eloquentia,"* trans. by A.G. Ferrers Howell, 1890); *Il convivio* (*c.* 1304–07; *Dante's Convivio*, trans. by W.W. Jackson, 1909).

BIBLIOGRAPHY. PAUL COLUMB DE BATINES, *Bibliografiia dantesca (1845–46)* (1888); CORNELL UNIVERSITY LIBRARY, *Catalogue of the Dante Collection . . .*, 2 vol. (1898–1900) and *Additions, 1898–1920* (1921); GIULIANO MAMBELLI, *Gli annali delle edizioni dantesche* (1931). For post-World War II studies: ALDO VALLONE, *Gli studi danteschi dal 1940 al 1949* (1950); ENZO ESPOSITO, *Gli studi danteschi dal 1950 al 1964* (1965). Annual bibliographies of Dante studies published in America are printed in *Dante Studies*, published by the Dante Society of America. Bodies specializing in Dante studies have been established in many countries. Apart from the Società dantesca italiana (founded 1888), of special interest to English-speaking readers are the Oxford Dante Society (founded 1876) and the Dante Society of America (founded 1881).

Collected editions: Complete critical editions of Dante's work include MICHELE BARBI *et al.* (eds.), *Le Opere di Dante. Testo critico della Società dantesca italiana*, 2 vol. (1921–22; 2nd ed., 1960); EDWARD MOORE (ed.), *Le Opere di Dante* (The Oxford Dante; 4th ed. rev. by PAGET TOYNBEE, 1924). Invaluable for textual scholarship, the National Edition of the Società dantesca italiana is in course of preparation (1932–).

Commentaries: For extracts from early commentaries, see GUIDO BIAGI *et al.*, *La Divina Commedia nella figurazione artistica e nel secolare commento* (1921–40), and its useful bibliography. Modern commentaries include those by GIUSEPPE VANDELLI, 16th ed. (1955); CARLO GRABHER, 26th ed., 3 vol. (1966); MANFREDI PORENA, new ed., 3 vol. (1954–55); ATTILIO MOMIGLIANO, 3 vol. (1964); and NATALINO SAPEGNO, 3 vol. (1957). See also FRANCESCO MAZZONI, *Saggio di un nuovo commento alla Divina Commedia* (1967). For English-

speaking readers, the commentary of CHARLES H. GRANDGENT, *La Divina Commedia di Dante Alighieri*, rev. ed. (1933), is excellent; as is CHARLES S. SINGLETON, *The Divine Comedy* (1970–).

Aids and introductory works: Of the general works available to English-speaking readers, see especially ERNEST H. WILKINS and THOMAS G. BERGIN, *A Concordance to the Divine Comedy of Dante Alighieri* (1965). EDWARD S. SHELDON and ALAIN C. WHITE, *Concordanza delle opere italiane in prosa e del Canzoniere di Dante Alighieri* to the minor Italian works (1905); and EDWARD K. RAND and ERNEST H. WILKINS, *Dantis Alagherii operum latinorum concordantiae* to the Latin works (1912), are also useful. PAGET TOYNBEE, *A Dictionary of Proper Names and Notable Matters in the Works of Dante* (1898; rev. by CHARLES S. SINGLETON, 1968), is invaluable. Excellent introductions to Dante include UMBERTO COSMO, *Guida a Dante* (1947; Eng. trans., *A Handbook to Dante Studies*, 1950); MICHELE BARBI, *Dante: vita, opere e fortuna* (1940; Eng. trans., *Life of Dante*, 1954); THOMAS G. BERGIN, *Dante* (1965). Also useful are JEFFERSON B. FLETCHER, *Dante* (1965); FRANCIS FERGUSSON, *Dante* (1966); THOMAS C. CHUBB, *Dante and His World* (1966); and ALDO VALLONE, *Dante* (1971).

General studies: EDWARD MOORE, *Studies in Dante* (1896–1917, reprinted 1968); PAGET TOYNBEE, *Dante Studies* (1921); BENEDETTO CROCE, *La poesia di Dante*, 3rd ed. rev. (1922; Eng. trans., *The Poetry of Dante*, 1922); T.S. ELIOT, *Dante* (1929; reprinted in *Selected Essays, 1917–1932*, 3rd ed., 1951); JOHN FRECCERO (ed.), *Dante: A Collection of Critical Essays* (1965); UBERTO LIMENTANI (ed.), *The Mind of Dante* (1965); the OXFORD DANTE SOCIETY, *Centenary Essays on Dante* (1965); WILLIAM J. DE SUA and GINO RIZZO, *A Dante Symposium* (1965); FRANCESCO MAZZONI, *Contributi di filologia dantesca* (1966).

Specialized studies: On the *Vita Nuova*, see CHARLES S. SINGLETON, *An Essay on the Vita Nuova* (1949); on the *Canzoniere*, PATRICK BOYDE, *Dante's Style in His Lyric Poetry* (1971); on Dante's philosophical thought, ETIENNE GILSON, *Dante et la philosophie* (1939; Eng. trans., *Dante the Philosopher*, 1948); on Dante's political thought, ALESSANDRO PASSERIN D'ENTREVES, *Dante As a Political Thinker* (1952); EWART K. LEWIS, *Medieval Political Ideas* (1954); CHARLES T. DAVIS, *Dante and the Idea of Rome* (1957); on the *Commedia*, W.H.V. READE, *The Moral System of Dante's Inferno* (1909, reprinted 1969); KARL VOSSLER, *Die göttliche Komödie*, 4 pt. (1907–10; Eng. trans., *Medieval Culture: An Introduction to Dante and His Times*, 2 vol., 1929, reprinted 1958); ERNST R. CURTIUS, *European Literature and the Latin Middle Ages* (1963); FRANCIS FERGUSSON, *Dante's Drama of the Mind* (1953); CHARLES S. SINGLETON, *Dante Studies:* vol. 1, *Commedia: Elements of Structure* (1954) and vol. 2, *Journey to Beatrice* (1958); JOHAN CHYDENIUS, *The Typological Problem in Dante* (1958); JOSEPH A. MAZZEO, *Structure and Thought in the Paradiso* (1958) and *Medieval Cultural Tradition in Dante's Comedy* (1960); IRMA BRANDEIS, *The Ladder of Vision: A Study of Dante's Comedy* (1960); HELEN F. DUNBAR, *Symbolism in Medieval Thought and Its Consummation in the Divine Comedy* (1961); THOMAS G. BERGIN, *Perspectives on the Divine Comedy* (1967) and *The Diversity of Dante* (1969).

Illuminated manuscripts: PETER BRIEGER, CHARLES S. SINGLETON, and MILLARD MEISS, *Illuminated Manuscripts of the Divine Comedy*, 2 vol. (1969).

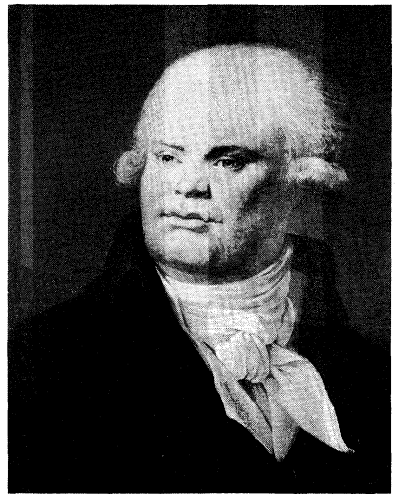
(F.Ma.)

Danton, Georges

Few leaders of the French Revolution have left so deep an impression on the collective memory of the French people as Georges-Jacques Danton. Almost 200 years after his death, Danton still evokes instinctive sympathy —for his energy, enthusiasm, and audacity. He is one of the most complex and controversial statesmen of the Revolutionary period.

A powerful orator who initially championed the extreme left and contributed to the fall of the monarchy, he was executed once his insistent calls for moderation seemed to threaten the extreme policies of the Revolutionary government headed by Maximilien de Robespierre. A compassionate defender of the rights of the oppressed, he was also an opportunist who did not scruple to enrich himself by selling the power of his office.

The tribune of the people (1789–92). Danton was born at Arcis-sur-Aube, in Champagne, on October 26, 1759, the son of Jacques Danton, an attorney, and his second wife, Marie-Madeleine Camus. After attending school in Champagne, Danton was from 1773 educated by the Oratorians at Troyes. After obtaining his law degree in 1784 at Reims, he went to Paris to practice and in 1787 bought the office of advocate in the *Conseil du roi* (council with legislative and judicial functions). He then married Antoinette Charpentier.



Danton, portrait by Mme Constance-Marie Charpentier (1767–1849). In the Musée Carnavalet, Paris.
J.E. Bulloz

At the outbreak of the Revolution in July 1789, Danton enrolled in the *garde bourgeoise* (civic guard) of the Cordeliers district and was elected president of the district in October. In the spring of 1790, with some militants from his district, he founded the popular association that was to become famous as the Club of the Cordeliers. So far, however, Danton's fame had been merely local. Elected a member of the provisional Paris Commune (city council) in January 1790, he was excluded from the council in its final form in September. Although elected administrator of the *département* of Paris in January 1791, he actually exercised no influence on that body.

Founder of the Club of the Cordeliers

Meanwhile, however, Danton shone at the Club of the Cordeliers and at another political association, the Club of the Jacobins, before both of which he frequently made speeches during 1791. During the crisis following Louis XVI's attempt to leave the country in June, he became increasingly prominent in the Revolutionary movement. His signature, however, does not appear on the famous petition of the Cordeliers demanding the abdication of Louis XVI, which, on July 17, resulted in the massacre of some of the petitioners by the national guard. During the repression following these events, Danton took refuge in London.

He returned to Paris to take part in the elections to the Legislative Assembly as elector for the Théâtre Français section, and in December 1791 he was elected second assistant to the *procureur* (public prosecutor) of the Paris Commune.

During the national crisis in the spring of 1792 (war was declared on Austria and Prussia on April 20), Danton resumed his role of tribune of the people. On June 18 he attacked the Marquis de Lafayette, an adviser of the King and a general, for using his position to play politics. Yet he took no part in the demonstrations before the royal palace of the Tuileries on June 20. Although his part in the overthrow of the monarchy by the insurrection of August 10, 1792, remains obscure, he was largely credited with its success.

The overthrow of the monarchy

Speaking before the Revolutionary Tribunal, Danton boasted that he had "been responsible for" the events of August 10; that insurrection, however, was not the result of the efforts of Danton or any other man but, rather, the collective act of obscure militants from all over the city. However small a part he played in removing the King, he was elected minister of justice by the Legislative Assembly. Though not officially its president, in the new government Danton dominated his colleagues by his

strength of character, the aura of his Revolutionary past, and his ability to make swift decisions. When the news arrived that Longwy had been taken by the invading allies on August 25, 1792, and Jean-Marie Roland, minister of the interior, proposed that the government should move from Paris to Blois, Danton objected vigorously. The proclamation he then caused the Executive Council to adopt bears his stamp: it was a summons to battle. On the morning of September 2, when it was learned in Paris that Verdun was besieged and while the populace broke into the prisons to search for suspects and traitors, Danton, in the Legislative Assembly, delivered the most famous of his speeches: "To conquer the enemies of the fatherland, we need daring, more daring, daring now and always, and France is saved!"

While Danton was delivering this speech, the prison massacres began for which the Girondins, the moderate wing of the Revolution, charged Danton with responsibility. There is no proof, however, that the massacres were organized by him or by anyone else, though it is certain that he did nothing to stop them. Just as in the case of the August insurrection, the September massacre was not the act of one man but of the people of Paris.

The massacres of September 1792

On September 6 Danton was elected deputy for Paris to the National Convention. He immediately made every effort to end all the disputes between the Revolutionary parties, but his policy of conciliation was thwarted by the Gironde, which demanded that he render an accounting when he left his post as minister of justice. Danton could not justify 200,000 livres of secret expenditures. He emerged from this conflict embittered and with his political prestige diminished.

Sent on a mission to Belgium, Danton took no part in the opening of Louis XVI's trial in the Convention. He was present, however, on January 15, 1793, and voted for death without reprieve. Although absent from the trial, Danton had played a part in it since the autumn of 1792. According to the *Mémoires* of Théodore, comte de Lamath, a former Revolutionary, Danton wanted to spare the King. It seems that having failed, despite strenuous efforts, to gain the support of the Girondins, Danton plotted with Gen. Charles-François du Périer Dumouriez to obtain the intervention of the English government by bribery. Only when the plan miscarried did he vote for the death of the King.

Danton remained in the mainstream of the Revolution, not without often engaging in intrigue. His dealings with Dumouriez, who commanded the army of Belgium, have never been clarified. After the defeat of Neerwinden (March 18, 1793), when Dumouriez went over to the Austrians, the Gironde accused Danton of complicity with the General. Boldly turning the tables, Danton made the same accusation against the Girondins. The break was irreparable.

Danton's Committee of Public Safety. On April 7, 1793, Danton became a member of the first Committee of Public Safety, which, created the previous day, became the executive organ of the Revolutionary government. For three months Danton was effectively the head of the government, charged especially with the conduct of foreign affairs and military matters. During this second period in the government he pursued a policy of compromise and negotiation. He tried in every direction to enter into diplomatic conversations with the enemy. No doubt he could in all honesty think it useful to negotiate in an attempt to dissolve the allied coalition or even to obtain a general peace. By the spring of 1793, however, a policy of negotiation was no longer conceivable: it was useless to try to disarm the enemy by concessions when he was victorious. On July 10, when the Committee of Public Safety's term expired, the Convention elected a new committee without Danton.

Leader of the moderate opposition. From that time Danton's political conduct became more complex. On various occasions he supported the policy of the Committee of Public Safety though at the same time refusing to play a part in it—which would have stabilized the political situation. Danton still reappeared from time to time as the tribune of the people, voicing the demands of the masses. He quickly showed, however, that he sought to stabilize the Revolutionary movement; very soon—whether he wanted it or not—he appeared as the leader of the Indulgents—the moderate faction that had risen out of the Cordeliers.

During the great Parisian popular demonstrations of September 4 and 5, 1793, Danton spoke eloquently in favour of all the popular demands. Yet at the same time he tried to set bounds to the movement and keep it under control. He demanded, for instance, that the meetings of the hitherto permanent sectional assemblies be reduced to two per week.

Danton's moderate position became more marked in the autumn of 1793. He did not, however, intervene personally but left it to his friends to criticize the policy of the government. His disapproval of the terrorist repression had become so strong that he withdrew from political life, alleging reasons of health or of family. Of the Girondins, he is reported to have said to a friend at the beginning of October 1793, "I shall not be able to save them," and to have burst into tears. On October 12 he obtained leave from the Convention and left for his native town. He returned on November 21, although the reasons for his return remain ambiguous.

Disapproval of terror

Danton at once resumed political activity. He vigorously supported the Committee of Public Safety against excesses of the anti-Christian movement and later opposed the abolition of the salaries of constitutional priests and hence the separation of church and state. Danton's support of the governmental policy of stabilization was doubtless not without ulterior motives, both personal and political; he was determined to save friends of his who had been arrested or who were in danger of arrest. But he also wanted to slow the Revolutionary drive of the government. The Dantonist policy was opposed in all points to the program of popular extremism supported by Jacques Hébert and his Cordeliers friends: extreme terror, war to the hilt.

Danton defined his moderate political line on December 1, 1793, when he informed the Revolutionary radicals that their role was ended. From then on, whether he wanted it or not, he was looked upon as the leader of the moderate opposition. At the beginning of 1794, Danton and his friends took an even more critical attitude, the Revolutionary journalist Camille Desmoulins, of *Le Vieux Cordelier*, serving as their spokesman. They were challenging not only the system of the terror of Robespierre but the whole policy of the Revolutionary government, while awakening the hopes of the opponents of the regime.

Once the government realized it could not allow itself to be overwhelmed from the right, however, the tide turned abruptly. When Fabre d'Églantine, the dramatist and zealous Revolutionary, compromised in the affair of the Compagnie des Indes, was arrested in January 1794, Danton tried to defend him obliquely by demanding that the arrested deputies should be judged before the people. "Woe unto him who sat beside Fabre and who is still his dupe!" cried a deputy, clearly threatening Danton himself.

The incident signalled more than the defeat of the offensive of the Indulgents, for, already compromised, they were themselves soon threatened by the counter-offensive of their adversaries, Hébert's ultra-left faction, the Exagérés, or Enragés. When the crisis, however, became more acute and the Exagéré opposition hardened its position, the government lost its patience: in March 1794, Hébert and the principal Cordeliers leaders were arrested. Sentenced to death, they were executed on March 24. The Indulgents, believing that their hour had come, increased their pressure. The government, however, had no intention of letting itself be overwhelmed by the moderate opposition of the right. Warned several times of the threats that hung over him, Danton remained unafraid: "They will not dare!" Finally, during the night of March 29–30, 1794, he and his friends were arrested. Before the Revolutionary tribunal, Danton boldly spoke his mind. To silence him, the Convention decreed that a suspect on trial who insulted national justice be excluded from the

Trial of Danton

debate. "I will no longer defend myself," Danton cried. "Let me be led to death, I shall go to sleep in glory." Danton was guillotined with his friends on April 5, 1794. "Show my head to the people," he said to the executioner. "It is worth the trouble."

Assessment. Denigrated during the first half of the 19th century, Danton was rehabilitated under the Second Empire and enshrined as a hero under the Third Republic. A chief controversy about him is the problem of his wealth and hence of his venality. To his contemporaries, Danton's venality was obvious, even though, for lack of documentation, it was not proved during his lifetime. Scholarly opinion in the 20th century was originally divided between two views: that Danton was a sincere democrat and sound patriot or that he was an unscrupulous politician capable of betraying the Revolution and selling himself to the court. It is now generally accepted that Danton was used as an informer by the court and that in return he received payments from the funds of the Civil List. At the same time, however, his attachment to the nation and to the Revolutionary cause is beyond doubt. The historian Georges Lefebvre, in stressing Danton's complexity, has tried to account for his behaviour:

> Interested flexibility, wily prudence, venality . . . ; but sometimes also the genuine realism of a statesman; then indomitable furies, careless neglects, sudden renunciations, due to a violent temperament which no moral or intellectual discipline ever tried to master; . . . too avid for pleasure to tolerate the somber reflections of suspicion and hatred.

Danton was a leader of men. More than any other Revolutionary leader, he could enter into communion with the sansculottes—the Revolutionary have-nots—to share their passions. He pleased the people by his generosity, his indulgence, his verve. All these were characteristics that won him the sympathy of the people and that, during the crisis of the summer of 1792, enabled him to serve the Revolution well.

BIBLIOGRAPHY. There is no reliable critical edition of Danton's addresses and speeches and no up-to-date biography. The best collection is ANDRE FRIBOURG, *Discours de Danton* (1910). For its historical interest the edition of the *Oeuvres de Danton* by AUGUSTE VERMOREL (1866) may be mentioned. The first important works on Danton appeared in the mid-19th century. The earliest well-documented study is ALFRED BOUGEART, *Danton: Documents authentiques pour servir à l'histoire de la Révolution française* (1865). Many more studies of Danton appeared under the Third Republic. Of special interest are JEAN ROBINET, *Danton, homme d'état* (1889); and the numerous writings of A. AULARD, favourable to Danton, published in the periodical *La Révolution Française* (see, for example, the year 1893). The most recent apologetic biography is JEAN LOUIS BARTHOU, *Danton* (1932), in French. ALBERT MATHIEZ answered Aulard's articles with studies critical of Danton in his journal, *Les Annales révolutionnaires* (such as, "La Fortune de Danton," 1912, and "Les Comptes de Danton," 1913). Mathiez's attacks centred on Danton's corruption; he collected his basic arguments in two volumes: *Danton et la paix* (1919) and *Autour de Danton* (1926). Today most historians agree on a middle position as adopted by LOUIS MADELIN in *Danton* (1914), in French; and in the excellent biography by HERMANN WENDEL, *Danton* (1930, in German; French trans., 1932). See also GEORGES LEFEBVRE, "Sur Danton," in *Annales historiques de la Révolution française* (1932), reprinted in *Études sur la Révolution française*, 2nd ed. (1963).

(Al.S.)

Danube River

The Danube is the second largest river of Europe after the Volga. It rises in the Black Forest mountains of West Germany and flows for approximately 1,770 miles (2,850 kilometres) to its mouth on the Black Sea. Along its course, it passes through eight countries under six variations of its name. In West Germany and Austria it is known as the Donau, in Czechoslovakia as the Dunaj, in Hungary as the Duna, in Yugoslavia and Bulgaria as the Dunav, in Romania as the Dunărea, and in the Soviet Union as the Dunay.

The Danube played a vital role in the settlement and political evolution of central and southeastern Europe. Its banks, lined with castles and fortresses, formed the boundary between great empires, and its waters served as a vital commercial highway between nations. The river's majesty has long been celebrated in music. The famous piece *An der schönen, blauen Donau* (*The Blue Danube*), by Johann Strauss the Younger, became the symbol of imperial Vienna. In the 20th century, the river has continued its role as an important trade artery. Its upper courses have been harnessed for their hydroelectric power, and the cities along its banks—including the national capitals of Vienna, Budapest, and Belgrade—depended upon it for their economic growth. Unfortunately, the Danube has been used as a disposal for urban, industrial, and agricultural waste and has joined the ranks of other rivers polluted by man.

History. During the 7th century BC, Greek sailors reached the lower Danube and sailed upstream, conducting a brisk trade. They were familiar with the whole of the river's lower course and named it the Ister. The Danube later served as the northern boundary of the vast Roman Empire and was called the Danuvius. A Roman fleet patrolled its waters, and the strongholds along its shores were the centres of settlements, among them Vindobona (later Vienna), Aquincum (later Budapest) Singidunum (later Belgrade), and Sexantaprista (later Ruse).

Ancient settlement

During the Middle Ages the old fortresses continued to play an important role, and new castles such as Werfenstein, built by Charlemagne, the 9th century Holy Roman emperor, were erected. When the Ottoman Empire spread from southeastern Europe to central Europe in the 15th century, the Turks relied upon the string of fortresses along the river for defense. The Habsburg dynasty recognized the navigational potential of the Danube. Maria Theresa, queen of Hungary and Bohemia from 1740 to 1780, founded a Department for River Navigation, and in 1830 a riverboat made a first trip from Vienna to Budapest, possibly for trading purposes. This trip marked the end of the river's importance as a line of defense and the beginning of its use as a channel of trade.

Regulated navigation on the Danube has been the subject of a number of international agreements. In 1616 an Austro-Turkish treaty was signed in Belgrade under which the Austrians were granted the right to navigate the Middle and Lower Danube. In 1774, under the Treaty of Küçük Kaynarca, Russia was allowed to use the Lower Danube. The Anglo-Austrian and the Russo-Austrian conventions of 1838 and 1840 promoted free navigation along the entire river, a principle that was more precisely formulated in the Paris Treaty of 1856, which also set up the first Danubian Commission with the aim of supervising the river as an international waterway. In 1921 and 1923 final approval of the Danube River Statute was granted by Austria, Germany, Yugoslavia, Bulgaria, Romania, Great Britain, Italy, Belgium, Czechoslovakia, Hungary, and Greece. The international Danube Commission was thus finally established as an authoritative institution with wide powers, including its own flag, the right to levy taxes, and diplomatic immunity for its members. It controlled navigation from the town of Ulm to the Black Sea and kept navigational equipment in good repair.

International control

During World War II free international navigation was interrupted and a consensus concerning its resumption was not reached until the Danubian Convention of 1948. Based on Soviet proposals, it allowed only the Danubian countries to participate in the new Danube Commission. Britain, France, and the United States refused to sign the convention.

The river's course. The Danube's vast drainage of 315,000 square miles (816,000 square kilometres) includes a variety of natural conditions that affect the origins and the regimes of its watercourses. They favour the formation of a branching, dense, deepwater river network that includes some 300 tributaries, 34 of which are navigable. The river basin expands unevenly along its length. It covers about 18,000 square miles at the Inn confluence, 81,000 square miles after joining with the Drava, and 228,000 square miles below the confluences of its most affluent tributaries, the Sava and the Tisza. In the lower course the basin's rate of growth decreases; at

the estuary the drainage area includes 315,000 square miles. About 56 percent of the entire Danube Basin is drained by its right-bank tributaries, which collect their waters from the Alps and other mountain areas and contribute up to 66 percent of the total river runoff or outfall.

Three sections are discernible in the river's basin. The upper course stretches from its source to the gorge in the Austrian Alps and the Western Carpathian Mountains called the Hungarian Gates. The middle course runs from the Hungarian Gates to the Iron Gate Gorge in the Southern Romanian Carpathians. The lower course flows from the Iron Gate to the delta-like estuary at the Black Sea.

The upper course. The Upper Danube springs as two small streams—the Breg and Brigach—from the eastern slopes of the Black Forest (Schwarzwald) mountains of West Germany, which partially consist of limestone. The headstreams lose some of their water because it leaks through the rock to the adjoining Rhine Basin. From Donaueschingen, where the headstreams unite, the Danube flows northeastward in a narrow, rocky bed. To the left of the valley rise the wooded slopes of the Swabian and the Franconian mountains. Between Ingolstadt and Regensburg the river cuts across them and forms a scenic canyon-like valley. To the right of the river course stretches the large Bavarian Plateau, covered with thick layers of river deposits from the numerous Alpine tributaries. The bank is low and uniform, composed mainly of fields, peat, and marshland.

The river's northernmost point

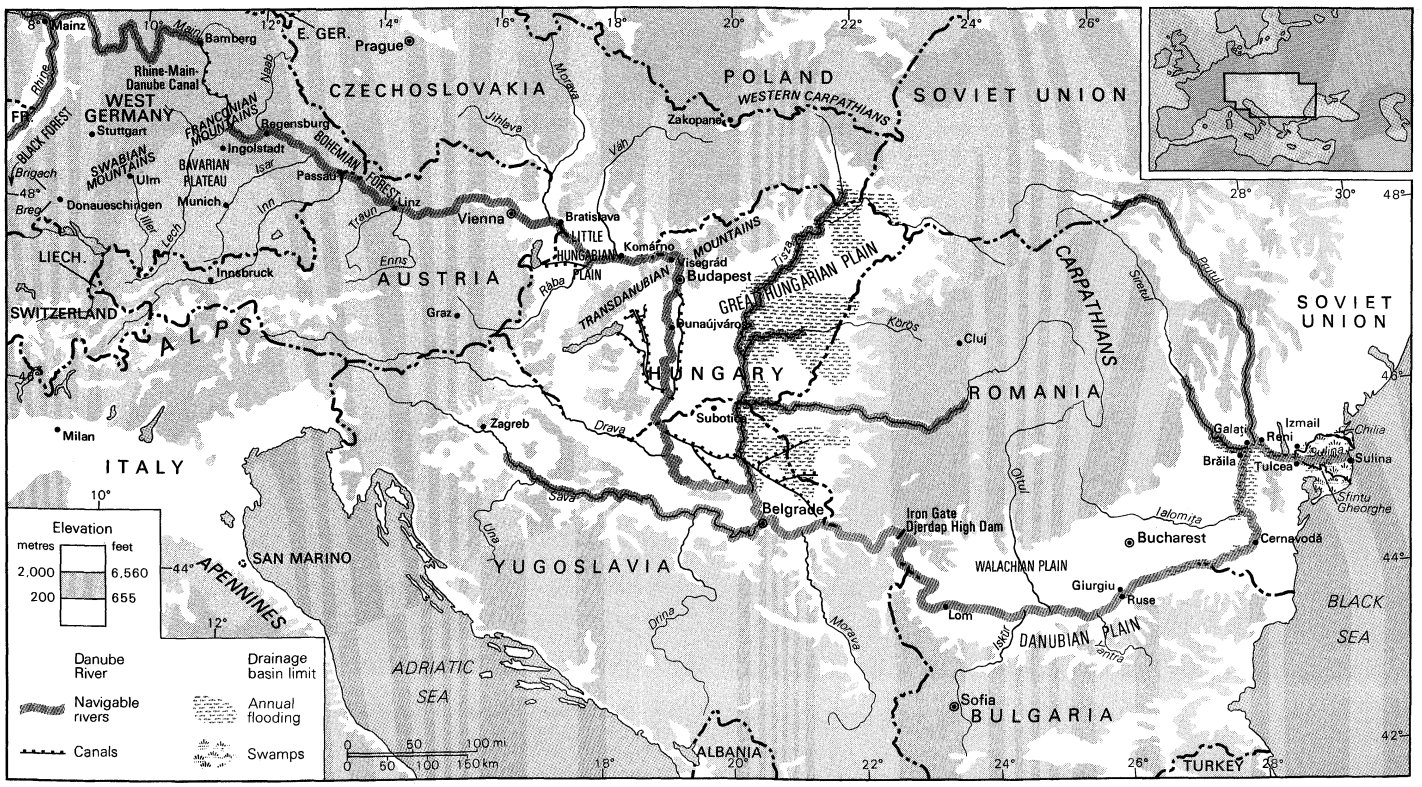
At Regensburg the Danube reaches its northernmost point. From there the river veers south and crosses wide, fertile, and level country. Shortly before it reaches Passau on the Austrian border, the river narrows and its bottom abounds with reefs and shoals. The Danube then flows through Austrian territory, where it cuts into the slopes of the Bohemian Forest (Böhmer Wald) and forms a narrow valley. In order to improve navigation, dams and protecting dikes have been built near Passau, Linz, and Ardagger. The Upper Danube, some 600 miles long, has a considerable average inclination of the riverbed (0.93 percent) and a rapid current of 2.2 to five miles per hour. Depths vary from 3 to 26 feet. The Danube swells substantially at Passau where the Inn River, its tributary, carries more water than the main river. Other major tributaries in the Upper Danube course include the Iller, Lech, Isar, Traun, Enns, and Morava rivers.

The middle course. In its middle course the Danube looks more like a flatland river, with low banks and a bed that reaches a width of over one mile. Only in two sectors—at Visegrád (Hungary) and the Iron Gate—does the river flow through narrow, canyon-like gorges. The basin of the Middle Danube exhibits two main features—the flatland of the Small and the Great Hungarian plains (Kis Alföld and Nagy Alföld), and the low peaks of the Western Carpathians and the Transdanubian mountains.

The Danube enters the Small Hungarian Plain immediately after emerging from the Hungarian Gates Gorge near Bratislava, Czechoslovakia. There the river stream slows down abruptly and loses its transporting capacity, so that gravel and sand settle on the bottom. Between Bratislava and Komárno 785,000 cubic yards of silt are deposited annually. A principal result of this deposition has been the formation of Great and Little Schütt (Vel'ký Žitný Ostrov and Szigetköz) Islands, an area of 734 square miles that supports 140,000 inhabitants in 110 settlements. The silting hampers navigation and occasionally divides the river into two or more channels. East of Komárno the Danube enters the Visegrád Gorge, where it is squeezed between the foothills of the Western Carpathian and the Hungarian Transdanubian mountains. The steep slopes of the right bank are crowned with fortresses, castles and cathedrals of the Hungarian Árpád dynasty of the 10th to the 15th centuries.

The Danube then flows past Budapest and across the vast Hungarian Plain until it reaches the Iron Gate Gorge. The riverbed is shallow and marshy, and low terraces stretch along both banks. River accumulation has built a large number of islands, including those of Margitsziget and Csepelsziget near Budapest. In this long stretch, the river takes on the waters of its major tributaries—the Drava, the Tisza, and the Sava—which create substantial changes in the river's regime. The average runoff increases from 3,074 to 7,639 cubic yards per second at the Iron Gate. The river valley looks most imposing here. Depths vary between five and 177 feet and the

The Iron Gate



The Danube River Basin and its drainage network.

current velocity ranges between 2.2 and 11.2 miles per hour. The rapids and reefs made the river unnavigable until a lateral navigation channel and a parallel railway allowed rivercraft to be towed upstream against the strong current.

The lower course. Beyond the Iron Gate, the Lower Danube flows across a wide plain. To the right, above steep banks, stretches the tableland of the Danubian Plain of Bulgaria. To the left lies the low Walachian Plain, which is separated from the main stream by a strip of lakes and swamps. The tributaries in this section are comparatively small and account for an increase in the total runoff of only 1,086 yards per second. They include the Oltul, the Siretul, and the Prutul. The width of the riverbed varies between 132 and 3,960 feet, the depth ranges between five and 30 feet, and the current speed is between one and 2.5 miles per hour.

The river is again obstructed by a number of islands. At Cernavodă, the Danube heads northward until it reaches Galaţi, where it veers abruptly eastward. Near Tulcea, some 50 miles from the sea, the river begins to spread out into its delta.

The delta. The river splits into three channels—the Chilia, which carries 67 percent of the total runoff; the Sulina, which accounts for 9 percent; and the Sfîntu Gheorghe (St. George), which carries 24 percent. Navigation is possible only along the 39-mile-long Sulina Channel, which has been straightened and dredged to a depth of 23 feet. Between the channels, a maze of smaller creeks and lakes are separated by oblong strips of land called *grinduri.* Most *grinduri* are arable and cultivated, and some are overgrown with tall oak forests. A large quantity of reeds that grow in the shallow water tracts are used in the manufacture of paper and textile fibres. The Danube Delta (Delta Dunării) covers an area of 1,660 square miles. It is a comparatively young formation. About 6,500 years ago the delta site was a shallow cove of the Black Sea coast. It was gradually filled as the river deposited 80,000,000 tons of silt a year. The delta continues to grow seaward at the rate of 80 to 100 feet annually.

The river regime. The different physical features of the river basin affect the amount of water runoff in its three sections. In the Upper Danube, the runoff corresponds to that of the Alpine tributaries, where the maximum comes in June when melting of snow and ice in the Alps is the most intensive. Runoff drops to its lowest point during the winter months.

In the middle basin, the phases last up to four months, with two runoff peaks in June and April. The June peak stems from that of the upper course, reaching its maximum ten to 15 days later. The April peak is local. It is caused by the addition of waters from the melting snow in the plains and from the early spring rains of the lowland and the low mountains of the area. Rainfall is important; the period of low water begins in October and reflects the dry spells of summer and autumn that are characteristic of the low plains. In the lower basin, all Alpine traits disappear completely from the river regime. The runoff maximum occurs in April, and the low point extends to September and October.

The river carries considerable quantities of solid particles, some 95 percent of which are quartz grains. The constant shift of deposits in different parts of the riverbed forms shoals. In the stretches between Bratislava and Komárno and in the Sulina Channel (Braţul Sulina), draglines are constantly at work to maintain the depth needed for navigation.

The temperature of the river waters depends on the climate of the various parts of the basin. In the upper course, where the summer waters derive from the Alpine snow and glaciers, the water temperature is low. In the middle and lower reaches, summer temperatures vary between 71° and 75° F (22° and 24° C) while winter temperatures near the banks and on the surface drop below freezing. Upstream from Linz the Danube never freezes entirely because the current is turbulent. The middle and lower courses, however, become icebound during severe winters. Between December and March, periods of ice drift combine with the spring thaw, causing floating ice blocks to accumulate at the river islands, jamming the river's course and often creating major floods.

The winter freeze

The natural regime of river runoff changes constantly as a result of the introduction of stream-regulating equipment, including dams and dikes. The chemical content also changes as city sewerage, fertilizers, and industrial wastes are dumped into the river.

Pollution. The mineral content of the river is greater during the winter than the summer. The content of organic matter is relatively low, but pollution increases as the waters flow past industrial areas.

Economic development. *Navigation.* In the early 1970s, eight countries operated trading ships upon the river. Listed in accordance with the amount of tonnage, in decreasing order, they were the Soviet Union, 727,000 tons; Yugoslavia, 619,000 tons; Romania, 397,000 tons; Hungary, 267,000 tons; Austria, 223,000 tons; Czechoslovakia, 198,000 tons; Bulgaria, 195,000 tons; and West Germany, 74,000 tons. The major river ports include Izmail and Reni, Soviet Union; Galaţi, Brăila, and Giurgiu, Romania; Ruse and Lom, Bulgaria; Belgrade, Yugoslavia; Dunaújváros and Budapest, Hungary; Komárno and Bratislava, Czechoslovakia; Vienna and Linz, Austria; and Regensburg, West Germany.

Power generation. By the early 1970s the Danube had been tapped for power in a rudimentary fashion, mainly in its upper course. The process, however, is spreading downstream. The largest hydroelectric project—the Djerdap High Dam and the Iron Gate power station—was built jointly by Yugoslavia and Romania. The project has a 1,052,000-kilowatt capacity and an annual output of 11,308,000,000 kilowatt-hours. Czechoslovakia and Hungary are jointly planning to dam the river at the Visegrád Gorge and harness its power, while Romania and Bulgaria plan to build a barrage on the lower reaches. Irrigated areas along the river in Czechoslovakia, Hungary, Yugoslavia, and Bulgaria continue to expand; much industrial use of Danube waters is made at Vienna, Budapest, Belgrade, and Ruse.

Dams and irrigation

Fishing. Until the late 1960s the Danube was a rich fishing ground. The most important fish varieties included the sturgeon, wels, pike, beluga, and Black Sea herring. Because of industrial pollution, however, animal life in the Danube has diminished and the fish have moved to side lakes and swamps.

The future. The Danube is to become part of an all-European navigational system. Originally conceived in 1921, the Rhine–Main–Danube Canal is to be commissioned in 1989. From the town of Mainz, it already follows the Main River upstream to Bamberg, where it will join a reconstructed section of the old Ludwigs Canal to Regensburg. From there to the West German–Austrian border, the Danube sector stretches for 130 miles. The total length of the canal will be 423 miles. The difference in heights between the Rhine level and the highest point on the way—1,080 feet—will be overcome by means of 47 sluice gates. Another wing of the canal to Passau will overcome a height of 35 feet through 13 gates.

A canal linking the Danube, Oder, and Elbe rivers was also being planned in the 1970s. Beginning at Bratislava, it will link up with the east–west navigable waterway of the Mittelland–Vistula–Bug–Dnepr rivers and will push further north to bring Danubian freight to the North and Baltic seas.

BIBLIOGRAPHY. Though over 50 years old, the ADMIRALTY NAVAL STAFF, NAVAL INTELLIGENCE DIVISION, *A Handbook of the River Danube,* 2 pt. (1915–19), is detailed on geographical and hydrological features, even including a section on diseases and itineraries. STEPHEN GOROVE, *Law and Politics of the Danube* (1964), examines the river's international importance, also treated in *La Commission européenne du Danube et son oeuvre de 1856 à 1931* (1931). A useful article is GUSTAV HOLZMANN, "Die Zukunft der Donauschiffahrt," *Z. Wirt Geogr.,* 8:18–25 (1964); while the river's importance at specific stages of its economic development is analyzed by ANTONIN BASCH in *The Danube Basin and the German Economic Sphere* (1944). EMIL LENGYEL, *The Danube* (1940), is largely historical. A good travel account is LOVETT F. EDWARDS, *Danube Stream* (1940); JOHN LEHMANN, *Down River: A*

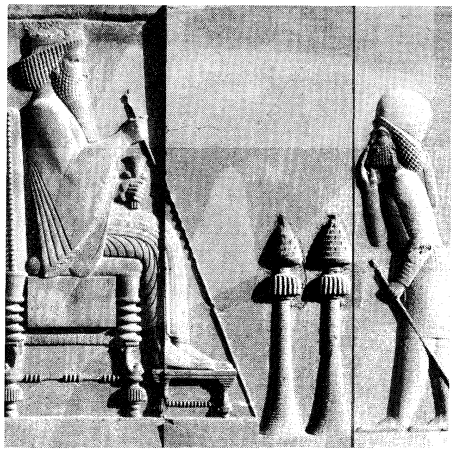
Danubian Study (1939); JOHN R. COLVILLE, *Fools' Pleasure: A Leisurely Journey down the Danube to the Black Sea, the Greek Islands and Dalmatia* (1935); and J.E.P. HUGHES, *Through Many Lands by Water: 1780 Miles down the Danube* (1930), are useful. An older, straightforward guide to places of interest along the river is *The Danube from Passau to the Black Sea*, a trans. from the German by MAY O'CALLAGHAN (1913); there are two portraits of the Danube in a more leisurely age by JOHN P. SIMPSON, *Letters from the Danube*, 2 vol. (1847); and ROBERT SNOW, *Journal of a Steam Voyage down the Danube to Constantinople* (1842). *The Danube Delta*, issued by the Foreign Languages Publishing House in Bucharest (1956), is a book of photographs.

(P.G.P.)

Darius I the Great, of Persia

One of the greatest of the Achaemenid kings of Persia (Iran), Darius I unified the administration of the Achaemenid Empire and began the Persian attempt to penetrate Europe. Darius, who reigned from 522 to 486 BC, was the son of Hystaspes, the satrap (provincial governor) of Parthia. The principal contemporary sources for his history are his own inscriptions, especially the great trilingual inscription on the Bīsitūn (Behistun) rock at the village of the same name, in which he tells how he gained the throne. The accounts of his accession given by the Greek historians Herodotus and Ctesias are in many points obviously derived from this official version but are interwoven with legends; *e.g.*, that Darius and his fellow conspirators left the question as to which of them should become king to the decision of their horses and that Darius won the crown by a trick of his groom.

By courtesy of the Oriental Institute, the University of Chicago



Darius I, seated before two incense burners, giving audience to a dignitary; detail of a bas-relief of the north courtyard in the Treasury at Persepolis, late 6th–early 5th century BC. In the Archaeological Museum, Tehrān.

Ascension to monarchy

According to Herodotus, Darius, when a youth, was suspected by Cyrus II the Great (who ruled from 559 to 529 BC) of plotting against the throne. Later, he was in Egypt with Cambyses II, the son of Cyrus and heir to his kingdom, as a member of the royal bodyguard. After the death of Cambyses in the summer of 522 BC, Darius hastened to Media, where, in September, with the help of six Persian nobles, he killed Bardiya (Smerdis), another son of Cyrus, who had usurped the throne the previous March. In the Bīsitūn inscription Darius defended this deed and his own assumption of kingship on the grounds that the usurper was actually Gaumata, a Magian, who had impersonated Bardiya after Bardiya had been murdered secretly by Cambyses. Darius therefore claimed that he was restoring the kingship to the rightful Achaemenid house. He himself, however, belonged to a collateral branch of the royal family, and as his father and grandfather were alive at his accession, it is unlikely that he was next in line to the throne. Some modern scholars consider that he invented the story of Gaumata in order to justify his actions and that the murdered king was indeed the son of Cyrus.

Darius did not at first gain general recognition but had to impose his rule by force. His assassination of Bardiya was followed, particularly in the eastern provinces, by widespread revolts, which threatened to disrupt the empire. In Susiana, Babylonia, Media, Sagartia, and Margiana, independent governments were set up, most of them by men who claimed to belong to the former ruling families. Babylonia rebelled twice and Susiana three times. In Persia itself a certain Vahyazdata, who pretended to be Bardiya, gained considerable support. These risings, however, were spontaneous and uncoordinated, and, notwithstanding the small size of his army, Darius and his generals were able to suppress them one by one. In the Bīsitūn inscription he records that in 19 battles he defeated nine rebel leaders, who appear as his captives on the accompanying relief. By 519 BC, when the third rising in Susiana was put down, he had established his authority in the east. In 518 Darius visited Egypt, which he lists as a rebel country, perhaps because of the insubordination of its satrap, Aryandes, whom he put to death. He also removed Oroetes, the disloyal satrap of Sardis.

Fortification of the empire

Having restored internal order in the empire, Darius undertook a number of campaigns for the purpose of strengthening his frontiers and checking the incursions of nomadic tribes. In 519 BC he attacked the Scythians east of the Caspian Sea and a few years later conquered the Indus Valley. In 513, after subduing eastern Thrace and the Getae, he crossed the Danube into European Scythia, but the Scythian nomads devastated the country as they retreated from him, and he was forced, for lack of supplies, to abandon the campaign. The satraps of Asia Minor completed the subjugation of Thrace, secured the submission of Macedonia, and captured the Aegean islands of Lemnos and Imbros. Thus, the approaches to Greece were in Persian hands, as was control of the Black Sea grain trade through the straits, of major importance to the Greek economy. The conquest of Greece was a logical step to protect Persian rule over the Asiatic Greeks from interference by their European kinsmen. According to Herodotus, Darius, before the Scythian campaign, had sent ships to explore the Greek coasts, but he took no military action until 499 BC, when Athens and Eretria supported an Ionian revolt against Persian rule. After the suppression of this rebellion, Mardonius, Darius' son-in-law, was given charge of an expedition against Athens and Eretria, but the loss of his fleet in a storm off Mt. Athos (492 BC) forced him to abandon the operation. In 490 BC another force under Datis, a Mede, destroyed Eretria and enslaved its inhabitants but was defeated by the Athenians at Marathon. Preparations for a third expedition were delayed by an insurrection in Egypt, and Darius died in 486 BC before they were completed.

Darius as an administrator

Although Darius consolidated and added to the conquests of his predecessors, it was as an administrator that he made his greatest contribution to Persian history. He completed the organization of the empire into satrapies, initiated by Cyrus the Great, and fixed the annual tribute due from each province. During his reign, ambitious and far-sighted projects were undertaken to promote imperial trade and commerce. Coinage, weights, and measures were standardized and land and sea routes developed. An expedition led by Scylax of Caryanda sailed down the Indus River and explored the sea route from its mouth to Egypt; and a canal from the Nile to the Red Sea, probably begun by the chief of the Egyptian delta lords, Necho I (7th century BC), was repaired and completed.

While measures were thus taken to unite the diverse peoples of the empire by a uniform administration, Darius followed the example of Cyrus in respecting native religious institutions. In Egypt he assumed an Egyptian titulary and gave active support to the cult. He built a temple to the god Amon in the Kharga oasis, endowed the temple at Edfu, and carried out restoration work in other sanctuaries. He empowered the Egyptian Udjahorresne, who had served under Cambyses, to re-establish the medical school of the temple of Sais, and he ordered

his satrap to codify the Egyptian laws in consultation with the native priests. In the Egyptian traditions he was considered as one of the great lawgivers and benefactors of the country. In 519 BC he authorized the Jews to rebuild the temple at Jerusalem, in accordance with the earlier decree of Cyrus. He also continued the privileges granted by Cyrus to the Greek sanctuaries: a Persian official was reprimanded for exacting tribute and forced labour from the priests of the temple of Apollo near Magnesia on the Maeander River. In the opinion of some authorities, the religious beliefs of Darius himself, as reflected in his inscriptions, show the influence of the teachings of Zoroaster, and the introduction of Zoroastrianism as the state religion of Persia is probably to be attributed to him.

Darius was the greatest royal architect of his dynasty, and during his reign Persian architecture assumed a style that remained unchanged until the end of the empire. In 521 BC he made Susa his administrative capital, where he restored the fortifications, built an audience hall (apadana), and a residential palace. The foundation inscriptions of his palace describe how he brought materials and craftsmen for the work from all quarters of the empire. At Persepolis, in his native country of Fars (Persis), he founded a new royal residence to replace the earlier capital at Pasargadae. The fortifications, apadana, council hall, treasury, and a residential palace are to be attributed to him, although not completed in his lifetime. He also built at Ecbatana and Babylon.

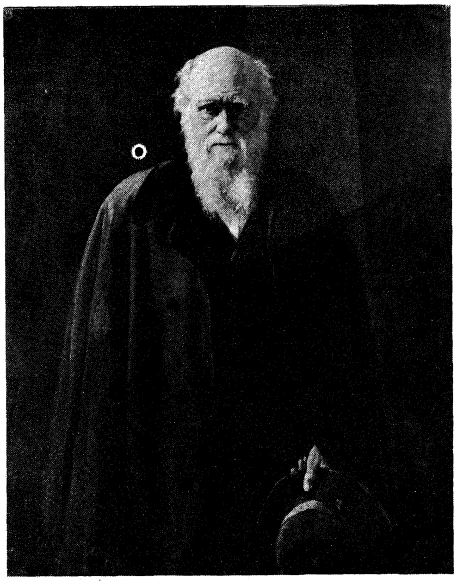
BIBLIOGRAPHY. ARTHUR T. OLMSTEAD, *History of the Persian Empire* (1948), an authoritative and detailed history of the Achaemenid period; DONALD N. WILBER, *Persepolis: The Archaeology of Parsa, Seat of the Persian Kings* (1969), a well-illustrated account of the history and monuments of Persepolis.

(J.M.M.-R.)

Darwin, Charles

Charles Robert Darwin, the English naturalist whose demonstration of organic evolution and its operating principle, natural selection, revolutionized human knowledge, was born on Feb. 12, 1809, at Shrewsbury. He was the son of Robert Waring Darwin and Susannah, daughter of Josiah Wedgwood I, the famous potter. His grandfather Erasmus Darwin was the polymath physician, poet, philosopher, and inventor who, by his second marriage, was also the grandfather of Francis Galton, founder of the science of eugenics.

By courtesy of the National Portrait Gallery, London



Darwin, oil painting by John Collier (1850–1934). In the National Portrait Gallery, London.

Youth and education. Darwin's mother died when he was eight years old, and he was brought up by his eldest sister, Caroline, to whom he was always grateful for instilling in him a humanitarian spirit. The boy developed very slowly: he was given, when small, to inventing gratuitous fibs and to daydreaming; and he was passionately fond of collecting seals, franks (equivalents of postage stamps), pebbles, and minerals—an important trait in his future as a naturalist.

In 1818 he entered Shrewsbury School. Later he complained that he was taught only classics, never realizing his debt to them for providing a sound basis for his education. He was a poor student, and in 1825 his father reproached him, saying, "You care for nothing but shooting, dogs, and rat-catching, and you will be a disgrace to yourself and all your family." He was then sent to Edinburgh University to study medicine, but that also was a failure: the lectures disgusted him with science, witnessing an operation nauseated him; and he only enjoyed collecting marine animals in tidal pools, accompanying fishermen trawling for oysters, and learning to skin and stuff birds from a Negro who had accompanied a naturalist in South America.

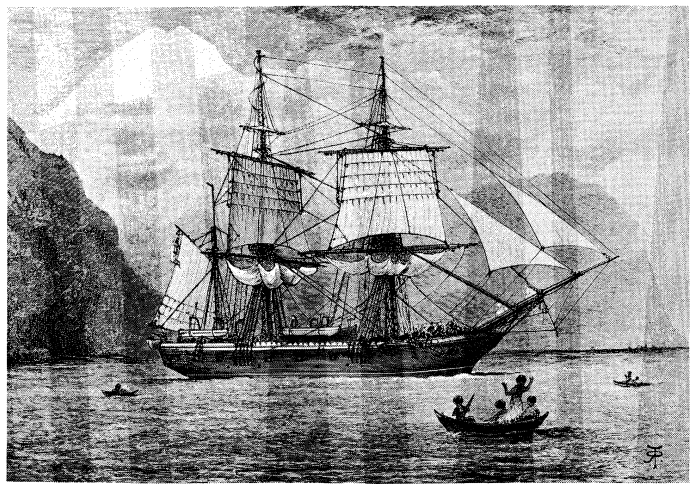
As there was no future for Darwin in medicine, he left Edinburgh in 1827 and was sent to Cambridge to prepare for Holy Orders in the Church of England. At Christ's College he paid little attention to his official studies and fell in with a set of sporting young men as keen on shooting, riding, and hunting as he was. But he also got to know some distinguished scientists—in particular, John Stevens Henslow, professor of botany, who influenced Darwin profoundly by stimulating his interest in natural history and by giving him confidence in himself.

Career. In 1831 the Admiralty asked for a naturalist to accompany Capt. Robert Fitzroy of the Royal Navy on a voyage in HMS "Beagle" to survey the coasts of Patagonia, Tierra del Fuego, Chile, and Peru, to visit some Pacific islands, and to establish a chain of chronometrical stations around the world. Henslow recommended Darwin, who wanted to accept, but whose father objected that it would only be another interruption in his checkered education. His uncle Josiah Wedgwood II, however, persuaded Robert Darwin that his objections were unsound and he withdrew them. Darwin sailed from Devonport in the "Beagle" on Dec. 27, 1831. He was to be away five years.

Voyage of the "Beagle"

The Cape Verde Islands provided him with his first object lesson of a volcano, on which he was able to test for himself the validity of Charles Lyell's *Principles of Geology*, which Henslow had advised him to take with him, but on no account to believe. Darwin did believe the results of his own observations, however, and these confimed him in his acceptance of Lyell's doctrine of uniformitarianism (*i.e.*, natural laws apply uniformly throughout time) and laid the foundations of his future as a scientist. In Brazil he saw his first tropical forest; in Argentina he found his first fossils—sloths, mastodons, and horses. In Tierra del Fuego he saw a race of men so savage, so devoid of any beliefs (and even occasionally cannibalistic) that they hardly seemed human. Three of them had been taken to England three years previously by Captain Fitzroy "to teach them the elements of Christianity and the use of tools," and they were now being repatriated. Darwin was astonished that "three years had been sufficient to change savages into as far as habits go complete and voluntary Europeans." But they soon reverted to savagery.

In Chile Darwin witnessed an earthquake and observed both its effects in raising the level of the land and its connection with volcanic eruption. Repeatedly when ashore he went on long, arduous, and dangerous expeditions on horseback, collecting and shooting, which showed that his addiction to sport had not been useless. On more than one occasion he saved the situation for his companions: once by running far and fast enough to save their boat from being destroyed by the wave raised by a glacier fall (they would all have been doomed had he failed) and another time by going on to get help when his captain and companions were exhausted and incapable of walking a step farther. Wherever he saw a mountain he



HMS "Beagle" in Straits of Magellan, engraving by unknown artist. Mount Sarmiento is in the distance.
The Bettmann Archive

climbed it, and on one journey from Chile to Argentina over high passes of the Andes, he was bitten massively by bugs. From the Galápagos Islands the "Beagle" sailed to Tahiti, New Zealand, Australia, Cocos Keeling Atoll, Mauritius, South Africa, St. Helena, Ascension Island, Brazil again (to check chronometers), and then home. Darwin landed at Falmouth on Oct. 2, 1836.

All Darwin's work stemmed directly from the observations and collections that he made during the voyage of the "Beagle." As shown in the title of his book *Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle, 1832–36* (1839), his main interests were at first geological (although natural history took precedence over geology in the second edition of his *Journal*, 1845), and his observations resulted in three further books: *Structure and Distribution of Coral Reefs* (1842), *Geological Observations on Volcanic Islands* (1844), and *Geological Observations on South America* (1846). In the eyes of posterity, these works were so eclipsed by Darwin's bombshell on evolution that they have been neglected, but they were fundamental to his later work.

Geological writings

Starting from the fact that coral polyps can live only in clear salt water less than 20 fathoms deep, at temperatures not less than 68° F, and that coral atolls and barrier reefs are all at about sea level, Darwin argued that such atolls and reefs could only have resulted from subsidence of the sea floor, the corals growing upward as their bases dropped. Darwin's view has since been confirmed by deep borings of coral reefs revealing at depths of nearly 5,000 feet dead corals that once lived within 120 feet of the surface.

Another kind of reef, which Darwin called the fringing reef, occurs above sea level and results from elevation of the sea floor. Plotting on the map the distribution of atolls and barrier reefs on the one hand and of fringing reefs on the other, he saw that great areas of the ocean bottom had undergone subsidence, and others elevation, and that all active volcanoes are in the latter. This agreed with the correlation that he had observed in South America between volcanic action and elevation of the ground. That such changes of level could be substantial he saw from his discovery in the Andes—at an altitude of 7,000 feet—of a fossil forest overlain by thousands of feet of sedimentary deposits laid down by the sea, thus proving the occurrence of earlier earth movements of the order of 10,000 feet vertical height.

In petrology his comparison of volcanic lavas with plutonic rocks showed that they were closely related. The minerals in crystalline granites and in glasslike lavas were similar. His studies of the direction of the strike and the angle of dip of strata showed that planes of cleavage and of foliation were constant over very wide areas, parallel to the direction of great axes along which elevation of land had taken place over hundreds of miles. Furthermore, these planes of cleavage and foliation had no relation to the planes of stratification of sedimentary deposits and had been superimposed on strata by pressure and recrystallization. This was the origin of the "deformation" theory of metamorphic rocks.

The collection of animals that he made was described by a team of specialists under Darwin's editorship and published in *The Zoology of the Voyage of the Beagle* (1840–43). But Darwin's biological contribution to science was of a different nature. When he started on the voyage, like everyone else he did not question the immutability of species. But several questions set him thinking: Why do so many similar animals exist so far apart geographically? Why does the South American rhea, for example, resemble so closely the African ostrich? On the other hand, why were adjacent areas populated by similar though not identical species? Why, for instance, were the birds and tortoises of each Galápagos island different, although the physical conditions of the islands seemed identical?

After his return Darwin saw, in 1837, that these questions and many more—in comparative anatomy, embryology, classification, geographical distribution, and palaeontology—could be satisfactorily explained if species were not immutable but had evolved into other species, many with a common ancestor. The evolutionary view of descent with modification from ancestral species, exhibited by descendant species, provides a complete explanation of all these questions, which otherwise remain inexplicable and without a common determining principle. Further details of the huge mass of evidence supporting this view from all branches of biology are given in the article on EVOLUTION.

Darwin's evolutionary concept

Darwin realized that it would be useless, in the state of opinion of his day, to try to convince anybody of the truth of evolution unless he could also explain how it was brought about. In searching for this cause he knew that the key to man's success in producing change in cultivated plants and domestic animals was careful selection of parents from which to breed the desired qualities, and he felt sure that selection must somehow also be operative in nature's creation of species. He knew that all individuals in a species were not identical but showed variation, and he realized that some individuals, well adapted to the places they occupied in the economy of nature (in the mid-20th century called ecological niches), would flourish, while others, less adapted, would perish. This was the principle of natural selection that he had grasped as early as 1837, but he still required to know how nature enforced it. On Sept. 28, 1838, he read Malthus' *Essay on the Principle of Population* in which the author tried to show that, as the rate of increase of human population was in a geometrical ratio, while that of increase in human food supply was only in arithmetical ratio, the result must be misery and death for the poor, unless population growth was checked. Malthus' argument was unsound because it has never been determined to what extent human food supply could be artificially increased if it were given sufficient priority and finance. But Darwin saw at once that this fallacious argument could be applied correctly to plants and animals, which are unable to increase their food supply artificially. He saw, too, that in these organisms mortality must be very high, thereby automatically enforcing the mechanism of selection of parents of successive generations. The note in telegraphic style which Darwin entered that day in his "Notebook on Transmutation of Species" is worth quoting:

The principle of natural selection

> On an average every species must have same number killed year with year by hawks, by cold, & c.—even one species of hawk decreasing in number must affect instantaneously all the rest. The final cause of all this wedging must be to sort out proper structure. . . . One may say there is a force like a hundred thousand wedges trying to force every kind of adapted structure into the gaps in the oeconomy of nature, or rather forming gaps by thrusting out weaker ones.

In these words Darwin showed that he had solved the problem of the origin and improvement of adaptation as

a result of selection pressure, and that modification during descent (*i.e.*, evolution) does not take place in a vacuum but is strongly related to the ecological niches occupied by the species. Darwin is therefore included among the founding fathers of the science of ecology.

Evolution had been advocated before, by some French speculative philosophers (Montesquieu, Maupertuis, Diderot), by Darwin's grandfather Erasmus Darwin, and by Lamarck, who was the first to draw up an evolutionary tree from unicellular organisms to man. But Lamarck's neglecting to provide evidence for evolution and his attempt to explain its cause by appealing to an "inner feeling"—the supposed existence of a tendency to perfection—and to his equally fanciful belief in the satisfaction of the needs of the organisms, led other scientists to reject his evolutionary theory. Darwin was the first to provide adequate evidence for evolution and to explain how the process of natural selection produces adaptation.

Delay in publishing his discoveries

After having discovered the greatest general principle in biology, Darwin kept it to himself. In 1842 he pencilled a "Sketch" of his results, which he expanded in an "Essay" in 1844, but he showed it only to his botanist friend Joseph Dalton Hooker. From 1846 to 1854 he devoted his attention and energies to a study of the different species of living and fossil barnacles, for the purpose of classifying them. This tedious work provided him with firsthand experience of the amount of variation found in species and of the problems of classification, essential for the study of how species originate. These researches were published between 1851 and 1854 in four specialized monographs on stalked and sessile, living and fossil Cirripedia.

In 1856 Darwin started to put on paper his discoveries about evolution and natural selection, adding all the time to his evidence by studies on the problem of divergence (*i.e.*, the greater variability of species belonging to wide-ranging genera containing many species), on geographical distribution and the function of sea and wind in disseminating the population of oceanic islands, and by discussions with his friends Lyell, Hooker, and Thomas Henry Huxley. The work went steadily on until, on June 18, 1858, out of the blue, Darwin received from Alfred Russel Wallace, a naturalist then in the Malay Archipelago, a succinct but complete statement of his own conclusions on evolution and natural selection. Darwin's shock at the danger of being forestalled in work on which he had been engaged for 20 years was great, but the situation was saved by Lyell and Hooker who insisted that a joint paper by Darwin and Wallace should be read before the Linnean Society of London on July 1, 1858. Darwin then started what he called an "abstract" of the full work on which he was engaged. This abstract was the *Origin of Species*, which was published on November 24, 1859, and sold out immediately. By 1872 the work had run through six editions.

Publication of *Origin of Species*

With this book Darwin brought down on himself enemies of two kinds. The first were old-fashioned scientists, some like Adam Sedgwick who refused to admit Darwin's method of using hypothesis as acceptable in science, and another like Richard Owen who had until then enjoyed the (very undeserved) reputation of being the leading English biologist and who, mad with jealousy and devoid of all scruple, realized that his former friend Darwin might eclipse him altogether and must, at any cost, be discredited. The other class of enemies were the upholders of orthodox religious beliefs, to whom Darwin administered two shocks: if evolution was true, the account of the Creation in the Book of Genesis was false or, at least, not literally true, and if evolution worked automatically by natural selection, there was no room for divine guidance and design in the production of living plants and animals, including man, on earth. The battle was joined at the Oxford meeting of the British Association for the Advancement of Science on June 30, 1860. Owen had carefully coached Samuel Wilberforce, bishop of Oxford, who attacked Huxley in a patronizing and contemptuous manner over Darwin's views. Huxley tore the bishop to pieces in his brilliant reply, and the Church of England never again formally attempted to cross swords with science.

Darwin published three more books as extensions and amplifications of the principles expressed in the *Origin of Species*. These were *The Variation of Animals and Plants Under Domestication* (1868), which was a detailed inquiry into the origins and varieties of cultivated plants and domestic animals; *The Descent of Man and Selection in Relation to Sex* (1871), which applied the principle of evolution to man, including his moral sense, and introduced the notion of sexual selection as complementary to natural selection; and *The Expression of the Emotions in Man and Animals* (1872), which was the culmination of an interest that Darwin had already shown in 1837 when he wrote "Seeing a dog, horse and man yawn, makes one feel how all animals are built on one structure." The latter book was a notable contribution to psychology and the foundation of the science of ethology.

Botanical writings

Darwin's remaining books dealt with plants, but all had running through them the theme of adaptation, its origins and improvements, and the survival value that it conferred. This is clearly seen in *On the Various Contrivances by Which British and Foreign Orchids are Fertilised by Insects* (1862), in which he showed that the flowers have a shape adapted to the landing on them of insects, which thrust their proboscises into the flower in search of nectar, and that the pollen sacs then become detached from the flower and attached to the insect, which flies away and pollinates other flowers. There are a number of fascinating points involved here: Darwin showed that all flowers that depart from radial symmetry are adapted to the reception of insects, whose role in pollinating flowers explains why flowering plants and insects both evolved rapidly in the Jurassic Period, when insects first appeared, and also why no plant that is pollinated by wind has coloured flowers, the function of which is to attract insects.

Having observed cross-pollination by insects and assuming that it provided a selective advantage, Darwin proceeded to investigate it. He raised two large beds of plants from cross-pollinated and self-pollinated seeds obtained from the same parent plant and found that the former were larger, more vigorous, and more fertile than the latter. The reasons for this are now well understood from Mendelian genetics. Darwin's results were published in *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom* (1876). He had found the reason why there are two sexes in the plant and animal kingdoms, for they are adaptations to ensure cross-fertilization and, as was learned later, interchange of genes.

In *Climbing Plants* (1875) Darwin studied not only the mechanism of twining, which he found to be a circular sweeping movement of growth by the apex of stems, but also the significance of this ability to twine and climb, which enables the plant to reach in one season a height where its leaves are more exposed to sunlight and air, without having to develop a thick woody trunk. Twining is therefore an adaptation, which is improved in some species by the conversion of leaves into tendrils and even further in others where the tips of the tendrils when they strike a solid surface are converted into adhesive disks, as in Virginia creeper. This Darwin viewed as a progressive series of adaptive improvements.

These observations led Darwin to experiment on the factors that cause growing tips of shoots and roots to bend; *i.e.*, to grow faster on one side than on the other. His great discovery here was that the growing tip of a shoot is sensitive to light, and that the tip turns toward the light because the stem grows on the side of the shoot away from the light; but this growth occurs at a place some distance down the stem, even if it is screened from light. Therefore, there is "some matter in the upper part which is acted upon by light, and which transmits its effects to the lower part." Similarly, in roots it is the tip that is sensitive to gravity and to contact with solid objects and that brings about differential growth further up the root as well as bending. From these experiments

sprang the whole science of growth hormones in plants. In *Insectivorous Plants* (1875) Darwin described the most remarkable adaptations. In the sundew (*Drosera rotundifolia*) the upper surface of each leaf bears some hundred tentacles ending in a gland that secretes a viscid substance. When a fly touches these glands it becomes stuck, and neighbouring tentacles bend toward it and their glands fix on the fly. Raw meat, ammonia, and other solutions, Darwin found, have the same effect as a fly. The secretion of the glands dissolves (digests) albumin, muscle, connective, and other tissues. The bending of the tentacles is due to the transmission of excitation through the cells of the leaf. The sensitivity of the glands is such that a tiny piece of hair weighing 0.0008 milligrams is sufficient to cause bending. The most extraordinary thing of all is that after a fly has been caught and the glands have secreted their digestive juice onto it the products of the digested insect are then absorbed by the glands into the plant. This fantastic state of affairs led Darwin to write jokingly to Hooker (Dec. 4, 1860), "By Jove, I sometimes think *Drosera* is a disguised animal." That it is adaptive was shown by Darwin's son Francis who grew *Drosera* plants "fed" on meat and compared them with "unfed" plants under the same conditions. The former were larger, had more leaves and more seed capsules. Later it became clear that *Drosera* can grow on very poor soil and has few roots because it supplements its nitrogen supply with animal food.

Darwin as an ecologist

The Formation of Vegetable Mould Through the Action of Worms (1881) was a pioneer study in quantitative ecology. The amount of soil, ground fine, brought up to the surface from the depth of a foot by earthworms, Darwin calculated (by weighing wormcastings) to amount to 18 tons per acre per year. The mold thus formed is not only aerated and most suitable for plant growth, but the wormholes make it easier for the rootlets to grow. The importance of this work was stressed in the mid-20th century by the fact that the chemical manures and pesticides that came to be used in agriculture kill the unpaid army of worms and reduce fertility.

Personality. The development of Darwin's mind and personality presents a number of problems. A few weeks before he served on the "Beagle" he did not even know what science was, until the Rev. Adam Sedgwick showed him that if a certain fossil shell of a tropical mollusk had really been found in a certain glacial deposit, it would overthrow all that was known of geology in England, for science is a consistent and organized body of knowledge. It was therefore no experienced scientist who sailed on the "Beagle" but an undistinguished candidate for Holy Orders equipped mainly with courage and horse sense. The man who returned, however, was the hardest headed biologist of the century. Part of this mental metamorphosis was due to his reading Lyell's *Principles of Geology* and applying them in the field to the facts that he observed. Another factor was his conviction from what he saw that the account of the Creation in the Bible was demonstrably false. He had learned to apply critical judgment.

At the same time he was astonishingly naïve in such general matters as methodology. In his day science was supposed to make progress only by inductive methods. Darwin wrote, "I worked on true Baconian principles and without any theory collected facts on a wholesale scale." This was invalidated by Darwin himself when he wrote to Lyell (June 1, 1860): "Without the making of theories, I am convinced there would be no observations." Thus he recognized that there are no such things as purely inductive observations, for if the observer had not already in his head an idea of what he was looking for, derived from deduction, he would not observe anything at all.

Darwin's method was to spin a hypothesis about anything that struck his attention (*i.e.*, anything that he was predisposed by ideas to see), and then to deduce from it consequences that should follow and could be refuted or verified. This "hypothetic-deductive" method, as Sir Peter Medawar has called it, is illustrated in Darwin's letter to F.W. Hutton (April 20, 1861): "I am actually weary of telling people that I do not pretend to adduce direct evidence of one species changing into another, but I believe that this view is in the main correct, because so many phenomena can thus be grouped and explained."

Method of work

As he went on and realized the enormous importance of his demonstration of evolution by natural selection, he began to concentrate more and more deeply on this field. He wrote to Hooker (Oct. 13, 1858): "It is an accursed evil to a man to become so absorbed in any subject as I am in mine." All his mental energy was focussed on "his subject," and that was why poetry, pictures, and music ceased in his mature life to afford him the pleasure that they had given him in his earlier days. As he himself said, his mind had become "a kind of machine for grinding general laws out of large collections of facts."

His technique for collecting facts, in addition to those that he observed for himself, was illustrative of his tenacity. Whenever he found any man whose practical experience in a field of study had made him expert, Darwin bombarded him unremittingly with questions and clung to him like a leech. He always gilded the pill of his questions with disarmingly sweet phrases: "if it would not cause you too much trouble" or "pray add to your kindness"; and he realized what a burden he must be to his correspondents, as he admitted to J. Jenner Weir (March 6, 1868), "If any man wants to gain a good opinion of his fellow men, he ought to do what I am doing, pester them with letters."

It is remarkable that, though in physical appearance he aged very markedly, he retained the most endearing childlike traits. On Jan. 20, 1839, he wrote to his fiancée, his first cousin Emma Wedgwood, nine days before their wedding, "I take so much pleasure in the [new] house, I declare I am just like a great overgrown child with a new toy; but then not like a real child, I long to have a copartner and possessor." He remained a great overgrown child all his life, and this accounts for his ever-present sense of fun, often expressed at his own expense, as when he accepted Jenner Weir's invitation to become a Patron of the Cat Show (Sept. 18, 1872), and added, "People may refuse to go and admire a lot of atheistical cats." His description of a good novel was that it must have in it some character that one can thoroughly love, and if a pretty woman, all the better. When he lost a game of cards he exploded in mock anger. He was also fond of teasing friends.

This childlike mentality was responsible for some astonishingly naïve views that he held. To Huxley he wrote (Jan. 9, 1860), "The history of error is quite unimportant." He had no historical or political sense whatever, as may be seen in what he wrote to the Austrian explorer Karl von Scherzer, (Dec. 26, 1869): "What a foolish idea seems to prevail in Germany on the connexion between Socialism and Evolution through Natural Selection." He was not unaware of Karl Marx, for when the latter asked permission to dedicate the English edition of *Das Kapital* to him, Darwin refused because he did not wish to be associated with attacks on religion. This needs explanation.

Darwin's agnosticism

The former candidate for Holy Orders had come to see that the Old Testament, "from its manifestly false history of the earth, . . . and from its attributing to God the feelings of a revengeful tyrant, was no more to be trusted than the sacred books of the Hindoos, or the beliefs of any barbarian." The New Testament did not fare any better, and he could "indeed hardly see how anyone ought to wish Christianity to be true; for if so, the plain language of the text seems to show that the men who do not believe, and this would include my Father, Brother and almost all my best friends, will be everlastingly punished. And this is a damnable doctrine." The key to understanding Darwin's thinking is his horror of the imposition of suffering—on slaves by their masters, on animals by men, and by "the clumsy, wasteful, blundering, low, and horribly cruel works of nature," as seen in the suffering caused by parasites and in the delight in cruelty

shown by some predators when catching and playing with their prey. If God is as almighty, omniscient, and possessed of inexhaustible compassion as he is painted, "it revolts our understanding to suppose that his benevolence is not unbounded." So Darwin became a reverent agnostic.

As a young man he was passionately fond of shooting, but after he discovered a bird that had been maimed but not killed on a previous shoot, he never shot again. But he did not give way to sentimentalism, as may be seen in the letter that he wrote to *The Times* (London) on June 23, 1876, about the controversy then (and since) raging over vivisection: "Women, who from the tenderness of their hearts and from their profound ignorance are the most vehement opponents of all such experiments, will I hope pause when they learn that a few such experiments performed under the influence of an anaesthetic, have saved and will save through all future time thousands of women from a dreadful and lingering death."

Darwin's illness

The young man who, during all the hardships and dangers of the voyage of the "Beagle" and his numerous hazardous journeys ashore, enjoyed stout health and great physical stamina, within a few months of his return began to show increasingly frequent symptoms of illness that reduced him to a state of semi-invalidism. These were great lassitude, painful intestinal discomfort with much flatulence, frequent vomiting, and sleeplessness. His doctors were quite unable to find any organic cause for his condition, in which he settled down to a daily routine of four hours' work, walks in the garden, and rests on a sofa smoking a cigarette while being read to. Frequently he was unable to work at all. On several occasions he went to take the water cure at a spa, but without lasting benefit.

Recently, psychiatrists have claimed to account for Darwin's condition by advancing diverse (and contradictory) explanations: that he had "poor nervous heredity on both sides," or "depressive obsessional anxiety and hysterical symptoms" due to "a distorted expression of aggression, hate, and resentment felt at an unconscious level by Darwin towards his tyrannical father"; that his work on evolution had killed the Heavenly Father and given him an Oedipus complex; and that his shunning of social life and acceptance of his wife's ministering care was evidence of his being a neurotic. All this specious and special pleading is unnecessary since S. Adler drew attention to the massive attack that Darwin suffered in 1835 from the bites of a bug, *Triatoma infestans*, the most important carrier of the trypanosome of Chagas' disease. This trypanosome, not discovered until 1909, is found in the blood even many years after infection; it causes lassitude and heart block and prevents normal functioning of the intestines. The case histories of patients with Chagas' disease and Darwin's symptoms fit like a glove: he suffered a heart attack in 1873 and died of another in 1882.

It is against this background of suffering that Darwin's scientific work was achieved and that his family life was lived. His home was an oasis of loving-kindness, as all visitors to Down House found, after the family had moved there in 1842 from London, where the strain of social life tired Darwin too much. Ten children were born, of whom two died in infancy. One, Anne, died at the age of ten under very tragic circumstances, and when Darwin had recovered somewhat from his grief he composed a little ode on her, the most beautiful thing that he ever wrote.

Of his other children, George, Francis, and Horace made such important contributions to science in astronomy, botany, and civil engineering, respectively, that they were knighted; Leonard was an economist and eugenicist. Subsequent generations have also shown great distinction, a tribute alike to Darwin and Wedgwood genes and to the family circles in which they grew up.

Darwin died at Down on April 19, 1882. Twenty members of Parliament petitioned the dean of Westminster for him to be buried in the Abbey. Agreement was reached immediately and the funeral took place on April 26. Among the pallbearers were Hooker and Huxley, as well as a great number of foreign dignitaries.

MAJOR WORKS

Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle, 1832–36, popular title, *The Voyage of the Beagle* (1839); *The Structure and Distribution of Coral Reefs* (1842); *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1859); *The Various Contrivances by Which Orchids are Fertilised by Insects* (1862); *The Movements and Habits of Climbing Plants* (1865); *The Variation of Animals and Plants Under Domestication* (1868); *The Descent of Man, and Selection in Relation to Sex*, 2 vol. (1871); *The Expression of the Emotions in Man and Animals* (1872); *The Insectivorous Plants* (1875); *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom* (1876); *The Different Forms of Flowers on Plants of the Same Species* (1877); *The Power of Movement in Plants* (1880); *The Formation of Vegetable Mould Through the Action of Worms* (1881).

BIBLIOGRAPHY. FRANCIS DARWIN (ed.), *Life and Letters of Charles Darwin*, 3 vol. (1887), is the standard life by Darwin's son. It contains most of the personal information about Darwin but was written at a time when the full significance of evolution of natural selection was not understood, and when the practice of transcribing and publishing letters was imperfect. *More Letters: Letters of Charles Darwin*, 2 vol., ed. by FRANCIS DARWIN and A.C. SEWARD (1903), is also a collection of essential documents but again suffers from inadequate methods of publication.

Works edited by SIR GAVIN DE BEER include: *Darwin and Huxley: Autobiographies* (1971), an edition collated from the original manuscript that represents the only accurate text; "Darwin's Sketch of 1842" and "Essay of 1844," in *Evolution by Natural Selection* (1958), contain Darwin's earliest exposition of his discoveries, first published by Francis Darwin in 1909 and long out of print; "Darwin's Notebooks on Transmutation of Species," *Bull. Br. Mus. (Nat. Hist.)*, Historical Series, vol. 2 (1960) and vol. 3 (1967), contain the evidence of the progress of Darwin's thoughts from 1837 to 1839. The only work to take into account Darwin's notebooks on transmutation of species and Mendel's attitude toward Darwin is *Charles Darwin* by SIR GAVIN DE BEER (1963).

The Cambridge University Library is the receptacle of most Darwin manuscripts, but an important collection is in the American Philosophical Society Library in Philadelphia, containing Darwin's letters to Sir Charles Lyell and a "Critical Chronology of Darwin's Life" on index cards.

(G.de B.)

Dating, Relative and Absolute

Dating is placing events in time—a necessary pursuit to understanding the ebb and flow of the past. In the context of dating, numbers such as AD 1492, when Columbus came to the New World, or AD 1066, the Battle of Hastings, come to mind. This kind of dating is called absolute dating, and a specific number of years is always given. Dating also can be done with relation to notable events or periods of history, without any numerical citation; this is described as relative dating. Examples are Renaissance literature and Colonial architecture, where the adjectives Renaissance and Colonial relate the nouns to past intervals that are noted more for what happened than for when it happened.

Distinction between absolute and relative dating

Ideally, it would be desirable to date a given event both absolutely and in a relative sense. When the extraordinary events and periods of history are absolutely dated, it becomes possible in many instances to date them relatively as well. If a certain English poem were known to have been written in AD 1580, for example, it would be—in relative terms—Elizabethan, because England's Queen Elizabeth I reigned from 1558 to 1603.

Unfortunately, the converse of the above principle does not always follow. That is, an event relatively dated is not necessarily dated absolutely. Two situations will demonstrate why the converse is not always so. First, the date of an event may be known only as having occurred before or after another event, with no knowledge of how long before or after. The discovery of fire, for example, is dated relatively as having occurred in prehistoric time—that is, before about 6,000 years ago. How much before is uncertain at present.

A second situation in which relative dating is possible without knowledge of absolute age occurs when the sequence of notable happenings to which dating is related has not been defined absolutely. In geology and in prehistoric archaeology, this situation existed for many decades before the application of radiometric dating. Such relative terms as Paleozoic and Neolithic originated when sequences of important events had been deciphered from layered deposits but without knowledge of the absolute dates of the events comprising each sequence.

This article treats dating outside the traditional field of history, which relies upon written records. Its focus instead is on human remains and artifacts that predate traditional history and on the geological rock record that reflects even more remote events. The dating of these events and these objects is of major concern here. This article also covers the several methods of dating and their application within the following format:

I. Relative dating
  Application to geological problems
    The rock cycle
    Determination of sequence
    Correlation
    The geological time scale
  Application to archaeological problems
    Archaeological sites
    Correlation
    The archaeological time scale
II. Absolute dating
  General considerations
    Meaning of absolute age
    Dating requirements
    Rate of record accumulation
  Radiometric dating
    Radioactivity and radioactive decay
    Principles of radiometric dating
    Time zero in radiometric dating
    Sources of error
    Major dating methods
    Minor dating methods
  Nonradiometric dating
    Geological processes as absolute chronometers
    Biological processes as absolute chronometers
  Applications of absolute dating
    The geological time scale
    The age of the Earth
    Sea-floor spreading
    Lunar history

For additional information on the several geological topics treated in this article, see EARTH, GEOLOGICAL HISTORY OF; GEOLOGICAL TIME SCALE; STRATIGRAPHIC BOUNDARIES; MARINE SEDIMENTS; IGNEOUS ROCKS; SEDIMENTARY ROCKS; METAMORPHIC ROCKS; ROCK MAGNETISM; and SEA-FLOOR SPREADING. See also ARCHAEOLOGY; CHEMICAL ELEMENTS; ISOTOPES; and RADIOACTIVITY.

I. Relative dating

Historically, determination of relative-age relationships preceded methods for absolute-age measurement. Such an order was not rationally predetermined; it simply reflects the fact that it is usually easier to decipher which of two events came first than to determine how many years ago either one happened.

The determination of the particular events that have happened came before their relative dating. This task generally depends on the existence of tangible things that are indicative of past events. In the field of archaeology, the tangible record consists of artifacts, skeletal remains, and structures such as buildings. In geology the documents of past events are rocks and the surface and subsurface features comprised of rocks. Because of the practical value of geological knowledge, systematic efforts at deciphering the geological past came before extensive archaeological endeavour. This gave rise to the principles of relative dating basic to both fields.

APPLICATION TO GEOLOGICAL PROBLEMS

The rock cycle. A concise description of crustal changes and rock origins is communicated through what is called the rock cycle. In interpreting the rock record, geologists observe the processes that are shaping the Earth today and assume that rocks came into being through similar processes in the past. More accurately, what is assumed rigorously is the temporal uniformity of the physical and chemical behaviour of matter. But applied practically, this usually means leaning very heavily on present processes while being fully aware that their intensity and scope may have been highly variable in the past.

Although all rocks are derived from pre-existing matter in some form, igneous rocks, those crystallizing from silicate melts, are the closest approximation to primordial earth material. Thus, an examination of the rock cycle—the idealized path traversed from molten material through various rock types and, ultimately, back to the bowels of the Earth—generally begins with magma, the molten material that cools on or within the Earth to give rise to igneous rocks.

The fate of almost any rock in extended contact with the atmosphere is to undergo change. When igneous rocks such as granite and basalt interact with atmospheric constituents—mainly oxygen, moisture, and carbon dioxide—the rocks are chemically altered and reduced in particle size. At some size, ranging from atoms to great blocks, particles are transported by water, winds, or glaciers and are deposited elsewhere when the transporting power diminishes. In certain favourable areas, deposition is extensive enough and of sufficient duration to result in the conversion of sediments into rock. Compaction and natural cementing, generally by precipitation of mineral matter from solution in pore spaces, are the main processes responsible for this change, called lithification. If stream sorting of particles has resulted in deposition of a bed of sand-size grains, for example, the rock formed by lithification will be a sandstone whose grains are bound by a mineral cement.

Formation of sedimentary rocks

Sedimentary rocks commonly are deposited in great troughs in the Earth's crust that are loci of subsequent mountain building. Mountain-building conditions involve both crustal deformation and the action of deep-seated igneous processes, and any rock present at depth may be altered; it is then called a metamorphic rock. Marble, a derivative of sedimentary limestone, is an example of such alteration. Under the high pressure and temperature of the metamorphic environment, changes of different sorts occur: the development of new minerals, enlargement of pre-existing crystals, and the rearrangement of crystals into more orderly patterns. If the temperature of metamorphism exceeds the rock melting point, partial or total melting will occur, and a new magma will be formed. Thus, the rock cycle may begin all over again.

Not every rock on Earth has run the gamut of changes included in the rock cycle. The cycle is merely a systematic representation of the sorts of changes, both observed and inferred, that have happened throughout most of Earth history. These are the changes that must be placed in proper sequence—that is, relatively dated in geological problems.

Determination of sequence. Relative geological dating focusses principally on sedimentary rocks. These rocks cover about three-quarters of the area of the continents, and unconsolidated sediments blanket most of the ocean floor. They provide evidence of former surface conditions and the life-forms then present. The sequence of a layered sedimentary series is easily defined because deposition always proceeds from the bottom to the top. This principle would seem self-evident, but its first enunciation more than 300 years ago represented an enormous advance in understanding. It is called the principle of superposition and means that in a series of sedimentary layers or superposed lava flows the oldest layer is at the bottom, and layers from there upward become progressively younger.

On occasion, however, deformation may have caused tilting of the rocks of the Earth's crust, perhaps to the point of overturning them. And, if erosion has blurred the record by removing considerable portions of the deformed sedimentary rocks, it may not be at all clear which edge of a given layer is the original top and which the original bottom.

Problems caused by tilting

Identifying top and bottom is clearly important in sequence determination—so important, in fact, that a considerable literature has grown up centring on this question alone. Many of the criteria of top–bottom determination are based on asymmetry in depositional features. Oscillation ripple marks, for example, are produced in sediments by water sloshing back and forth. When such marks are preserved in sedimentary rocks, they define the original top and bottom by their asymmetric pattern. Certain fossils also accumulate in a distinctive pattern or position that serves to define the topside.

Another factor which causes complications in sequence determination is the action of thrust faulting, which slides older rocks on top of younger ones. Although each layer in a sedimentary series may continue to be oriented with its top side up and bottom side down, their relative ages will no longer accord with the principle of superposition. This occurrence has taken place in many of the world's mountain ranges. In Glacier Park, Montana, for example, a block of sedimentary layers 1,000,000,000 years old has been thrust over layers only one-tenth as old.

Cross-cutting relationships

The principle of crosscutting relationships is another important criterion in sequence determination. It states that, if rock feature *A* cuts across rock feature *B*, then *A* is younger than *B*. Commonly, *B* is a rock mass, whereas *A* is generally a quasi-planar feature, such as a dike (tabular body of igneous rock) or a fault or an unconformity (interruption of the sequence by erosion or nondeposition). Thus, a fault that offsets the stratification of a sedimentary sequence occurred after the layers were deposited. And a dike running up the face of a limestone quarry was injected as magma along a crack in the limestone after the limestone was deposited as sediment. Similarly, an unconformity that terminates a series of inclined layers is a buried erosion surface created after the layers were tilted.

A rock mass encasing a relatively small fragment of different rock type must be younger than the fragment it contains, and this provides another tool for deciphering the rock record. If a conglomerate, for example, consists of pebbles of a distinctive granite, then the conglomerate formed as a rock unit after the parent granite mass solidified from magma. Another example concerns the relative age of a sill formed of magma forced between adjacent layers of sedimentary rock. In the magma injection, chunks of overlying sedimentary materials often break loose and are trapped within the upper part of the sill magma, perhaps later to be seen in a cliff exposing a cross section of the layered series. The fact that the sill is younger than the overlying sedimentary layer becomes clear from the sedimentary fragments encased in the upper sill rock. Without the fragments, other criteria—often more subtle—would be needed to distinguish the sill from a lava flow that had been subsequently buried by younger sedimentary material.

On the basis of the foregoing principles, the proper sequence of the layers exposed in any particular outcrop—be it as extensive as the Grand Canyon or as limited as a small road cut—can be determined. A hypothetical outcrop in which these principles are applied is shown in Figure 1.

From *Journal of Geological Education* (December 1967)

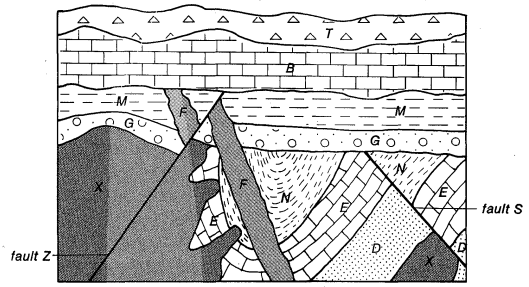


Figure 1: Hypothetical outcrop to which sequence-determining techniques can be applied. The correct sequence, from oldest to youngest, is: *D,E,N,X,S,G,M,F,Z,B,T.*

As the only tangible record of a certain past interval at a certain place, every sedimentary outcrop is essentially the sole source of whatever is to be learned about that interval in that place. Geologic deduction must then extract from the rock at hand as much information about past conditions as is possible—such information as the regional climate, the depositional environment, the nature of the source area that provided sediment grains, and the living forms then extant.

Unfortunately, when the geologist finishes interpreting a single sedimentary outcrop, he has only the story of an unknown time interval from somewhere in the vastness of the past. Beneath the lowermost layer in the outcrop, rocks that would extend that time interval may be present, but they are out of sight. In addition, erosion may have removed overlying rocks and the record that they contain. It is even possible that within a given outcrop there may not be a continuous record of deposition. Whatever the situation, it is clear that a single outcrop falls far short and is inadequate for representing all of the past. The solution is found in the technique called correlation.

Correlation. Correlation is the technique of piecing together the information content of separated outcrops. When two outcrops are pieced together, the time interval they represent is probably greater than that of each alone. Presumably, if all the world's outcrops were integrated, sediments representing all of geological time would be available for examination. This optimistic hope, however, must be tempered by the realization that much of the Precambrian (older than 570,000,000 years) record is missing. Correlating two separated outcrops means establishing that the two share certain characteristics that imply their formation at the same time. The most useful indication of time equivalence is similar fossil content. The basis for assuming that similar fossils indicate contemporary formation is called faunal succession.

Unlike the principles of superposition and crosscutting, faunal succession is a secondary principle. That is, it depends on other sequence-determining principles for establishing its validity. Suppose there exist a number of fossil-bearing outcrops each composed of sedimentary layers that can be arranged in relative order, primarily based on superposition. Suppose, too, that all the layers contain a good representation of the animal life existing at the time of deposition. From an examination of such outcrops with special focus on the sequence of animal forms comes the empirical generalization known as faunal succession—that the faunas of the past have followed a specific order of succession, so that the relative age of a fossiliferous rock is indicated by the types of fossils it contains.

William Smith, an English geologist and engineer, first noticed around 1800 that the different rock layers he encountered in his work were characterized by different fossil assemblages. Using fossils simply for identification purposes, Smith constructed a map of the various surface rocks outcropping throughout England, Wales, and southern Scotland. Smith's geologic map was extremely crude, but in its effect on Earth study it was a milestone.

Following Smith's pioneering work, subsequent generations of geologists have confirmed that similar and even more extensive fossil sequences exist elsewhere. To this day, fossils are useful as correlation tools to geologists specializing in stratigraphy. In dating the past, the primary value of fossils lies within the principle of faunal succession: each interval of geologic history had a unique fauna that associates a given fossiliferous rock with that particular interval.

Index fossils

The basic conceptual tool for correlation by fossils is the index, or guide, fossil. Ideally, an index fossil should be such as to guarantee that its presence in two separated rocks indicates their synchroneity. This requires that the life-span of the fossil species was but a moment of time relative to the immensity of geological history. In other words, the fossil species must have had a short temporal range. On the practical side, an index fossil should be distinctive in appearance so as to prevent misidentifica-

tion, it should be cosmopolitan both as to geography and as to rock type, and its fossilized population should be sufficiently abundant for discovery to be highly probable. Such an array of attributes represents an ideal, and many problems of stratigraphic geology are rendered difficult because of departure of the natural fossil assemblage from this ideal. Nevertheless, there is no greater testimony to the validity of fossil-based stratigraphic geology than that provided by absolute dates made possible through radioactive measurements. Almost without exception, the relative order of strata defined by fossils has been confirmed by radiometric ages.

Correlation based on the physical features of the rock record has also been used with some success, but it is restricted to small areas that generally extend no more than several hundred miles. The first step is determining whether similar beds in separated outcrops can actually be traced laterally until they are seen to be part of the same original layer. Failing that, the repetition of a certain layered sequence (*e.g.*, a black shale sandwiched between a red sandstone and a white limestone) lends confidence to physical correlation. Finally, the measurement of a host of rock properties may well be the ultimate key to correlation of separated outcrops. The more ways in which two rocks are physically alike, the more likely it is that the two formed at the same time.

Only a partial listing of physical characteristics is necessary to indicate the breadth of approach in this area. Such features as colour, ripple marks, mud cracks, raindrop imprints, and slump structures are directly observable in the field. Properties derived from laboratory study include (1) size, shape, surface appearance, and degree of sorting of mineral grains; (2) specific mineral types present and their abundances; (3) elemental composition of the rock as a whole and of individual mineral components; (4) type and abundance of cementing agent; and (5) density, radioactivity, and electrical–magnetic–optical properties of the rock as a whole.

Rock properties determined in the laboratory

With the development of miniaturized analytical equipment, evaluation of rock properties down a small drill hole has become possible. The technique, called well logging, involves lowering a small instrument down a drill hole on the end of a wire and making measurements continuously as the wire is played out in measured lengths. By this technique it is possible to detect depth variations in electrical resistivity, self-potential, and gamma-ray emission rate and to interpret such data in terms of continuity of the layering between holes. Subsurface structures can thus be defined by the correlation of such properties.

Field geologists always prize a layer that is so distinctive in appearance that a series of tests need not be made to establish its identity. Such a layer is called a key bed. In a large number of cases, key beds originated as volcanic ash. Besides being distinctive, a volcanic-ash layer has four other advantages for purposes of correlation: it was laid down in an instant of geological time; it settles out over tremendous areas; it permits physical correlation between contrasting sedimentary environments; and unaltered mineral crystals that permit radiometric measurements of absolute age often are present.

Correlation may be difficult or erroneous if several different ash eruptions occurred, and a layer deposited in one is correlated with that from another. Even then, the correlation may be justified if the two ash deposits represent the same volcanic episode. Much work is currently under way to characterize ash layers both physically and chemically and so avoid incorrect correlations.

The geological time scale. The end product of correlation is a mental abstraction called the geological column. It is the result of integrating all the world's individual rock sequences into a single sequence. In order to communicate the fine structure of the geological column, it has been subdivided into smaller units. Lines are drawn on the basis of either significant changes in fossil forms or discontinuities in the rock record (*i.e.*, unconformities); the basic subdivisions of rock are called systems, and the corresponding time intervals are called periods. In the upper part of the geological column, where fossils abound, these rock systems and geological periods are the basic units of rock and time. Lumping of periods results in eras, and splitting gives rise to epochs. In both cases, a threefold division into early–middle–late is often used, although those specific words are not always applied. For example, the three eras with abundant life have prefixes paleo ("old"), meso ("middle"), and ceno ("recent"). Similarly, many periods are split into three epochs. Names assigned to individual epochs follow no single worldwide standard except for the seven epochs comprising the last two periods (Figure 2).

Divisions of the geological column

Adapted by permission from Don L. Eicher, *Geologic Time* (1968); Prentice-Hall

relative duration of eras	era	period	epoch	duration in 000,000 of years (approx.)	000,000 of years ago (approx.)
Cenozoic			Holocene	approx. lasting	10,000 years
Mesozoic	Cenozoic	Quaternary	Pleistocene	2.5	2.5
Paleozoic			Pliocene	4.5	7
		Tertiary	Miocene	19	26
			Oligocene	12	38
			Eocene	16	54
			Paleocene	11	65
	Mesozoic	Cretaceous		71	136
		Jurassic		54	190
		Triassic		35	225
	Paleozoic	Permian		55	280
		Carbon: Pennsylvanian		45	325
		Carbon: Mississippian		20	345
		Devonian			395
		Silurian		35	430
		Ordovician		70	500
		Cambrian		70	570
Precambrian	Precambrian			4030	
	formation of Earth's crust				4,600

Figure 2: Geological time scale.

Over the interval from the Paleozoic to the present, about 35 epochs are recognized in North America. This interval is represented by approximately 250 formations, discrete layers thick enough and distinctive enough in lithology to merit delineation as units of the geological column. Also employed in subdivision is the zone concept, in which it is the fossils in the rocks rather than the lithologic character that defines minor stratigraphic boundaries. The basis of zone definition varies among geologists, some considering a zone to be all rocks containing a certain species (usually an invertebrate), whereas others focus on special fossil assemblages.

The lower part of the geological column, where fossils are almost absent, was at one time viewed in the context of two eras of time, but subsequent mapping has shown the provincial bias in such a scheme. Consequently, the entire lower column is now considered a single unit, the Precambrian. Hopefully, the results of radiometric dating will someday provide finer Precambrian subdivisions that have worldwide applicability.

The geological column and the relative geological time scale are sufficiently defined to fulfill the use originally envisioned for them—providing a framework within which to tell the story of Earth history. Just as human history has its interweaving plots of warfare, cultural development, and technological advance, so the Earth's rocks tell another story of intertwined sequences of events. Mountains have been built and eroded away, seas have advanced and retreated, a myriad of life-forms have inhabited land and sea. In all these happenings the geological column and its associated time scale spell the difference between an unordered series of isolated events and the unfolding story of a changing Earth.

APPLICATION TO ARCHAEOLOGICAL PROBLEMS

The fields of archaeology and geology are significantly different in subject matter. To the archaeologist, man is the central figure, his cultural and skeletal remains are the materials for study, and knowledge of his developing culture is the goal. By contrast, man is a minor character in the geological drama. He appeared almost at the end of the story, when the topography and continental outlines had achieved essentially their present configurations and life on Earth was much like that existing today. In methodology, however, geological and archaeological investigations have many similarities. Both focus on a record within the Earth and exhume objects for laboratory study; more importantly, both depend on superposition and correlation to introduce chronological order among widely scattered remains of ancient events.

Archaeological sites. Although archaeological sites do not display the almost perfect layering characteristic of most sedimentary outcrops, they often show enough layering to permit application of the principle of superposition for sequence determination. Much of the layering is the result of natural sedimentary processes that take place at the sites of human occupation. Dust and volcanic ash settle from the air, flood deposits cover living sites adjacent to rivers, and cave floors accumulate ceiling rocks, dripstone coatings, or animal excrement. The growth of vegetation may make a significant contribution.

Layering in archaeological sites

In towns occupied for long intervals, streets gradually rise by serving as rubbish receptacles, while degenerating houses collapse alongside them. Every so often, fire or earthquake or invader strikes with destruction, forcing residents to rebuild from scratch on freshly smoothed-out rubble. Such events have raised the multilayered tells of the Middle East as much as 100 feet (30 metres) above their surrounding plains.

Man, of course, is as likely to stir up layers as he is to form them. In burying his departed brethren or the day's garbage, he may be forcing relatively young objects into levels of far greater antiquity. Failure to recognize such intrusions may lead to erroneous sequences and stand in the way of reconstructing a realistic history.

In its ideal configuration, an archaeological site is a series of layers that must be stripped away one at a time. The artifacts or structures in each layer are separated from the sedimentary matrix, after which their locations in three-dimensional space are recorded. Hopefully, such locations in the horizontal time plane have cultural significance rather than merely reflecting random scattering.

Because no site is ideally layered, one or more trenches are often dug in order to see the stratigraphy in cross section. Trenches indicate the approximate thickness of culture-bearing deposits and provide some idea whether the site is worth the digging required to examine it.

Correlation. Artifacts are to archaeological correlation what fossils are to rock correlation. As correlation tools, both artifacts and fossils are undergirded by the principle of superposition for defining their order of appearance. Both include special types that are valuable because their temporal range is very short. Both have successions of forms that developed over time. And the limitations and problems that beset one are like those that plague the other.

Major artifacts used in archaeological correlation

Chief among the artifacts useful in archaeological correlation are tools, hand weapons, and pottery fragments (potsherds). Generally made of durable materials, such objects can perform their intended function through a multitude of designs. Like fossils, there are no a priori grounds on which to predict the evolution in design of a pot or a spear point or even whether change in style necessarily occurred.

The geographical spread of an ancient pottery style depended on the degree to which trade homogenized a region. To the extent that individual groups were isolated from other societies by natural and cultural barriers, their artifacts would almost certainly develop along different lines. Isolation, then, clearly limits the distance of correlation and introduces an undesirable element of provincialism in archaeological study.

As a way to overcome the spatial limitations imposed by correlating with artifacts, there has been a concerted effort by archaeologists to search out climatic indicators associated with cultural material—such features as characteristic sediments, distinctive weathering patterns, and climate-sensitive plants and animals. The goal is to match various archaeological sequences with the climate chronology that has been determined for the Ice Age and subsequent time (see Figure 3). It follows that two widely separated archaeological layers would be correlated if they fitted identically into the standard relative chronology of climatic change.

From William Lee Stokes, *Essentials of Earth History: An Introduction to Historical Geology*, 2nd ed. (© 1966); Prentice-Hall, Inc.

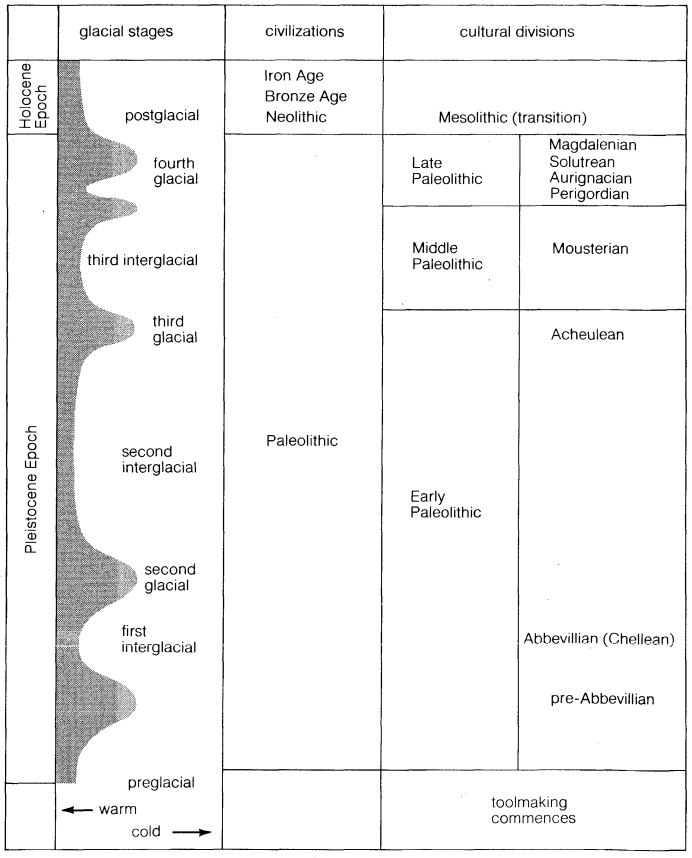


Figure 3: Archaeological time scale.

Associated fossils are a further aid in correlation, provided they are of extinct organisms. In fact, it was the association of man-made implements with the bones of such extinct mammals as the mammoth and rhinoceros that was instrumental in winning the day for those propounding the great antiquity of man. The crowning evidence came in 1864 with the discovery in a French cave of a mammoth's portrait engraved on a mammoth tusk.

Despite a century of developing principles and methods, archaeological correlation in 1950 clearly needed a new technique with sufficient versatility to integrate hundreds of isolated sites. It was then that carbon-14 dating first became feasible, evoking an overwhelmingly positive response from archaeologists. In carbon-14 dating they were offered an absolute dating technique spanning the last 50,000 years and applicable to the sorts of organic material often plentiful in culture layers. If its numerical ages were correct, then worldwide correlation and sequence determination of separated sites was possible, as well as the unscrambling of stratigraphically complex deposits.

Importance of carbon-14 dating

On the basis of more than 20 years of experience, it is now safe to say that carbon-14 dating is indeed the Rosetta Stone of prehistoric archaeology. A measure of chronological order has been introduced where previ-

ously there was often chaos. Furthermore, it has provided the yardstick necessary to establish and calibrate additional chronometers based on time-dependent physical properties such as thickness of obsidian rinds and magnetic intensity. Only with fossil men, whose ages are measured in hundreds of thousands of years, does carbon-14 dating fail to provide insights. And here potassium–argon dating gives evidence of picking up where carbon-14 leaves off. In combination, potassium–argon and carbon-14 hold out the promise of understanding mankind's history to a degree undreamed of before their introduction. Hopefully, the sharp distinction between history and prehistory will be blurred by this new chronologic light on man's ancient past. Details concerning the two techniques are discussed under *Radiometric dating: Major dating methods.*

The archaeological time scale. Like its geological counterpart, the archaeological time scale (Figure 3) is a product of Western intellectual thought that crystallized chiefly within the 19th century. Parallels between the two time scales are remarkable in many ways. First, there were ideological barriers to overcome. This meant at the outset that stone artifacts had to be properly recognized as products of man rather than freaks of nature or the works of elves and fairies. That accomplished, a more formidable task remained—convincing Western man that the history of his ancestors was far longer than that allowed for in Archbishop Ussher's pronouncement that creation occurred in 4004 BC.

Thomsen's efforts to create a time scale

The foundation of an archaeological time scale can be ascribed chiefly to C.J. Thomsen of Denmark. While working in his family's shipping business, Thomsen spent one day a week as the unpaid secretary of the Royal Commission for the Preservation of Danish Antiquities. His job was to put some order to the correspondence and specimens that had accumulated for several years. As a rank amateur, Thomsen carried out his task unfettered by any presuppositions on how best to classify material. The artifacts themselves formed the basis of his system. Specifically, they were placed in broad categories according to construction material (stone, metal, and pottery). Further subdivisions were made on the basis of use intended, such as tools, weapons, containers, and cult objects. When his museum was opened for public view some three years later, Thomsen displayed selected specimens in three cabinets housing, respectively, objects of stone, bronze, and iron.

At the grand opening Thomsen probably had not reached the conclusion that material type indicated relative age. But an extensive correspondence with archaeologists in the coming years brought a conviction that prehistory could be divided chronologically into Stone Age, Bronze Age, and Iron Age. Twenty years after taking over as commission secretary, Thomsen published his classic account entitled *Guide to Scandinavian Antiquities*, in which he not only presented his three-age hypothesis but even provided such details as criteria to discern objects of bronze made in the Iron Age. Undergirding Thomsen's chronology were observations on the artifacts that occurred together and, in some cases, stratigraphic relationships. There was even an awareness that certain artifact types suggested evolution of form.

The three ages of stone, bronze, and iron formed only the superstructure of prehistory, however. As expected, closer scrutiny of the physical record brought an avalanche of evidence pointing toward greater complexity. First came the recognition in France that the Stone Age needed subdivision into an old period (Paleolithic) and a new period (Neolithic); this conclusion was based on the recognition that chipped stones antedated polished ones. Further excavations suggested that the Paleolithic–Neolithic boundary corresponded to the termination of the Ice Age and the onset of postglacial conditions. Later, the Mesolithic, or Middle Stone, Age was introduced in the early 20th century to correct a misconception that there had been a hiatus in European human history during the transition from glacial to nonglacial conditions.

In the late 1800s prehistorians reached the high point in their preoccupation with chronological classification. It was then that a whole series of French archaeological sites gave rise to new subdivisions or cultural traditions within the Paleolithic. Such adjectives as Abbevillian (formerly called Chellean), Acheulean, Mousterian, and Aurignacian were introduced, each derived from a different location in France where an important find had been made.

So long as the world of archaeology lay totally in Europe and the Middle East, there was almost an implicit faith that the three-age time scale had a universality to it in both space and time. The pace and course of cultural development have clearly been different in different places, however, and today different time scales are being developed in different parts of the world, each with its own set of period names. In North America, for example, the major time divisions of prehistory bear the names Paleo-Indian, Meso-Indian, and Neo-Indian, and cultural traditions within these broad intervals have been given such names as Clovis and Hopewellian. Multiplying terminology is clearly a danger if it obscures similar sequences of events in different places. Yet it is an obvious necessity if geographical barriers have resulted in many totally unique lines of development. The hope is that recently developed absolute-dating schemes will soon unify global prehistory to the point where the jargon explosion will be curbed. In the meantime, the archaeological time scale of Figure 3 will continue to be the standard for some time to come.

II. Absolute dating

GENERAL CONSIDERATIONS

Meaning of absolute age. Absolute age, the elapsed time since a past event occurred, must be clearly understood in terms of that event. For people, age is almost universally measured from birth, but in some cultures it is conceivable that age might be reckoned from the time of puberty rites. That is, what is implicit in one context may not be so in another. Similarly, when absolute ages are given for an object of geological or archaeological significance, a knowledge of what happened at time zero is of the utmost importance. Failure to be specific in this regard may introduce misunderstanding.

In most cases, the age of a rock is the number of years that have elapsed since it reached its present condition. To say that a rock is 1,000,000,000 years old means that it underwent crystallization, deposition, or its last metamorphism 1,000,000,000 years ago—depending on whether it is an igneous, sedimentary, or metamorphic rock. On the other hand, if the rock is a granite pebble that has been fashioned into an axhead, the most significant age is the time since the tool was made, not the interval since the granite crystallized. Similar comments apply to a piece of wood used for an ax handle; photosynthesis and shaping occurred at different times. In more subtle situations, this sort of distinction may be crucial to understanding or application.

Dating requirements. Essential to dating is a physical record of time. Beginning with such a record, all schemes of absolute dating are comprised of three elements: (1) a discernible starting point that defines the meaning of absolute age, (2) a discernible termination point, and (3) a known rate for spanning the two. A freshly cut tree that exposes a record of time in its cross section of concentric rings provides a good example. The starting point is the inner ring and the termination point the outer ring; the two are spanned at a rate of one ring per year. Hence, 100 rings indicate a tree that grew continuously from a time 100 years ago until recent cutting. The tree is said to be 100 years old, even though all rings but the inner one are younger than this age.

Essential elements in absolute dating schemes

For 100 years to be the correct age, however, two premises that were tacitly assumed must be true. First, there must be a physical record for every year—in other words, the tree is presumed to have deposited rings continuously. Second, the rate of ring addition is assumed to be exactly one per year. How valid these two premises are depends on a detailed examination of the growth patterns of many trees representing different species and different environments. It has been deter-

mined that the two premises are generally true, although exceptions have been noted. A small uncertainty is therefore associated with the 100-year age, its magnitude being determined by species of tree and growth environment. Implicit in this discussion is the assumption that tree growth during observation is typical of that in previous years.

Another assumption involves the date of the start of growth. If a tree stump is produced not by cutting but, rather, by natural causes—such as felling by a heavy wind in the unknown past—a counting of rings would establish only the age of the tree when it was killed, not the time when the tree began to grow. The record of ring widths provides a 100-year chronology of climatic variation in either case, but the chronology obtained from the man-made stump is fixed, whereas that from the wind-created stump is without an anchor in time. It is said to be floating. Floating chronologies are usually far less valuable than those that are fixed.

Questions asked in evaluating dating methods

The foregoing example points to five questions that must be posed in evaluating any method proposed for determining absolute ages: (1) what physical record of time is available? (2) is the record complete? (3) are the end points of the record clearly discernible? (4) what point in time corresponds to the terminal point of the record? and (5) is the rate of buildup or accumulation of the record determinable and, if so, how reliably?

Rate of record accumulation. The physical records used in absolute time measurement leave much to be desired in the way of completeness and clarity. But the crucial problem in absolute dating is generally the rate at which the record accumulated. The usual procedure is to measure the present rate and assume that this has prevailed throughout the period of record accumulation. If there were good theoretical and practical reasons for constancy of rate, this procedure would be satisfactory, but there exist in nature many changing conditions that can alter process rates. It is for this reason that radiometric dating claims such a supreme position; there seems to be no possible way for environmental change to cause a change in the rate of radioactive decay.

Support for the constancy of radioactive change comes from both nuclear theory and experiment. In the laboratory, for example, it is impossible to alter the rate of radioactive decay by any combination of pressure and temperature known to exist within the Earth's crust. The same is true with respect to gravitational, magnetic, and electric fields as well as the chemical state in which a given radioactive element is found. In short, the process of radioactive decay is immutable under all conditions significant to geology and archaeology.

RADIOMETRIC DATING

Radioactivity and radioactive decay. The word radioactivity almost defines itself, indicating a sustained process in which radiation is continuously emitted from certain types of matter. An examination of radioactivity at the atomic level, however, shows that the outward appearance of continuity is merely the combined result of many random atomic events called disintegrations, or decays. When an individual radioactive atom disintegrates, two things happen: (1) one or more particles are ejected from the nucleus of the atom, contributing to the total sample radiation; and (2) the nucleus changes character as a result of particle ejection, becoming a new kind of matter in most cases.

The two results of atomic disintegration

It is impossible to predict the instant when any given radioactive atom will disintegrate. But, when enough radioactive atoms of a certain type are placed together, observation shows that the number of disintegrations per unit time is proportional to the number of radioactive atoms present. The situation is analogous to the death rate among the human population insured by an insurance company. Although it is impossible to predict when a given policy holder will die, the company can count on paying off a certain fraction of beneficiaries every month.

The foregoing behaviour of radioactive atoms can be expressed mathematically as follows:

$$\underset{\text{rate of disintegration}}{R} \quad \underset{\text{is proportional to}}{\propto} \quad \underset{\text{number of radioactive atoms.}}{N} \tag{1}$$

Converting this proportion to an equation incorporates the additional observation that different radioisotopes have different disintegration rates even when the same number of atoms are observed undergoing decay. In other words, each radioisotope has its own decay constant, abbreviated λ (lambda), which provides a measure of its intrinsic rapidity of decay. Proportion (1) becomes:

$$R = \lambda N. \tag{2}$$

Stated in words, this equation says that the rate at which a certain radioisotope disintegrates depends not only on how many atoms of that isotope are present but also on an intrinsic property of that isotope represented by lambda (λ), the so-called decay constant. Values of lambda vary widely—from 10^{20} reciprocal seconds (*i.e.*, the unit of 1/second) for a rapidly disintegrating isotope such as helium-5 to less than 10^{-25} reciprocal seconds for slowly decaying cerium-142.

In the calculus, the rate of decay R in equation (2) above is written as the derivative dN/dt, in which dN represents the small number of atoms that decay in an infinitesimally short time interval dt. Replacing R by its equivalent dN/dt results in the differential equation

$$\frac{dN}{dt} = -\lambda N. \tag{3}$$

Solution of this equation by techniques of the calculus yields one form of the fundamental equation for radiometric age determination,

$$\frac{N}{N_0} = e^{-\lambda t}, \tag{4}$$

in which N_0 is the number of radioactive atoms present in a sample at time zero, N is the number of radioactive atoms present in the sample today, e is the base of natural logarithms (equal to about 2.72), λ is the decay constant of the radioisotope being considered, and t is the time elapsed since time zero.

Two alterations are generally made to equation (4) in order to obtain the form most useful for radiometric dating. In the first place, since the unknown term in radiometric dating is obviously t, it is desirable to rearrange equation (4) so that it is explicitly solved for t. Second, the more common way to express the intrinsic decay rate of a radioisotope is through its half-life (abbreviated $t_{1/2}$) rather than through the decay constant lambda (λ). Half-life is defined as the time period that must elapse in order to halve the initial number of radioactive atoms. The half-life and the decay constant are inversely proportional, because rapidly decaying radioisotopes have a high decay constant but a short half-life. With t made explicit and half-life introduced, equation (4) is converted to the following form, in which the symbols have the same meaning:

Definition of half-life

$$t = \frac{t_{1/2}}{0.693} \times \log_e \left(\frac{N_0}{N}\right). \tag{5}$$

Principles of radiometric dating. Among the 103 known chemical elements, there exist approximately 1,650 different species of atoms known as isotopes. Of this total, about 335 are naturally occurring (as opposed to man-made), and more than 1,350 are radioactive (as opposed to stable). In nature there are about 65 radioactive isotopes. Those most useful in dating are listed in Table 1.

In radiometric dating as currently practiced, the dominant cumulative effect that is measured is the amount of daughter isotope present (the daughter is the isotope produced by decay of some original, long-lived isotope called the parent). Figure 4 summarizes pictorially the four modifications of the basic decay principle central to radiometric dating. The simple hourglass (Figure 4A), which includes the majority of rock-dating methods, is readily related to equation (5) above. Because the half-life

Table 1: Natural Radioactive Isotopes of Possible Use in Absolute Dating

radioisotope	mode of origin*	half-life (years)	particle emitted	daughter
Aluminum-26 (^{26}Al)	cosmogenic	7.4×10^5	beta-plus	magnesium-26 (^{26}Mg)
Argon-39 (^{39}Ar)	cosmogenic	269	beta-minus	potassium-39 (^{39}K)
Beryllium-10 (^{10}Be)	cosmogenic	2.7×10^6	beta-minus	boron-10 (^{10}B)
Carbon-14 (^{14}C)	cosmogenic–man-made	5,730†	beta-minus	nitrogen-14 (^{14}N)
Chlorine-36 (^{36}Cl)	cosmogenic	3.07×10^5	beta-minus	argon-36 (^{36}Ar)
Hydrogen-3 (^{3}H)	cosmogenic–man-made	12.3	beta-minus	helium-3 (^{3}He)
Iodine-129 (^{129}I)	primordial	16×10^6	beta-minus	xenon-129 (^{129}Xe)
Lead-210 (^{210}Pb)	radiogenic	22	beta-minus	bismuth-210‡ (^{210}Bi); ultimately to lead-206
Lutetium-176 (^{176}Lu)	primordial	36×10^9 (?)	beta-minus	hafnium-176 (^{176}Hf)
Potassium-40 (^{40}K)	primordial	1.28×10^9	electron capture beta-minus	calcium-40 (^{40}Ca) and argon-40 (^{40}Ar)
Protactinium-231 (^{231}Pa)	radiogenic	3.24×10^4	alpha	actinium-227‡ (^{227}Ac)
Rhenium-187 (^{187}Re)	primordial	43×10^9 (?)	beta-minus	osmium-187 (^{187}Os)
Rubidium-87 (^{87}Rb)	primordial	50×10^9	beta-minus	strontium-87 (^{87}Sr)
Silicon-32 (^{32}Si)	cosmogenic	650	beta-minus	phosphorus-32‡ (^{32}P); ultimately to sulfur-32
Sodium-22 (^{22}Na)	cosmogenic	2.6	beta-minus	neon-22 (^{22}Ne)
Strontium-90 (^{90}Sr)	man-made	28.9	beta-minus	yttrium-90‡ (^{90}Y); ultimately to zirconium-90
Thorium-230 (^{230}Th)	radiogenic	7.8×10^4	alpha	radium-226‡ (^{226}Ra)
Thorium-232 (^{232}Th)	primordial	14.1×10^9	alpha	radium-228‡ (^{228}Ra); ultimately to lead-208
Uranium-234 (^{234}U)	radiogenic	2.47×10^5	alpha	thorium-230‡ (^{230}Th)
Uranium-235 (^{235}U)	primordial	0.71×10^9	alpha	thorium-231‡ (^{231}Th); ultimately to lead-207
Uranium-238 (^{238}U)	primordial	4.5×10^9	alpha	thorium-234‡ (^{234}Th); ultimately to lead-206
Uranium-238 (^{238}U)	primordial	1.0×10^{16}	spontaneous fission	many fission fragments‡ plus 2–3 neutrons

*Primordial isotopes have existed as long as the Earth; radiogenic isotopes are the products of radioactive decay; cosmogenic isotopes result from cosmic rays impinging on matter; and man-made isotopes are mainly produced during nuclear bomb testing. †See text for variation. ‡Radioactive.

is constant for any particular radioisotope being considered and is known from physical measurement, dating laboratories focus on measurements of the ratio N_0/N, the reciprocal of which is merely the fraction of radioactive atoms still remaining in the sample being dated. When the numerical value of N_0/N is inserted into equation (5) along with the appropriate half-life, the age t is readily calculated.

Although the number (N) of radioactive atoms present in a sample is directly measurable, the comparable number at time zero (N_0) must be arrived at indirectly. In the simple hourglass (Figure 4A), the initial number of radioactive atoms is just the sum of the stable daughter atoms and the radioactive parent atoms now in the sample. This means that analyses for two different sample isotopes are required. As an example, suppose a uranium sample is shown to contain four trillion (4×10^{12}) atoms of uranium-238 and one trillion stable lead-206 atoms. It follows that there were five trillion uranium-238 atoms at time zero and that the ratio (N_0/N) is 5:4, or 1.25. Insertion of 1.25 into equation (5) along with a half-life of 4.5 billion (4.5×10^9) years results in an age of 1.45 billion years.

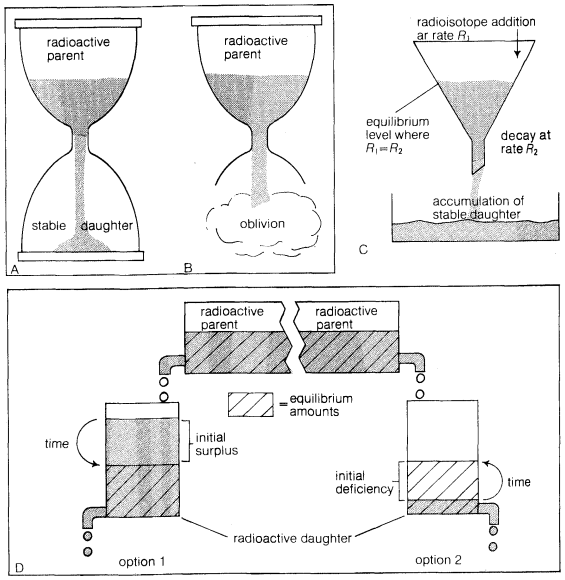


Figure 4: *Principles of various radioactive dating methods:* (A) Simple hourglass. (B) Bottomless hourglass. (C) Funnel at equilibrium. (D) Cascade series returning to equilibrium (two options).

In Figure 4B the so-called bottomless hourglass, equation (5), is again used, but here there is no possibility of analyzing for stable daughter atoms. The reason is simply that these atoms are swallowed up in a sea of primordial isotopes of the same type. Consider the most important example of B, namely, radiocarbon dating, in which carbon-14 decays to nitrogen-14, the dominant isotope of nitrogen. Because nitrogen is quite common and is always present in natural situations in which carbon-14 is disintegrating, nitrogen-14 produced by decay is immediately mixed in with environmental nitrogen-14 and completely dwarfed by it. There is no possibility whatsoever of distinguishing identical isotopes of different origin. What must be done in the case of the bottomless hourglass is to look about for other systems involving the same radioisotope but having essentially zero age. Hopefully, there will be good reason to believe that the sample of unknown age began its life with a similar level of radioactivity. In radiocarbon dating, for example, the carbon-14 level in living material is presumed to represent the time-zero level of older wood and charcoal samples submitted for dating. Such an assumption and others like it are clearly not made without first marshalling evidence that they have a high probability of being valid. Failure of the assumption may introduce a significant error.

In Figure 4C is a third radiometric dating method, one in which a radioisotope is being added to a sample as fast as it is departing by decay. The experience of using a funnel to pour liquid through a small opening is a common one and readily demonstrates that any given equilibrium level in the funnel has a corresponding rate of addition. If that rate increases, the level rises until the outgoing rate also increases enough to balance inflow, thus establishing a new and higher equilibrium level.

In the natural situation, radioactive atoms are added

to a sample through cosmic radiation. Generally, the material being irradiated is meteoritic matter in interplanetary space or rocks from the lunar surface. Mainly composed of high-energy protons, cosmic rays bombard the nuclei of sample atoms with such high intensity that a whole series of so-called spallation reactions occurs. These involve nuclear fragmentation and the creation of over 30 different, brand-new stable and radioactive isotopes. The abundance of each radioisotope builds up to an equilibrium level at which it decays as fast as it forms. From the outset of irradiation, stable daughter products of the various radioisotopes steadily accumulate.

When a meteorite falls or a lunar sample is brought back to Earth, the sample is removed from its radiation environment, and radioisotope synthesis ceases. But if the decay rates of its several radioisotopes are measured immediately, they approximate both the production and decay rates in space. Such rates in combination with a quantitative measurement of the accumulated stable daughter isotopes permit a calculation of age. For example, the age of a sample would be 10^{15} seconds (about 30,000,000 years) if it contained 10^{15} stable daughter atoms built up through a disintegration rate of one per second.

Exposure age of a meteorite

The age described here is a so-called exposure age—how long the sample spent being irradiated in space. Implicit in exposure ages of meteorites is this view of their history: they are fragments of even larger objects broken up by space collisions long ago; at that time, they took on essentially their present configuration and were laid bare to a cosmic-ray flux that has remained constant ever since. Clearly, then, meteorite exposure ages are also breakup ages. For lunar samples, exposure ages presuppose a time when rocks were placed on the Moon's surface in a certain position so as to expose an upper surface to an invariant cosmic-ray flux.

The fourth variation of radiometric dating—Figure 4D—is more subtle and mathematically more complex than the first three cases. This situation is restricted to natural decay chains with several intermediate radioactive daughters—specifically, the chains headed by uranium-238 and uranium-235. In uranium deposits that have been in place for at least several million years, the decay chains exist in what is called secular equilibrium. This may be likened to a cascade series in which water flows at constant rate out the bottom of one tank into the next one and so on down through a number of tanks. Under equilibrium conditions, with water flowing into each tank as fast as it leaves through the bottom, any variation in the size of a drain hole from one tank to another must be compensated for by differences in height of water in the various tanks. The bottom pan will be exceedingly large and without a bottom hole, so that with time it becomes more full.

The water in each tank represents the relative number of atoms of an isotope in the decay chain, the passage of water from tank to tank corresponds to radioactive decay down the chain, the size of hole in each tank is analogous to an individual decay constant or half-life, and the final large tank without a hole represents the steady accumulation of stable daughter atoms. A constant water supply being fed into the cascade parallels the uranium parent's insertion of decay products into the chain.

In the natural situation there is obviously no flow of material. A more appropriate image would have uranium atoms as red balls fixed in a crystal lattice; their successive transitions to different isotopes down the decay chain would then be visualized as a sequence of intermittent colour changes, finally ending with a green ball, symbolic of stable lead.

So long as the equilibrium situation just pictured prevails, there is no chronometer apart from the disappearance of uranium and the accumulation of lead. This is simply the case shown in Figure 4A. The basis of case D, however, is an upset in the equilibrium by some natural process and a subsequent period of return to equilibrium. If the extent of upset is known, how far the system has progressed back to equilibrium is an indication of the time since the upsetting conditions occurred.

The two options of case D show two different ways of upsetting equilibrium. On the left is the situation in which the level in one of the tanks down the cascade is somehow made higher than the equilibrium height. The effect on the tank involved is to make it drain faster than it was draining at equilibrium. Because the inlet rate remains constant, the net result is a gradual drop in the tank level until equilibrium height is restored. In the right-hand option, water is removed from a tank lower down, so that the tank then gains water from above faster than it drains. With time, the level rises to the previous equilibrium position. In either case it is assumed that both the initial displacement from equilibrium and the rate of return to equilibrium are known. Thus, the level at any moment is a key to the time since equilibrium was disturbed.

Time zero in radiometric dating. Although the time-zero events of significance in radiometric dating appear quite diverse, they share one common factor: they all involve the assemblage of a group of atoms into a configuration, which, with the exception of nuclear disintegration effects, is ideally retained to this day. In most cases, the atomic assemblages of interest exist in the solid state, into which they were assembled by such processes as crystallization, surface adsorption, photosynthesis, and animal metabolism. Among these, crystallization is of greatest importance geologically and encompasses such specific mechanisms as the cooling of magma to become igneous rock, metamorphism of preexisting rock material, and organic or inorganic precipitation of matter previously dissolved in different aqueous environments (*e.g.*, lakes, groundwater, or the oceans). In a few situations involving short-lived radioisotopes such as carbon-14 and hydrogen-3 (tritium), the sample material is liquid—specifically, water. At time zero such water was removed from intimate contact with the atmosphere, which was the source of its radioactive components. Table 2 lists the time-zero events for different sample types dated by radioactivity.

Sources of error. If a certain radiometric age proves to be wrong, the error stems from one of three sources: the measurement, interpretation, or assumptions made.

Measurement errors

Measurement errors can be made in many ways. Most of the daughter isotopes of importance are present in exceedingly small amounts—for example, often at microgram levels—and the problem of contamination from dust in the air or impurities dissolved in laboratory reagents can be a very significant one. Human or instrumental error is a constant problem, made even more significant by the low concentrations involved. Contamination and human fallibility are handled not only by taking certain procedural and environmental precautions arrived at through a priori reasoning but also by making duplicate measurements that serve to point up accidental errors. In order to discern systematic errors, interlaboratory comparisons and employment of different dating methods are two possible approaches.

Errors of interpretation

The real value of a radiometric age generally resides in its being a key to the time of a past event. In exploiting this value, however, there exists a danger that what is actually dated and what is thought to have been dated may be two different things. As an illustration of error resulting from interpretation, imagine an ancient group of people who collected driftwood and burned it in their fires. Thousands of years later, an archaeologist collects the charred wood remaining and has it dated by radiocarbon. In the published account of his work he quotes the radiocarbon age as the time when the tribe resided at the site, not the time when the wood was photosynthesized. Although the author may have been unaware of what a radiocarbon age represents, it is more likely that he made a certain assumption without stating it—that the interval between photosynthesis of driftwood and its conversion to charcoal was negligible compared to the measurement error inherent in the radiocarbon age. If in reality the driftwood were 500 years old when burned, he would unknowingly be wrong by

Table 2: Time-Zero Events for Different Sample Types Dated by Radioactivity

sample type	time-zero event	qualifying remarks
Igneous rock	time of crystallization from magma	actually, time zero is the moment the rock cooled to the point at which it became a closed system; although the rock body dated did not experience its time-zero event simultaneously in all parts, the interval is almost always negligible in relation to subsequent time
Metamorphic rock	time of metamorphism	same remarks as for igneous rock
Sedimentary rock	time of sedimentation	repeat of the second remark under igneous rock; those sediment grains derived without alteration from igneous or metamorphic rocks date the time of crystallization or metamorphism, not the time of sedimentary accumulation
Meteorite	time of first exposure to cosmic-ray flux	because meteorites are almost certainly igneous, they also have a crystallization age that must exceed exposure age
Organic material (wood, shell, etc.)	time of plant photosynthesis or animal metabolism	the rate of turnover of different biochemical components of organisms is highly variable; organism lifetime, however, is usually so short as to be essentially a point event relative to time since death (exception: inner rings in large, old trees)
Deep-sea sediment (inorganic fraction)	time of settling through water during which dissolved radionuclides were removed by adsorption on grain surfaces	the organic fraction of deep-sea sediment has the usual time zero of organisms; a primary volcanic ash component in deep-sea sediment has a crystallization time zero essentially identical to that defined by settling
Groundwater, ocean water, and glacier ice	time when water or ice ceased to have contact with atmospheric CO_2	flow patterns of water and ice must be known for time zero to have much meaning

500 years. The error would be solely one of interpretation, having nothing to do with the measurement made or the assumptions upon which the radiocarbon method rests. Similar examples could be cited for other dating methods.

Errors of assumption

Most of the effort that has gone into making radiometric dating a reliable enterprise focusses on overcoming errors of assumption. Of all the assumptions, none is so significant as the one that presupposes a closed system—*i.e.*, that a sample has remained as it was originally assembled except for changes produced by radioactive decay. This is a reasonable premise, but it may not hold for all natural samples.

A second important assumption is that a sample began its life free of daughter atoms. If this were not so, there would be the appearance of a certain age even at time zero. At a later time such a "built-in" age would be the error inherent in a measured age. This second assumption is never valid in the strictest sense, and there are some cases in which it is not even valid practically. A simple chemical formula such as NaCl (symbolizing sodium chloride, common table salt), for example, implies that a search of all salt crystals would uncover atoms of just two elements, sodium and chlorine. But, in the real world of crystals, the substitution of foreign atoms for a few of the dominant ones is the rule rather than the exception. This is particularly true when the medium out of which crystals are forming (*e.g.*, magma or seawater) has great compositional diversity. Of course, crystalline substances vary greatly in their inherent power of excluding foreign atoms during their formation, but no substance forms absolutely pure. In samples used for dating, an isotope present at the time of crystallization becomes significant only when it dwarfs or is comparable in amount to the same isotope accumulated by radioactive decay.

As applied to materials used in radiometric dating, any daughter element present at time zero is labelled with the adjective common. For instance, the lead crystallizing along with uranium in the same mineral would be called common lead. Without corrections, common lead would cause a radiometric age to be too great; its presence must be recognized and corrected for.

Laboratories that measure radiometric ages do not generally measure half-lives. Instead, dating laboratories simply insert into their age calculations whatever half-life is currently held in highest esteem among the physicists who specialize in measuring absolute decay rates. This involves the tacit assumption that the half-life values are correct. In all probability, future physical measurements will not change half-lives beyond the percentages noted in Table 3.

Major dating methods. The decay schemes of concern and the specific material types and ranges appropriate to each major method of radiometric dating are listed in Table 3. The principle behind each method has been discussed earlier (see Figure 4).

Uranium–thorium to helium–lead. Methods that focus on uranium and thorium as geochronometers have been pushed into the background today by the potassium–argon and rubidium–strontium techniques. The reasons for this involve limited applicability and questionable reliability. Uranium and thorium have low average abundances in the Earth's crust—about two parts per million for uranium and seven parts per million for thorium. Although natural processes of enrichment provide higher levels in some rocks, it is only in rare bodies such as pegmatites (very coarse crystalline rocks) and vein deposits that actual igneous minerals of uranium and thorium are found. In common igneous rocks, the two elements exist primarily as substitutes for other atoms in minor minerals such as zircon ($ZrSiO_4$). The net effect of this overall distribution is to limit severely the number of rock types to which uranium–thorium geochronometers can be applied. Nonetheless, the method offers considerable reward, because the occurrence of two uranium isotopes, each providing an independent age, makes available an internal check on reliability. Unfortunately, such checks have

Table 3: Major Methods of Radiometric Dating

parent isotope	daughter isotope	parent half-life (in years)	half-life uncertainty (percentage)	dating principle	approximate dating range (years)	applicable material types*
^{238}U	^{206}Pb	4.51×10^9	1	A	10^7–age of Earth	igneous and metamorphic rocks containing such minerals as zircon, uraninite, monazite, pitchblende, xenotime, samarskite, thorianite, thorite†
^{235}U	^{207}Pb	0.71×10^9	2	A	10^7–age of Earth	
^{232}Th	^{208}Pb	14.1×10^9	1	A	10^7–age of Earth	
U, Th	^{4}He	(as above)	—	A	10^5–age of Earth	magnetite in igneous and metamorphic rock; shells of mollusks and coral (maximum range of 10^7 years)
^{40}K	^{40}Ar	1.28×10^9	3	A	10^5–age of Earth	igneous and metamorphic rocks containing muscovite, biotite, phlogopite, lepidolite, sanidine, hornblende†; sedimentary rocks containing glauconite and sylvite
^{87}Rb	^{87}Sr	50×10^9	5	A	10^7–age of Earth	igneous and metamorphic rocks containing muscovite, biotite, lepidolite microcline†; sedimentary rocks containing glauconite
^{14}C	^{14}N	5,730	1	B	500–50,000	wood, charcoal, peat, shells, charred bones, animal and plant tissue, cloth, tufa, groundwater, ocean water, glacier ice

*Inclusion of a material type means that some successful applications are known, not that success is the general rule. †Minerals in moderately coarse-grained rocks are usually isolated by special techniques following rock crushing. Fine-grained rocks such as rhyolite, basalt, tuff, slate, and phyllite are generally handled whole with no attempt to separate discrete mineral fractions.

painted a generally gloomy picture for those seeking a chronometric tool—albeit a most revealing picture for those interested in geochemical processes. The gloom arises from the general discordance among isotopic ages based not only on uranium-235 and uranium-238 but also on thorium-232. The last radioisotope usually accompanies uranium because of its similar ionic charge and radius.

The problem of discordance in atomic ages

Experience shows that, with the exception of results from the mineral uraninite, the three uranium–thorium–lead ages are almost always different. Why this should be so is the subject of many scientific papers and much experimental work. Clearly, discordance is not usually due to common lead. The presence of such lead is revealed through its nonradiogenic isotope lead-204 and is corrected for by calculations involving its isotopic composition; the latter is usually obtainable from cogenetic lead minerals (*e.g.*, galena) found in physical association with the uranium mineral being dated.

Where discordance generally seems to lie is in the failure of uranium-bearing minerals to be closed systems. This situation appears to prevail whether uranium is the dominant metal in the mineral or is merely a trace constituent. Whatever invalidates the closed-system assumption is beyond direct observation and must be inferred from analytical data on natural and laboratory-treated samples. It seems probable that lead loss is the most significant mechanism causing discordance. Of less certainty is the chronology of lead loss—specifically, whether it involved continuous diffusion beginning at the time of crystallization or was a relatively short-term process active only during a metamorphic event.

The escape of helium is also significant. Derived from alpha particles, helium accumulates within a uranium mineral in a time-dependent manner similar to lead buildup. The main difference is that for every atom of uranium or thorium decaying there are six, seven, or eight helium atoms produced, depending on whether the parent atom is thorium-232, uranium-235, or uranium-238. The escape of some helium is predictable on the basis of its small mass and incompatibility with its environment in a crystal lattice. Furthermore, there is no isotopic way to recognize common helium, for it is also alpha-particle derived. As a consequence, uranium–helium dating has received far less attention than the various lead methods. At the present time, its application to the mineral magnetite seems reliable, provided that acid leaching first removes surficial material not held within the tight magnetite lattice. Increasingly significant is the application of uranium–helium dating to the shells of marine organisms that incorporate small numbers of uranium atoms into their calcium carbonate lattices.

Potassium-40 to argon-40. Unlike uranium and thorium, the element potassium is of high average crustal abundance (about 2.5 percent) and occurs widely at concentrations where analysis is relatively easy. The fact that potassium is radioactive was recognized as early as 1905. Over 30 years went by, however, before potassium-40 was revealed as the radioactive isotope and another 20 years before its half-life was accurately determined. Because two decay modes are possible for potassium-40, one yielding stable argon-40 and the other stable calcium-40, potassium minerals accumulate two daughter elements. These build up in the ratio of eight calcium-40 atoms to one argon-40 atom. Clearly, once the production ratio of the two daughters has been accurately determined, both need not be measured in a radiometric-dating analysis. For example, the detection of eight trillion calcium-40 atoms in a sample implies the presence of one trillion argon-40 atoms and therefore the decay of nine trillion potassium-40 atoms over the life of the sample.

Despite the relatively greater abundance of calcium-40, it is argon-40 that is almost always measured. The reason is simply that the calcium-40 isotope in common calcium is almost always so abundant that it overwhelms radiogenic calcium-40. Only a few potassium-bearing minerals, such as lepidolite and sylvite, form under conditions where the common calcium level is low enough not to mask radiogenic calcium-40.

The two main problems that plague potassium–argon dating are like those of the uranium–lead method—common argon (sometimes called excess, or extraneous, argon) and argon leakage. Because of its inert chemical behaviour, argon tends to be almost totally excluded from most potassium-bearing minerals at the time they crystallize. Even so, minute amounts may enter the lattice and form a significant fraction of the argon-40 in samples younger than 100,000 years. Furthermore, there is no isotopic way to recognize the common argon, for it is pure argon-40 held in the crust since its formation by even earlier potassium-40 decay.

Problems in the potassium–argon method

Still another source of argon contamination is an atmospheric contribution adsorbed on surfaces at the time samples come in contact with air. A significant part of this is removable by heating in a vacuum at a temperature low enough to suppress the escape of radiogenic argon-40. Through measurements of nonradiogenic argon-38 and argon-36, what is left of the atmospheric portion can be recognized and corrected for.

Far more serious than argon contamination is argon leakage. As was the case with the uranium–thorium methods, recognition is the first problem. It centres on age comparisons involving different decay schemes applied to the same rock samples. Specifically, search is made for potassium-bearing minerals associated with uraninite specimens that have shown concordant uranium–thorium–lead ages. The ubiquitous association of rubidium and potassium offers a second comparison of radiometric chronometers. As a result of such comparisons, supplemented by laboratory and field studies on leakage rates, potassium minerals have been evaluated with respect to the seriousness of argon loss. The results are much more encouraging than is true for uranium methods. Micas (biotite, muscovite, phlogopite) are especially good in retaining argon despite a priori arguments to the contrary. Hornblende is even better; potassium feldspar is poorer. Elevated temperatures—in excess of 300° C (about 600° F)—can change the rankings and ultimately cause intolerable leakage in all potassium minerals. It follows that the thermal history of a dated sample is very significant in determining how valid its potassium–argon age is.

Rubidium-87 to strontium-87. The geochemical similarity of rubidium to potassium is most fortunate for radiometric geochronometry. The two elements occur not only in the same rock types but also in the same minerals. Moreover, from silicate meteorites to the various common types of terrestrial igneous rocks, the atomic ratio of potassium to rubidium remains more or less constant at about 600 (about 0.25 if just the two radioisotopes are considered). Consequently, almost every potassium-bearing sample carries two geochronometers within it—one involving the decay of potassium-40, the other based on the decay of rubidium-87. The age comparisons made possible are of incomparable value.

The rubidium–strontium method is not without its problems—specifically, loss of the radiogenic daughter (strontium-87), presence of common strontium, and uncertainty about the half-life. Very little experimental work is available on which to judge strontium-87 retention by minerals. Experience in dating samples by different methods suggests the following order of decreasing retentivity, however: potassium feldspar > muscovite > biotite. Ages that are based on analyses of whole rocks show them to have the highest retention of all, an indication that those strontium atoms that do escape their parent mineral grains do not wander far.

Common strontium is the rule rather than the exception. It is recognized through its nonradiogenic isotope strontium-86 and defined isotopically by measurements of associated rubidium-free minerals such as plagioclase feldspar and apatite. Calculation then suffices to remove the nonradiogenic strontium-87 contribution from the total strontium-87 content of a sample.

The half-life of rubidium-87 has been a thorny problem that is still not fully resolved. The uncertainty centres on the difficulty of detecting every disintegration in a sample of rubidium-87. At a time when physicists were quot-

Half-life of rubidium-87

ing 60 billion years as the half-life, geochronologists were arguing that a 50-billion-year value was needed to make rubidium-87 ages agree with those from the potassium-40 and uranium methods. Recent physical measurements suggest that the geological half-life is closer to the truth, and 50 billion years is now quoted in the chart of nuclides prepared by the U.S. Atomic Energy Commission.

At the present time, applying the rubidium–strontium method to igneous and metamorphic rocks generally involves treating a number of sample analyses in the context of a so-called isochron diagram (Figure 5). This diagram is the graphical expression of the basic dating equation (5) after certain modifications are made. In order to use the isochron diagram, it is necessary first to isolate different potassium minerals from a rock and then to measure each mineral for two ratios: strontium-87/strontium-86 and rubidium-87/strontium-86. Each pair of ratios defines a point on the diagram; if measurements are correct and each mineral has been a closed system, the several points should define a straight line. From the inclination of the line, rock age t is related to the inclination angle A by the equation $\tan A = e^{\lambda t} - 1$. The point of intersection of the line with the vertical axis defines the isotopic ratio of common strontium.

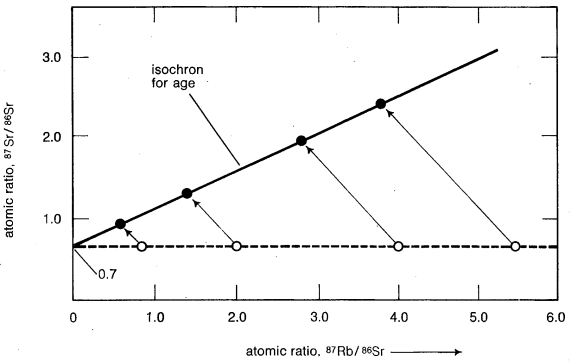


Figure 5: Rubidium–strontium isochron diagram. The four solid circles represent four different rubidium-bearing minerals separated from the same rock and analyzed for the two atomic ratios shown on the axial coordinates. Arrows for each mineral show how subsequent decay of rubidium-87 to strontium-87 changed the ratios from their initial values (open circles) to their final values.

An age determined from an isochron diagram represents the most recent time when all of the strontium in the rock was of the same isotopic composition. Since that time, different rock minerals have changed that composition by decay of their contained rubidium. The greater the rubidium content, the greater the isotopic change of strontium in each mineral. The process responsible for homogenizing the strontium initially was either metamorphism or melting. Without data such as field evidence, however, the exact cause of homogenization may be unknown.

Carbon-14. Radioactive-carbon dating stands in marked contrast to the previously described rock-dating methods. Application of carbon-14 dating is mainly to material synthesized by organisms within the past 50,000 years, and the principle of carbon-14 dating is that of the bottomless hourglass rather than the simple hourglass (see Figure 4).

Discovery of carbon-14

The discovery of natural carbon-14 by the U.S. scientist Willard Libby began with his recognition that a process that had produced radiocarbon in the laboratory was also going on in the Earth's upper atmosphere—namely, the bombardment of nitrogen by free neutrons. Once created in the atmosphere, newly created carbon-14 atoms were presumed to react with atmospheric oxygen to form carbon dioxide (CO_2) molecules. Radioactive carbon thus was visualized as gaining entrance wherever atmospheric carbon dioxide enters—into land plants by photosynthesis, into animals that feed on the plants, into marine waters and freshwaters as a dissolved component, and from there into aquatic plants and animals. In short, all parts of the carbon cycle were seen to be invaded by the isotope carbon-14.

Table 4: Age Concordances Possible Under Ideal Conditions

location	rock dated	mineral dated	dating method	age*
Redstone, N.H.	granite	zircon	^{238}U–Pb	187
		zircon	^{235}U–Pb	184
		zircon	Th–Pb	190
		biotite	K–Ar	182
		biotite	Rb–Sr	185
Spruce Pine, N.C.	granite	zircon	^{238}U–Pb	370
		zircon	^{235}U–Pb	375
		biotite	K–Ar	349
		biotite	Rb–Sr	375
Romteland, Norway	pegmatite	uraninite	^{238}U–Pb	890
		uraninite	^{235}U–Pb	892
		uraninite	Th–Pb	900
Black Hills, S.D.	—	uraninite	^{238}U–Pb	1,610
		uraninite	^{235}U–Pb	1,615
		microcline	Rb–Sr	1,630
		muscovite	K–Ar	1,590
Northern Minnesota	pegmatite	muscovite	K–Ar	2,480
		muscovite	Rb–Sr	2,500
Cooke City, Mont.	pegmatite	uraninite	^{238}U–Pb	2,600
		uraninite	^{235}U–Pb	2,640
		mica	Rb–Sr	2,750
		mica	K–Ar	2,500

*Error on ages is between one and three percent.

Invasion is probably not the proper word for a component that Libby calculated should be present only to the extent of about one atom in a trillion stable carbon atoms. So low is such a carbon-14 level that no one had detected natural carbon-14 until Libby, guided by his own predictions, set out specifically to measure it. His success initiated a series of measurements designed to answer two questions: is the concentration of carbon-14 uniform throughout the plant and animal kingdoms? And, if so, has today's uniform level prevailed throughout the recent past?

After showing the essential uniformity of carbon-14 in living material, Libby sought to answer the second question by measuring the radiocarbon level in organic samples dated historically—materials as old as 5,000 years from sources such as Egyptian tombs. With correction for radioactive decay during the intervening years, such old samples hopefully would show the same starting carbon-14 level as exists today. This was just what Libby's measurements showed. His conclusion was that over the past 5,000 years the carbon-14 level in living materials has remained constant within the 5 percent precision of measurement. A dating method was thus available, subject only to confirmation by actual application to specific chronologic problems.

Since Libby's foundational studies, tens of thousands of carbon-14 measurements of natural materials have been made. Expressed as a fraction of the contemporary level, they have been mathematically converted to ages through equation (5). Archaeology has been the chief beneficiary of radioactive-carbon dating, but late glacial and postglacial chronological studies in geology have also been aided greatly.

Variations of the carbon-14 level in nature

With improvements in accuracy of measurement and ever-mounting experience in applying carbon-14 dating, superior and more voluminous data are accumulating to provide better answers to Libby's original questions. It is now clear that carbon-14 is not homogeneously distributed among today's plant and animal kingdoms. The occasional exceptions all involve nonatmospheric contributions of carbon-14-depleted carbon dioxide to organic synthesis. Specifically, volcanic carbon dioxide is known to depress the carbon-14 level of nearby vegetation; dissolved limestone carbonate occasionally has a similar effect on freshwater mollusks, as does upwelling of deep ocean water on marine mollusks. In every case, the living material affected gives the appearance of built-in age.

In addition to spatial variations of carbon-14 level, the question of temporal variation has received much study. A 2 to 3 percent depression of the atmospheric radioactive-carbon level since 1900 was noted soon after Libby's pioneering work, almost certainly the result of the dump-

ing of huge volumes of carbon-14-free carbon dioxide into the air through smokestacks. Of more recent date is the overcompensating effect of man-made carbon-14 injected into the atmosphere during nuclear-bomb testing. The result has been to raise the atmospheric carbon-14 level by over 50 percent. Fortunately, neither effect is significant in the case of older samples submitted for carbon-14 dating, but there is a clear warning to investigate the past for comparable natural effects. Libby hoped to do just that in his original work on historically dated samples as old as 5,000 years.

Since then, precision of measurement has improved, and tree rings as old as 8,200 years have become available for analysis. The results reveal that the atmospheric radiocarbon level prior to 1000 BC deviates measurably from the contemporary level. In the year 6200 BC it was 8 percent above today's level. In the context of carbon-14 dating, this departure from the present level means that samples with a true age of 8,200 years will be dated by carbon-14 as 7,500 years old. The situation between 5500 BC and 9000 BC is currently being investigated through comparison with varved sediments (sharply defined sedimentary layers deposited in special environments at the rate of two per year). Comparable modes of checking times even more remote than 9000 BC are currently a hope rather than a reality.

The ultimate cause of carbon-14 variations with time has generally been attributed to temporal fluctuations in the cosmic rays that bombard the upper atmosphere and create terrestrial carbon-14. Whatever the cause, however, it is clear that carbon-14 dates lack the accuracy that traditional historians would like to have. Hopefully, there will come a time when all carbon-14 ages rest on firmer knowledge of the sample's original carbon-14 level than is now available. Until then, the inherent error from this uncertainty must be recognized.

Carbon-14 contamination

A final problem of importance in carbon-14 dating is the matter of sample contamination. If a sample of buried wood is impregnated with modern rootlets or a piece of porous bone has recent calcium carbonate precipitated in its pores, failure to remove the contamination will result in a carbon-14 age between that of the sample and that of its contaminant. Consequently, numerous techniques for contaminant removal have been developed. Among them are the removal of humic acids from charcoal and the isolation of cellulose from wood and collagen from bone. Today, contamination as a source of error in samples younger than 25,000 years is extremely rare. Beyond that age, however, the fraction of contaminant needed to have measurable effect is quite small, and, therefore, undetected or unremoved contamination may occasionally be of significance.

Uranium-series disequilibrium. Uranium disequilibrium methods offer the tantalizing prospect of checking carbon-14 dating within its range and then extending this range by an additional 200,000 years. The underlying principle was described earlier in terms of a cascade series returning to equilibrium (see Figure 4D). Currently there are only two series of practical importance, those started by the two major isotopes of the element uranium. The important transitions involve the decay paths of ^{238}U (uranium-238) to ^{206}Pb (lead-206) and ^{235}U (uranium-235) to ^{207}Pb (lead-207), which are (with half-lives in years bracketed):

$$^{238}U\ [4.51 \times 10^9] \rightarrow {}^{234}U\ [2.48 \times 10^5] \rightarrow {}^{230}Th\ [8.0 \times 10^4] \rightarrow \cdots \rightarrow {}^{206}Pb\ \text{(stable)}$$

$$^{235}U\ [7.1 \times 10^8] \rightarrow {}^{231}Pa\ [3.43 \times 10^4] \rightarrow \cdots \rightarrow {}^{207}Pb\ \text{(stable)}.$$

Within uranium-bearing minerals of the Earth's crust a stable condition called secular equilibrium is established within the first few million years after mineral formation. Such a state involves the maintenance of a constant number of atoms of each radioactive member of the two chains shown above, a situation made possible when the decay of each radioisotope is just balanced by formation from its radioactive precursor. In terminology used previously, there exist two cascade series at equilibrium.

By disrupting this equilibrium the dating potential is introduced. What occurs is a two-step sequence that culminates in a solid sample having either a deficiency or a superabundance of uranium relative to other radioactive elements in the chain. Disequilibrium then prevails in the newly formed solid.

Steps in the creation of datable uranium-bearing rock samples

The first step in the creation of datable samples is the breakdown of uranium-bearing rocks by chemical weathering, thereby permitting rainfall to move the two uranium decay series into freshwater or ocean water. Step two is the reassembly of solid samples within these water bodies by one of two mechanisms that create disequilibrium: (1) preferential incorporation of dissolved uranium-238 and uranium-235 into a calcium carbonate solid such as coral, at the same time excluding protactinium-231 and thorium-230 (the latter is sometimes called ionium), or (2) preferential adsorption of protactinium-231 and thorium-230 on suspended clay particles while excluding uranium isotopes. In most cases the radioisotopes are present at levels of one to ten parts per million.

With time the uranium isotopes in calcium carbonate decay into thorium-230 and protactinium-231 and build up these two daughters toward their equilibrium amounts. Before this steady state is reached, the nearness to equilibrium is time-dependent and thus an indicator of carbonate age.

In the case of clay particles, their fate is to settle out on the floor of an ocean or a lake and to undergo decay of adsorbed thorium-230 and protactinium-231. Based on a constant rate of clay sedimentation, the amounts of these two radioisotopes per gram of clay decrease downward because of increasing age. If a zero age is ascribed to surface material, the age of material farther down is obtainable by comparing concentrations of thorium-230 and protactinium-231 within a sample core to those in surface sediment.

The foregoing account describes the uranium-series methods as one would wish them to be rather than as they are. Four assumptions are involved, each of which is generally or frequently invalid: (1) carbonates and clay sediments are completely closed systems, (2) clay sedimentation rates are uniform in each sediment column, (3) all radioactive elements incorporated in clay grains are adsorbed during settling, and (4) there is total exclusion of uranium from clay sediments and of thorium-230 and protactinium-231 from calcium carbonate solids. In addition, a further complicating factor has been ignored, namely, the discovery that uranium-238 and its daughter uranium-234 are usually not at secular equilibrium within natural aqueous solutions or in the precipitates derived from them.

Taken together, these complications explain the great amount of experimental work currently under way in uranium-series dating. The goals are, first, to determine whether there are compensating steps that will get around failures in the above assumptions and, second, to identify those materials and situations that offer the greatest promise of successful dating.

Uranium in carbonates

At present, uranium in carbonates looks very favourable for corals, unfavourable for mollusk shells, and promising for oolites, inorganic marls, and speleothems (*i.e.*, layers in stalagmites). The chief problem centres on the closed-system assumption. Evidence suggesting that a carbonate sample has been a closed system includes: (1) age agreement between the uranium-234–thorium-230 and uranium-235–protactinium-231 clocks as well as with related carbon-14 and potassium–argon dates; (2) uranium-series ages consistent with stratigraphic relationships; and (3) no evidence of calcium carbonate recrystallization that has turned aragonite into calcite.

In the case of clay-sediment ages, a measure of success has come in dating marine sediments from deep-sea cores. Almost total failure, however, has resulted from strict adherence to the assumed constancy in clay sedimentation rate. As an alternative, it is often assumed that radiogenic thorium-230 and the main thorium isotope thorium-232 are adsorbed from seawater in a constant ratio. Because thorium-232 has such a long half-life (14,000,000,000 years), its decay with depth in a

sediment core is negligible. Consequently, an index of sediment age is derived from the extent to which the surface value of thorium-230–thorium-232 has decreased with depth. A somewhat better approach has been to focus on how the ratio protactinium-231–thorium-230 changes down a core. Using this ratio for dating marine sediments requires the fewest and the most probable assumptions concerning temporal constancy of adsorption variables. The greatest dating success has thus been obtained through protactinium-231–thorium-230 ratios, subject mainly to uncertainties in how closely marine sediments approximate closed systems.

An unsuspected disequilibrium in the uranium series was discovered in the early 1960s. Natural waters and precipitated solids derived from them were shown to have an excessive amount of uranium-234 relative to the parent uranium-238. In contemporary ocean water as well as marine carbonates, this excess is uniformly 15 percent, whereas much greater values have been found in waters and precipitates of the continents. The mechanism responsible for disequilibrium is preferential leaching of uranium-234 by groundwater acting upon uranium-bearing surface rocks. Such rocks have the two uranium decay chains at equilibrium. When a uranium-238 atom ejects an alpha particle to produce what soon becomes a uranium-234 atom, however, there is a recoil that loosens the uranium-234 atom in its lattice position and makes it more susceptible to being leached away.

Although the excess of uranium-234 has introduced a complication in uranium-series methods, it has simultaneously offered another possibility of dating carbonates. Dating hinges on a firm knowledge of the percent of excess at time zero, for it is from the time-zero value that uranium-234 drops toward equilibrium (that is, toward no excess). At present, the 15 percent excess shown by all ocean-water samples gives hope for dating marine carbonates. It is not at all certain, however, whether this uniformity in space is matched by uniformity in time. In other words, today's oceanic value of 15 percent excess uranium-234 may not have prevailed in the past. This uncertainty is typical of the uranium-series methods; they all require comparisons at many points in order to provide confidence in the ages obtained.

Spontaneous-fission tracks. In place of simple radioactive decay, a small fraction of uranium-238 atoms undergo a process known as spontaneous fission. The effect of a single fission event is to create a minute region of crystal disruption called a track. The density of spontaneous-fission tracks in uranium-bearing samples is dependent on the concentration of uranium and the duration of fission. It follows, then, that measurements of track density and uranium content permit calculation of sample age. Because uranium-235 fissions readily when bombarded with slow neutrons in a nuclear reactor, the track density so induced is a measure of uranium content for any particular neutron dose. Consequently, a complete fission-track age measurement requires just two track-density determinations combined with a measured neutron dose.

Steps in fission-track dating

The actual procedure of fission-track dating begins with preparation of a smooth sample through either polishing or mineral cleavage. Immersion in an etching solution follows, resulting in enlargement of spontaneous-fission tracks to the point of microscopic visibility. Because of chemical variation from sample to sample, a number of acids and alkalis find use in etching. Hydrofluoric acid is used for most silicates, although strong potassium and sodium hydroxides are occasionally employed. The mineral apatite is etched with nitric acid, calcite with hydrochloric acid. Trial and error determine the optimum etching conditions of time (seconds to hours), temperature (ambient to several hundred degrees), and solution strength (usually concentrated).

In order to count tracks, a high-powered microscope is required. This means magnifications of from 250 to 1,800. How much sample area must be examined for tracks depends on their abundance. The fewer the tracks, the more area must be counted in order to obtain a statistically reliable track density.

After a count of tracks made by spontaneous fission, the sample is placed in a nuclear reactor for a period of hours or days until from 10^{16} to 10^{18} slow neutrons per square centimetre have passed through it. Etching is repeated, followed by counting of the new tracks that formed by the induced fission of uranium-235. Values of the two track densities and the neutron dose are inserted into the following equation in order to obtain sample age: age (in years) $= 6 \times 10^{-8} \times N \times (Ds/Di)$, in which N is the total neutron dose expressed as neutrons per square centimetre, Ds is the track density for spontaneous fission of uranium-238, and Di is the track density for induced fission of uranium-235.

Many minerals and rocks are datable by fission-track counting. These include the common silicate minerals of igneous rocks (different micas, feldspar, and hornblende) as well as accessory minerals such as zircon and apatite. Natural glasses such as obsidian, tektites, and basaltic glass also have been used successfully. Even man-made glasses have been dated, especially those in which uranium minerals were added to give colour. Satisfactory results on meteorites have also been obtained.

The dating range of fission-track dating is governed mainly by the uranium content of samples. At a uranium level of one part per million, for example, samples as young as 300,000 years can be measured by track counting for just one hour. This can be lowered to 8,000 years if the investigator is willing to invest a 40-hour week in counting. Most natural glasses contain uranium in parts-per-million-amounts, while zircon and apatite may go as high as 100–1,000 parts per million. Man-made glasses coloured with 10 percent uranium are easily dated even when they are as young as 30 years. At the other end of the range, rock ages of 1,000,000,000 years have been obtained.

Problems in fission-track dating

The major drawback to fission-track dating lies in the tendency of tracks to fade away when elevated temperatures are maintained for long periods. One-hour experiments in the laboratory show that different materials have different critical temperatures at which tracks quickly disappear. The temperature extremes vary from 60° C (140° F) for the uranium mineral autunite up to 1,050° C (1,920° F) for quartz. Held at 20° C (68° F) for its complete lifetime, any substance with a critical temperature exceeding 400° C (750° F) should be able to preserve fission tracks for periods comparable to the age of the Earth. The most promising rocks, therefore, are those that have always been at or near the Earth's surface—that is, volcanic rocks. By contrast, rocks originating in the hotter interior, such as granite and metamorphic rocks, almost universally show smaller fission-track ages than those obtained through rubidium–strontium and potassium–argon methods. Confined to its proper niche, however, fission-track dating has made and will continue to make important contributions to chronologic studies.

Thermoluminescence dating. Hope rather than accomplishment mainly characterizes the status of thermoluminescence dating at the present time. The phenomenon of thermoluminescence is the emission of light from heated crystals previously exposed to radiation. The longer the time allowed for radiation to build up an inventory of trapped electrons, the greater the light energy released when the crystals are heated and electrons return to their normal positions. In theory, it is possible to date the onset of radiation by measuring the amount of light released, the total radiation dose, and the rate at which that dose was applied.

In applying thermoluminescence dating to such materials as pottery, bones, and minerals, the sample material is first ground to a powder. A gram or less is then spread out thinly over an electric heating plate. Next, the sample is heated at 20° C (36° F) per second within a nitrogen atmosphere, during which time the light emitted is measured by a photomultiplier tube. The final record is a graph of light intensity plotted against sample temperature. Total glow is usually taken as the light emitted between 350° and 450° C (650° and 850° F). Finally, radioactive impurities responsible for the glow are mea-

sured by standard radiochemical techniques applied to another portion of the sample.

Assumptions of thermoluminescence dating

Three assumptions are basic to thermoluminescence dating. The first is that the glow recorded is an accurate measure of age. This requires that no light be lost previous to laboratory heating and that nothing but sample radiation is responsible for the glow recorded. Here the thermal, chemical, and physical histories of the sample are of central importance. If the sample spent a significant part of its past at depths exceeding several thousand feet, the elevated temperature might have caused a slow drain of light. Sample recrystallization would cause a similar effect. Furthermore, luminescence from both chemical reactions and sample grinding may occasionally augment the glow of radiation-induced thermoluminescence.

The second assumption is that the laboratory radiation dose simulates the natural dose in all ways except rate, which is patently false. Natural radiation is a mixture of alpha, beta, and gamma radiation from impurities within the sample, whereas the artificial source is outside the sample and generally consists of cobalt-60 gamma rays or strontium-90 beta rays. Despite these differences, there often are ways to adjust for them and thus secure acceptable ages.

Only the total radiation dose is considered to be an influence on sample glow. The rate of radiation is supposedly immaterial. Although almost certainly not true in a rigid sense, this assumption cannot be far from the truth for those ages that receive independent confirmation.

Where thermoluminescence dating appears to have the greatest potential is in dating pottery, fossil bones, and teeth. Young deposits of volcanic ash also show promise, provided that small quartz and feldspar crystals (about one millimetre in diameter) can be isolated. To date, a number of potsherds as old as 8,000 years have given reliable ages when checked by associated carbon-14 dates. Dating of bones and teeth is just beginning but may one day be of great value. Unfortunately, bone samples younger than 100,000 years are currently out of reach because of chemiluminescence originating from residual organic matter. Successful extraction of the organic fraction might lower the dating limit to a point at which checks with carbon-14 dating become possible.

Minor dating methods. As analytical instrumentation becomes more sophisticated and scientific understanding improves, putting the label minor on a particular dating method may ultimately be shown quite presumptuous. Any review, therefore, must be considered as a status report prefaced with the qualifying words "as things now stand."

Table 5 contains a brief summary of a number of radiometric dating methods that thus far have had only limited application. Some of the reasons for such limitations are as follows: (1) uncertain half-life; (2) low concentrations resulting in difficult analysis and large errors; (3) theoretical uncertainties in the specific geochemical pathways followed by parent or daughter isotopes; and (4) haphazard contributions of man-made isotopes to the atmosphere and hydrosphere.

Three additional notes are in order. First, there is nothing minor about the importance of estimating the interval between synthesis of the chemical elements and their incorporation into crystallizing meteorites. The fact that Table 5 includes the methods behind such estimates merely reflects their limited application and their order-of-magnitude results. Secondly, the isotopes used in estimating meteorite exposure ages are so numerous that almost all have been excluded from Table 5. Several such isotopes are listed in Table 1, however (*e.g.*, ^{22}Na [sodium-22], ^{26}Al [aluminum-26], ^{36}Cl [chlorine-36]). Finally, although man-made fission products such as cesium-137 and strontium-90 have been useful in defining motions and rates within the atmosphere and oceans, none has been included in Table 5.

NONRADIOMETRIC DATING

In addition to radioactive decay, many other processes have been investigated for their potential usefulness in absolute dating. Unfortunately, they all occur at rates that lack the universal consistency of radioactive decay. Sometimes human observation can be maintained long enough to measure present rates of change, but it is not at all certain on a priori grounds whether such rates are representative of the past. This is where radioactive methods frequently supply information that may serve to calibrate nonradioactive processes so that they become useful chronometers. Nonradioactive absolute chronometers may conveniently be classified in terms of the broad areas in which changes occur—namely, geological and biological processes, which will be treated here.

Geological processes as absolute chronometers. *Weathering processes.* During the first third of the 20th

Table 5: Minor Methods of Radiometric Dating

parent isotope	daughter isotope	parent half-life (in years)	dating principle	dating range (in years)	chronometric applications	problems and limitations
^{187}Re (rhenium)	^{187}Os (osmium)	4.3×10^{10} ($\pm 0.5 \times 10^{10}$)	A	40×10^6 up to the oldest samples	dating iron meteorties and the mineral molybdenite	uncertain half-life uncertain crustal distribution coupled with low abundance even in geochemically favorable minerals
^{176}Lu (lutetium)	^{176}Hf (hafnium)	3.3×10^{10} ($\pm 0.5 \times 10^{10}$)	A	50×10^6 up to the oldest samples	dating rare-earth minerals, such as gadolinite, that are present in pegmatites	uncertain half-life there are few minerals rich enough in lutetium to measure the minute amounts of radiogenic hafnium
^{210}Pb* (lead)	^{210}Bi (bismuth)	22	B	last few hundred years	dating glacier ice samples	short dating range downward migration of ^{210}Pb within ice owing to surface melting
^{3}H (hydrogen)	^{3}He (helium)	12.3	B	last hundred years or so	dating old wines deciphering ground water movements	contaminating effect of relatively great amounts of artificial ^{3}H produced in nuclear bomb testing
^{10}Be (beryllium)	^{10}B (boron)	2.7×10^6	B	0.5×10^6 up to 10×10^6	determining cosmic-ray exposure ages of iron meteorites and times since their fall to Earth determining deep-sea sedimentation rates	exceedingly low levels of ^{10}Be resulting in difficult analysis and high analytical uncertainties
^{32}Si (silicon)	^{32}P (phosphorus)	650	B	last few thousand years	determining ocean water mixing rates determining deep-sea sedimentation rates dating glacier ice samples	uncertain distribution pathways exceedingly low levels of ^{32}Si resulting in difficult analysis and high analytical uncertainties
^{129}I (iodine)	^{129}Xe (xenon)	16×10^6	inverse of B	—	determines the interval between element synthesis and age of meteorite crystallization—in other words, the age of the elements	uncertainties in theoretical assumptions concerning how, when, and in what relative amounts the chemical elements were made
^{244}Pu (plutonium)	several isotopes of xenon	82×10^6	inverse of B	—		

*The lead-210 in this method is derived from the radioactive decay of the gas radon-222, which escapes into the atmosphere from uranium-bearing surface rocks and within a few days becomes lead-210.

century, several presently obsolete weathering chronometers were explored. Most famous was the attempt to estimate the duration of Pleistocene interglacial intervals through depths of soil development. In the American Middle West, thicknesses of gumbotil and carbonate-leached zones were measured in the glacial deposits (tills) laid down during each of the four glacial stages (see Figure 3). Based on a direct proportion between thickness and time, the three interglacial intervals were determined to be longer than postglacial time by factors of 3, 6, and 8. To convert these relative factors into absolute ages required an estimate in years of the length of postglacial time. When certain evidence suggested 25,000 years to be an appropriate figure, factors became years—namely, 75,000, 200,000, and 150,000 years. And, if glacial time and nonglacial time are assumed approximately equal, the Pleistocene Epoch lasted about 1,000,000 years (the presently accepted duration of the Pleistocene is from 2,500,000 to 10,000 years ago, approximately).

Obsidian hydration

Only one weathering chronometer is employed widely at the present time. Its record of time is the thin hydration layer at the surface of obsidian artifacts. Although no hydration layer appears on artifacts of the more common flint and chalcedony, obsidian is sufficiently widespread that the method has broad application.

In a specific environment the process of obsidian hydration is theoretically described by the equation $D = Kt^{1/2}$, in which D is thickness of the hydration rim, K is a constant characteristic of the environment, and t is the time since the surface examined was freshly exposed. This relationship is confirmed both by laboratory experiments at 100° C (212° F) and by rim measurements on obsidian artifacts found in carbon-14 dated sequences. Practical experience indicates that the constant K is almost totally dependent on temperature and that humidity is apparently of no significance. Whether in a dry Egyptian tomb or buried in wet tropical soil, a piece of obsidian seemingly has a surface that is saturated with a molecular film of water. Consequently, the key to absolute dating of obsidian is to evaluate K for different temperatures. Ages follow from the above equation provided there is accurate knowledge of a sample's temperature history. Even without such knowledge, hydration rims are useful for relative dating within a region of uniform climate.

Like most absolute chronometers, obsidian dating has its problems and limitations. Specimens that have been exposed to fire or severe abrasion must be avoided. Furthermore, artifacts re-used repeatedly do not give ages corresponding to the culture layer in which they were found but instead to an earlier time, when they were fashioned. Finally, there is the problem that layers may flake off beyond 40 microns (0.002 inch) of thickness—that is, over 50,000 years in age. Measuring several slices from the same specimen is wise in this regard, and such a procedure is recommended regardless of age.

Erosional processes. It is beyond the scope of this article to describe the reasoning and results of quantitative geomorphology, that branch of geology directed toward estimating the pace at which landforms develop and continental surfaces are lowered by erosion. One example will suffice to illustrate the dating possibilities inherent in processes of erosion, however. That case is the cutting of a deep gorge by recession of the lip of Niagara Falls.

Over the past 100 years of observation, the falls have retreated at a rate of several feet per year. The entire gorge, therefore, represents 25,000 years' worth of erosion, provided that today's rate has always prevailed. If the onset of cutting occurred very shortly after departure of the most recent ice sheet, postglacial time has lasted for 25 millennia.

This period of time was the key to assigning absolute durations to interglacial ages. But all of this reasoning has been shown wrong by carbon-14 dating, which places the departure of the most recent glacier at around 11,000 years ago at this latitude. Either today's rate of falls retreat is unusually low, or part of the gorge was cut before postglacial time began (see further WATERFALLS).

Accumulational processes. Sediment in former or present water bodies, salt dissolved in the ocean, and fluorine in bones are three kinds of natural accumulations and possible time indicators. To serve as geochronometers, the records must be complete and the accumulation rates known.

The fossiliferous part of the geologic column includes perhaps 400,000 feet of sedimentary rock, if maximum thicknesses are selected from throughout the world. In the late 1800s, attempts were made to estimate the time over which it formed by assuming an average rate of sedimentation. Because there was great diversity among the rates assumed, the range of estimates was also large—from a high of 2,400,000,000 years to a low of 3,000,000 years. Despite this tremendous spread, most geologists felt that time in the hundreds of millions of years was necessary to explain the sedimentary record.

If the geological column were made up entirely of annual layers, its duration would be easy to determine. Limited sedimentary deposits did accumulate in this way, and they are said to be varved; one year's worth of sediment is called a varve, and, in general, it includes two laminae per year.

Varve dating

Varves arise in response to seasonal changes. New Mexico's Castile Formation, for example, consists of alternating layers of gypsum and calcite that may reflect an annual temperature cycle in the hypersaline water from which the minerals precipitated. In moist, temperate climates, lake sediments collecting in the summer are richer in organic matter than those that settle during winter. This feature is beautifully seen in the seasonal progression of plant microfossils found in shales at Oensingen, Switzerland. In the thick oil shales of Wyoming and Colorado, the flora is not so well defined, but layers alternating in organic richness seem to communicate the same seasonal cycle. These so-called Green River Shales also contain abundant freshwater-fish fossils that confirm deposition in a lake. At their thickest, they span 2,600 vertical feet. Because the average thickness of a varve is about 1/2,000 of a foot, the lake is thought to have existed for more than 5,000,000 years.

Each of the examples cited above is of a floating chronology—that is, a decipherable record of time that was terminated long ago. In Sweden, in contrast, it has been possible to tie a glacial varve chronology to present time and so create a truly absolute dating technique. Where comparisons with radiocarbon dating are possible, there is general agreement (see further VARVED DEPOSITS).

Fluorine dating

As early as 1844, an English chemist named Middleton claimed that fossil bones contain fluorine in proportion to their antiquity. This idea is sound in principle, provided that all the other natural variables remain constant. Soil permeability, rainfall, temperature, and the concentration of fluorine in groundwater all vary with time and location, however. Fluorine dating is therefore not the simple procedure that Middleton envisioned.

Still, the idea that hydroxyapatite in buried bone undergoes gradual change to fluorapatite is a correct one. In a restricted locality where there is uniformity of climate and soil, the extent of fluorine addition is at least a measure of relative age and has been so used with notable success in dating certain homonid remains. Both the Piltdown hoax, for example, and the intrusive burial of the Galley Hill skeleton were exposed in part by fluorine measurements. Supplementing them were analyses of uranium, which resembles fluorine in its increase with time, and nitrogen, which decreases as bone protein decays away.

Fluorine changes could conceivably be calibrated if bone samples were found in a radiometrically dated sequence. But conditions governing fluorine uptake are so variable even over short distances that it is risky to use fluorine content as an absolute chronometer much beyond the calibration site itself. In short, fluorine dating is not now and probably never will be an absolute chronometer. Even when used in relative dating, many fluorine analyses on diverse samples are needed, and these must be supplemented by uranium and nitrogen measurements to establish confidence in the chronological conclusions.

Geomagnetic variations. Based on three centuries of direct measurement, the Earth's magnetic field is known to be varying slowly in both its intensity and direction. In fact, change seems to have been the rule throughout all of the Earth's past. Magnetic minerals in rocks and in articles of fired clay provide the record of ancient change, for they took on the magnetic field existing at the time of their creation or emplacement.

Archaeomagnetic dating

The technique called archaeomagnetic dating rests on a carbon-14 calibration of secular variations that have been determined from fired clay objects. In certain areas tree-ring dating and historic records also provide calibration ages. Material from ancient kilns and fire pits is the best source of data, because these are stationary and thus preserve the directional character of the magnetic field existing at the time they were last heated up. Measurements on several samples are averaged to obtain the necessary magnetic values indicative of time. At present, much work remains in calibrating curves for the many areas in which dating is desired. The goal is a dating method useful as far back as the Neolithic Period (about 10,000 years ago).

Polar reversals were originally discovered in lava rocks and since have been noted in deep-sea cores. In both cases the time dimension is added through radiometric methods applied to the same materials that show the reversals. Potassium–argon is the commonest chronometer used. A so-called paleomagnetic time scale has been proposed along the line of the geologic time scale; time divisions are called intervals, or epochs (see further ROCK MAGNETISM).

Biological processes as absolute chronometers. *Tree-ring growth.* In the early 1900s an Arizona astronomer named Andrew E. Douglass went looking for terrestrial records of past sunspot cycles and not only found what he sought but discovered a useful dating method in the process. The focus of his attention was on the growth rings in trees—living trees, dead trees, beams in ancient structures, and even large lumps of charcoal.

The key documents for tree-ring dating (or dendrochronology, as it is more formally called) are those trees that grow or grew where roots receive water in direct proportion to precipitation. Under such a situation, the annual tree rings vary in width as a direct reflection of the moisture supplied. What is important in tree-ring dating is the sequence in which rings vary. Suppose, for example, that a 100-year-old tree is cut down and its ring widths are measured. The results can be expressed graphically, and, if a similar graph were made from a small stump found near the 100-year-old tree, the two graphs could be compared until a match of the curves was obtained. The time when the small stump was made would thereby be determined from the position of its outer ring alongside the 100-year record.

Not every tree species nor even every specimen of a suitable species can be used. In the American Southwest, success has been achieved with yellow pine, Douglas fir, and even sagebrush. Unfortunately, the giant sequoia of California does not live in a sufficiently sensitive environment to provide a useful record. The even older bristlecone pine in California's White Mountains does have a climate-sensitive record, but its area of growth is so limited and so inaccessible that no bristlecone specimens have so far appeared in archaeological sites. Despite this shortcoming, dead bristlecone pine trees are presently providing rings as old as 8,200 years for dating by carbon-14. The purpose is to check the carbon-14 method.

Coral growth. Certain fossil corals have long been used to date rocks relatively, but only recently has it been shown that corals may also serve as absolute geochronometers. They may do so by preserving a record of how many days there were in a year at the time they were growing. The number of days per year has decreased through time because the rate of rotation of the earth has decreased; geophysical evidence suggests that days are currently lengthening at the rate of 20 seconds per million years. If this were typical of the slowdown during the past, a year consisted of 423 days about 600,000,000 years ago.

Readings from horn corals

It is thought that horn corals indicate the number of days per year by means of their exceedingly fine external ridges of calcium carbonate, each of which is thought to represent a day's growth. Several hundred of the fine ridges also seem to cluster as a unit that presumably corresponds to one year. In certain modern West Indian corals the number of fine ridges in a presumed annual increment is approximately 360, suggesting that coral patterns are being properly interpreted.

Not many fossil corals are in a state of preservation that permits the counting of ridges. But those that are give encouragement that an absolute chronometer may be in the offing. Several Middle Devonian corals indicate between 385 and 410 ridges, with an average of about 400. It remains to be seen whether this method of dating, so elegant in concept and so simple in application, will blossom or wither away in the years to come.

APPLICATIONS OF ABSOLUTE DATING

The geological time scale. At the time radioactivity was discovered in 1896, the relative geologic time scale was essentially in the form shown in Figure 2—minus the numbers, of course. The problem was to obtain those numbers. Because the relative time scale is based on the geologic column, the obvious first approach is to sample the column at intervals and date the material using radiometric techniques. Unfortunately, this approach is limited, because the geologic column is almost totally composed of sedimentary rocks, and ironically these are least amenable to radiometric dating.

Use of glauconite in sandstone samples

The picture is not totally bleak, however. In the first place, there is promise in the mineral glauconite, a green potassium-bearing component in a number of fossiliferous sandstones with ages ranging from Cambrian (500,000,000 to 570,000,000 years ago) to the present. Thought to precipitate from seawater, glauconite gives radiometric ages that presumably measure the time of sedimentation. Clouding the picture, however, is evidence that some glauconite samples gradually lose radiogenic argon. This means that their ages are suspect and define only minimum values.

Of more importance than glauconites are volcanic rocks that interrupt the sedimentary column. Once they were ash-falls or lava flows, now they are layers of tuff or rhyolite interbedded with sedimentary rocks. Because they are well defined stratigraphically and generally provide reliable radiometric ages, they constitute some of the best reference points in the whole time scale. The main problem is that they are rare.

In the absence of interlayered datable rocks, the procedure changes to setting maximum or minimum ages on fossil-bearing sediments that have a clear-cut age relationship with igneous bodies. If a dike cross-cutting a Middle Cambrian shale has a potassium-argon age of 400,000,000 years, the shale is at least 400,000,000 years old. Elsewhere the shale may be eroded so as to reveal the granite on which it was originally deposited. If the granite is 800,000,000 years old, the shale age is then bracketed between 400,000,000 and 800,000,000 years. Pinpointing the age further may require dating a multitude of samples until finally an approximate numerical age can be assigned to the Middle Cambrian.

The age data shown on the geological time scale will probably change only in minor detail in the future. Where major change will occur through future radiometric dating is in the vast reaches of Precambrian time, now in desperate need of a time framework like that given by fossils to the last half-billion years.

The age of the Earth. The age of the Earth is a single point in the absolute geological time scale but one so important philosophically as to merit special consideration. First, it is possible to set upper and lower limits on the Earth's age. The minimum value is obviously the age of the oldest rock on Earth, about 3,500,000,000 years. For a maximum age, a more theoretical approach is taken; it is based on the belief that at the time of element synthesis prior to the existence of the Earth as a discrete planet, the two uranium isotopes were approximately equal in abundance. Because uranium-235 decays faster

Minimum and maximum ages of the Earth

than uranium-238, the ratio is now one to 137.7 rather than the original one-to-one. Calculation shows that 6,000,000,000 years are necessary to bring about the change. Hence, an earth age in the range 3,500,000,000 to 6,000,000,000 years is indicated.

If meteorites and the Earth had similar beginnings—and all evidence is consistent with that assumption—then the age of meteorites is relevant here. On the basis of many measurements using several radiometric methods, meteorite ages fall very close to 4,500,000,000 years. This value almost splits the range given above.

Finally, the isotopic composition of terrestrial lead provides valuable evidence. Of the four lead isotopes, only lead-204 has no radioactive precursor; the other three are produced, in part, by the radioactive decay of uranium-235, uranium-238, and thorium-232. Over time the isotopic composition of lead has therefore changed appreciably. To go from primordial lead (represented by that in iron meteorites) to modern lead (represented by that in recent deep-sea sediments) requires 4,500,000,000 years of decay by the uranium and thorium in the crust.

The conclusion, then, from several lines of evidence is that the age of the Earth—that is, the time since it differentiated into layers of core, mantle, and crust—is about 4,500,000,000 years.

Sea-floor spreading. Although the idea of drifting continents has been debated for decades, discoveries at sea since 1960 appear to have supplied the long-missing element, a mechanism that would raft land masses about. That mechanism is called sea-floor spreading.

Magnetic surveys at sea have revealed patterns of stripes denoting high and low values of magnetic intensity. Especially noteworthy is a survey south of Iceland, where the stripes run parallel to the rift zone at the crest of the Mid-Atlantic Ridge. Moreover, the rift divides the pattern into mirror-image halves. This is thought to have been caused by upwelling of molten material at the rift followed by sideward spreading. The stripes of uniform magnetic intensity are like time lines and their magnetic contrast with adjacent bands is ascribed to repeated reversals of the Earth's magnetic field. It is noteworthy that the first few stripes adjacent to the Mid-Atlantic rift have widths in proportion to the corresponding epochs of the paleomagnetic time scale. When time from the scale is combined with the widths from the magnetic pattern, the result is a velocity of about two centimetres (0.8 inch) per year. On the premise that spreading has been constant at two centimetres per year, distance from the Mid-Atlantic rift becomes a direct indicator of age. Thus, some 80 magnetic bands spanning 1,600 kilometres (1,000 miles) are interpreted as 80 polar reversals in the past 80,000,000 years.

Evidence of spreading from deep-sea cores

An independent check on the foregoing reasoning has resulted in dramatic confirmation. Deep-sea drilling has brought up complete sediment cores at various places in the Atlantic Ocean. If sea-floor spreading is real, bottom sediment in each core was deposited at the time the underlying basalt began spreading outward from the rift. Thus, index microfossils in basal sediment should increase in age the farther a core was taken from the rift. This is exactly what is found. Furthermore, absolute ages of the microfossils agree with ages based on a spreading rate of two centimetres per year (see further SEA-FLOOR SPREADING).

Lunar history. Before man set foot on the Moon, its history had been a matter of study for a number of years. Photographs through telescopes and later from Ranger and Orbiter unmanned spacecraft are still the source for most of the raw data from which astrogeologists, as they are known, seek to determine lunar events and their sequence. Their relative dating techniques are the tried and true principles of superposition and cross-cutting. Lunar seas appear to be great lava flows that lap over the base of adjacent highlands and almost bury certain craters, thus being younger than both. Rilles, steep-walled valleys resembling Grand Coulee in the state of Washington, are clearly younger than the rock through which they cut. So are escarpments that appear to be faults in the lunar crust. At places there are overlapping craters, the younger having removed a rim section from the older. Giant rays of light-coloured material radiate out from a number of fresh-appearing craters and so must be younger than the features they lie across. Moreover, craters run a continuum from very sharp to very subdued and thus provide relative ages for the surfaces on which they have formed.

Gradually, astrogeologists have been able to piece together the sequence of inferred events for much of the lunar face that perpetually is toward the Earth. Like their geological forbears, they have proposed a relative time scale complete with names such as Imbrian, Eratosthenian, and Copernican, terms originated in accord with the geographic tradition of Earth-based stratigraphy. Maps of most of the Moon's front face are presently available to show the different rock units, their presumed origins, and their relative ages.

Without an atmosphere, lunar history is clearly different from that on Earth. Meteorite impact is inferred to be the main agent of change, excavating in one place and dispersing the material outward in all directions. Volcanism is thought to be the agent responsible for some craters and domes as well as sheets that end in lobate forms typical of lava flows. Weathering is a fact on the Moon, but so far its complete nature is beyond full understanding. Micrometeoroid bombardment as well as cosmic radiation may be significant in this regard.

The lunar rock samples

The rocks obtained from the Moon are very old. Basalts from Apollo 11 show potassium–argon and rubidium–strontium ages of approximately 3,700,000,000 years. The Apollo 12 basalts are about 400,000,000 years younger. Certain materials from both Apollo 11 and 12 give ages approximating 4,600,000,000 years, however. Rocks from the Fra Mauro region collected on the Apollo 14 mission have been dated at 3,800,000,000 years, whereas the so-called genesis rock of Apollo 15 gave an age of 4,200,000,000 years. Although it is far too early to understand the full meaning of lunar rock ages, one thing is clear: most of the events that brought the Moon to its present state occurred within the first 1,000,000,000 years of its history. The history of the Moon is slowly emerging through the application of the relative and absolute dating principles developed in studying the earth. There is no reason why they will not succeed in unlocking the secrets of Mars, and perhaps of other planets as well.

BIBLIOGRAPHY. The range of topics treated in this article is great, encompassing several kinds of dating methodology and several areas of applicability—*e.g.*, archaeology, the geological sciences, and astrogeology. Accordingly, a diverse array of books of relevance must be cited here. The titles are largely self explanatory.

WILLIAM B.N. BERRY, *Growth of a Prehistoric Time Scale Based on Organic Evolution* (1968); DON R. BROTHWELL and ERIC HIGGS (eds.), *Science in Archeology*, rev. ed. (1969); G. BRENT DALRYMPLE and MARVIN A. LANPHERE, *Potassium-Argon Dating* (1965); DON L. EICHER, *Geologic Time* (1968); HENRY FAUL (ed.), *Nuclear Geology* (1954); *Ages of Rocks, Planets, and Stars* (1966); E.I. HAMILTON, *Applied Geochronology* (1965); W.B. HARLAND (ed.), *The Phanerozoic Time-Scale* (1964); PATRICK M. HURLEY, *How Old Is the Earth?* (1959); RUTH E. MOORE, *Man, Time, and Fossils* (1953); F. FRITZ OSBORNE (ed.), *Geochronology in Canada* (1964); *Radioactive Dating: Symposium Proceedings*, International Atomic Energy Agency, Vienna (1963); *Radioactive Dating and Methods of Low-Level Counting: Symposium Proceedings*, International Atomic Energy Agency, Vienna (1967); ROBERT R. SHROCK, *Sequence in Layered Rocks* (1948); H.E. WRIGHT, JR. and DAVID G. FREY (eds.), *The Quaternary of the United States* (1965); F.E. ZEUNER, *Dating the Past: An Introduction to Geochronology*, 4th ed. rev. (1958).

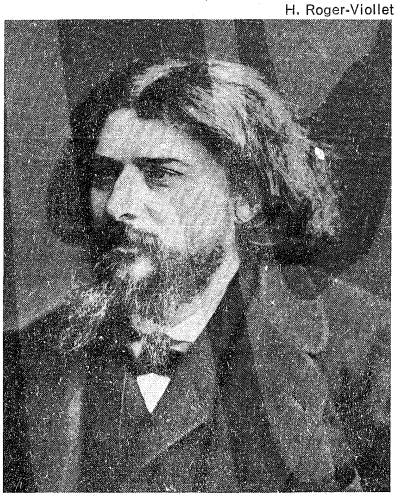
(E.A.O.)

Daudet, Alphonse

Remembered today chiefly as an outstanding short story writer and the author of sentimental tales of provincial life in the south of France, Alphonse Daudet was chiefly known in his own day as a dramatist and novelist. Although usually classed with the Naturalist school of writers in his documentation of the 19th-century contemporary scene, particularly among the lower classes of soci-

ety, Daudet is distinguished from other members of the school by his more sensitive approach and by his impressionistic description.

H. Roger-Viollet



Daudet.

He was born at Nîmes, France, on May 13, 1840, the son of a silk manufacturer. In 1849 his father was obliged to sell his factory and to move to Lyon where Alphonse was at first repelled by the damp-laden atmosphere but soon found compensations in escapades and in the life of the Rhône. He wrote his first poems and his first novel at the age of 14. In 1857 his parents lost all their money, and Daudet had to give up his hopes of matriculating. His work as an usher at a school at Alès for six unhappy months culminated in his dismissal but later furnished the theme, with embellishments and omissions, for his semi-autobiographical novel *Le Petit Chose* (1868). At the end of the year he went to join his elder brother, Ernest, in Paris.

Early works

Daudet now threw himself into writing and began to frequent literary circles, both Bohemian and fashionable. A handsome young man, he formed a liaison with a model, Marie Rieu, to whom he dedicated his only book of poems, *Les Amoureuses* (1858). His long and troubled relationship with her was to be reflected, much later, in his novel *Sapho* (1884). He also contributed articles to the newspapers, in particular to *Figaro*. In 1860 he met Frédéric Mistral, the leader of the 19th-century revival of Provençal language and literature, who awakened his enthusiasm for the life of the south of France, which was regarded as inherently passionate, artistic, and sensuous as opposed to the moral and intellectual rigour of the north. In the same year he obtained a secretarial post under the duc de Morny.

His health undermined by poverty and by the venereal disease that was eventually to cost him his life, Daudet spent the winter of 1861–62 in Algeria. One of the fruits of this visit was *Chapatin le tueur de lions* (1863), whose lion-hunter hero can be seen as the first sketch of the author's future Tartarin. Daudet's first play, *La Dernière Idole*, made a great impact when it was produced at the Théâtre de l'Odéon in Paris in 1862. His winter in Corsica at the end of 1862 is recalled in passages of his *Lettres de mon moulin* (1869; "Letters from My Mill"). His full social life over the years 1863–65 (until Morny's death) provided him with the material that he analyzed mercilessly in *Le Nabab* (1877). In January 1867 he married Julia Allard, herself a writer of talent, with whom he was deeply in love and who gave him great help in his subsequent work. They had two sons, Léon and Lucien, and a daughter, Edmée.

In the Franco-German War, which had a profound effect on his writing (as can be judged from his second volume of short stories *Les Contes du lundi*, 1873; "Monday Tales"), Daudet enlisted in the Army, but he fled from Paris during the terrors of the Commune of 1871. His *Les Aventures prodigieuses de Tartarin de Tarascon* (1872) was not well received, though its adventurous hero is now celebrated as a caricature of naiveté and boastfulness. His play *L'Arlésienne* was also a failure (although its 1885 revival was acclaimed). His next novel, *Fromont jeune et Risler aîné* (1874), which won an award from the Académie Française, was a success, and for a few years he enjoyed prosperity and fame—though not without some hostile criticism.

In his last years Daudet suffered from an agonizing ailment of the spinal cord caused by his venereal disease. *La Doulou* (not published until 1931) represents his attempt to alleviate his pain by investigating it. With admirable self-control he continued to write books of all sorts and to entertain Parisian literary and musical society. He was a kindly patron of younger writers—for instance, of Marcel Proust. In 1895 he visited London and Venice. He died suddenly on Dec. 16, 1897.

Evaluation

Psychologically, Daudet represents a synthesis of conflicting elements, and his actual experience of life at every social level and in the course of travels helped to develop his natural gifts. A true man of the south of France, he combined an understanding of passion with a view of the world illuminated by Mediterranean sunlight and allowed himself unfettered flights of the imagination without ever relaxing his attention to the detail of human behaviour. All his life he recorded his observations of other people in little notebooks, which he used as a reservoir of inspiration: a novel, he held, should be "the history of people who will never have any history." Yet there was nothing unfeeling in his approach (he has even been accused of sentimentality), and he was free from preconceived ideas: unlike his fellow naturalists, he believed that the world in its diversity was misrepresented by novelists who concentrated only on its uglier aspects.

At the same time his objective interest in external detail went hand in hand with the expression of an extraordinarily compassionate personality and a reverence for the mystery of things and of individuals. Everything in his world had an inner reality that he reproduced no less faithfully than he did the material phenomena. Finally, he saw passion as endowed with something like the force of destiny, and this conception, which bore fruit in many of his writings, tempers his satire with pity and brings him into kinship with Dickens as well as with Maupassant.

Daudet's work as a whole reveals not so much a continuous evolution as an episodic process in which various literary tendencies found expression successively. Even so, the antiromantic irony of *Tartarin de Tarascon* gave place to a realism akin to that of the Pointillist and Impressionist painters in *Lettres de mon moulin*, which was followed by the tragic tone of *L'Arlésienne* as a corrective to his earlier mockery of southern characteristics; also there is more sympathy and anxiety than irony in *Le Petit Chose* and *Les Contes du lundi*. As he grew older Daudet became more and more preoccupied with the great conflicts in human relationship: *Jack* (1876) presents a woman torn between physical and maternal love; *Numa Roumestan* (1881), the antagonism between the northern and the southern character in man and woman; *L'Évangéliste* (1883), filial affection struggling against religious fanaticism; and *La Petite Paroisse* (1895), the contrarieties of jealousy. In *Sapho* (1884), underlying the moral issue, there is Daudet's evaluation of a whole generation of young men, together with a statement of the age-old dilemma of the lover who must choose between freedom and pity for the girl he leaves. *Le Trésor d'Arlatan* (1897), *Notes sur la vie*, and *Nouvelles notes* show Daudet as a bold psychologist, anticipating Freud in his analysis of complexes. Truth and fantasy, merciless delineation and poetry, clear-sighted seriousness and a sense of humour, irony, and compassion, all the contrasting elements of which man's dignity is made up are to be found harmonized in Daudet's best work.

MAJOR WORKS

VERSE: *Les Amoureuses* (1858).

PLAYS: *La Dernière Idole* (1862); *Les Absents* (1863); and *L'Oeillet blanc* (1865), three plays written with E. L'Épine; *Le Frère aîné* (1867; Eng. trans., 1930); *Le Sacrifice* (1869);

L'Arlésienne (1872); *Lise Tavernier* (1872; Eng. trans., 1890); *La Lutte pour la vie* (1890); *L'Obstacle* (1891), with L. Hennique; *La Menteuse* (1892).

NOVELS AND SHORT STORIES: *Le Roman du Chaperon-Rouge* (1862); *Le Petit Chose* (1868; *My Brother Jack*, 1877; *The Little Good-for-Nothing*, 1878; *The Little Weakling*, 1917); *Lettres de mon moulin* (1869; *Letters from My Mill*, 1880); *Lettres à un absent* (1871); *Aventures prodigieuses de Tartarin de Tarascon* (1872; *The New Don Quixote or the Wonderful Adventures of Tarascon*, 1875; the first of several English translations); *Contes du lundi* (1873; *Monday Tales*, 1927; another Eng. trans., 1950); *Les Femmes d'artistes* (1874; *Artists' Wives*, 1890); *Fromont jeune et Risler aîné* (1874; *Fromont the Younger and Risler the Elder*, 1880); *Robert Helmont* (1874; Eng. trans., 1888); *Jack* (1876; Eng. trans., 1877); *Le Nabab* (1877; *The Nabob*, 1878); *Les Rois en exil* (1879; *Kings in Exile*, 1880); *Numa Roumestan* (1881; Eng. trans., 1884); *L'Evangéliste* (1883; Eng. trans., 1883); *Sapho* (1884; Eng. trans., 1886; one of several translations); *Tartarin sur les Alpes* (1885; *Tartarin on the Alps*, 1887); *La Belle-Nivernaise* (1886; Eng. trans., 1887); *L'Immortel* (1888); *Port-Tarascon* (1890; Eng. trans. by Henry James, 1891); *Rose et Ninette* (1892; Eng. trans., 1892); *La Petite Paroisse* (1895); including *L'Enterrement d'une étoile*, *La Fédor* (1896); *Le Trésor d'Arlatan* (1897); *Soutien de famille* (1898; *The Hope of the Family*, 1898).

MEMOIRS: *Souvenirs d'un homme de lettres* (1888; *Recollections of a Literary Man*, 1889); *Trente ans de Paris* (1888; *Thirty Years of Paris and of My Literary Life*, 1888); *Notes sur la Vie* (1899); *Premier voyage, premier mensonge* (1900; *My First Voyage, My First Lie*, 1901); *La Doulou* (1931), notebooks.

LITERARY CRITICISM: *Pages inédites de critique dramatique, 1874–1880* (1923).

BIBLIOGRAPHY. J. BRIVOIS, *Essai de bibliographie des oeuvres de M. Alphonse Daudet* (1895, reprinted 1967). See also H. TALVART and J. PLACE, *Bibliographie des auteurs modernes de langue française*, vol. 4 (1933).

Editions: *Oeuvres complètes de Alphonse Daudet*, definitive edition, 20 vol. (1929–31).

Biography and criticism: For Daudet's own reminiscences, see *Trente ans de Paris* (1888; Eng. trans., *Thirty Years of Paris and of My Literary Life*, 1888) and *Souvenirs d'un homme de lettres* (1888; *Recollections of a Literary Man*, 1889). For general biography: LUCIEN DAUDET, *Vie d'Alphonse Daudet* (1941); R.H. SHERARD, *Alphonse Daudet: A Biographical and Critical Study* (1894); G.V. DOBIE, *Alphonse Daudet* (1949). For Daudet's early life: J.H. BORNECQUE, *Les Années d'apprentissage d'Alphonse Daudet* (1951), together with Bornecque's edition of *Le Petit Chose*, *Lettres de mon moulin*, and *Contes du lundi*, 2 vol. (all 1947), which provide a complete survey of Daudet's life and work down to 1870, with bibliography; see further, M. BRUYERE, *La Jeunesse d'Alphonse Daudet* (1955). On his personality and literary significance: LEON DAUDET, *Alphonse Daudet* (1898); E. ZOLA, *Les Romanciers naturalistes* (1881); J.H. ROSNY, *Torches et lumignons* (1921); M. SACHS, *The Career of Alphonse Daudet: A Critical Study* (1965). For special aspects: MARY BURNS, *La Langue d'Alphonse Daudet* (1916); G.R. SAYLOR, *Alphonse Daudet as a Dramatist* (1940).

(J.-H.B.)

Daumier, Honoré

Honoré Daumier, a French caricaturist and painter, is best known for his thousands of satirical lithographs that portray the society of his time as incisively as the novels of *La Comédie humaine* of his friend Honoré de Balzac. No less important, though hardly known during his lifetime, were his paintings, which helped introduce techniques of Impressionism into modern art.

Background and early life. Daumier was born on February 20 or 26, 1808, in Marseilles. Traits of his ancestry—a violent temperament, a generous and rather fanciful turn of mind, and an easily aroused capacity for pity—all form part of his character. His mother's family was from a village in which samples of unique ancient sculptured reliefs—fierce primitive human heads—had been found. His grandfather and father both worked in Marseilles as "glaziers"—that is to say, dealers in frames (or passe-partout pictures) and decorative tableaux that they painted themselves. His godfather was a painter. When Daumier was seven, his father abandoned his business in order to go to Paris and, like so many Provençals, seek his fortune as a poet. He was presented



"Self-Portrait," painting by Honoré Daumier. In the Musée Calvet, Avignon, France.
Giraudon

to the king, Louis XVIII; but his swift fall from favour—he was famous only for a fortnight—unbalanced him mentally. After apparently being confined for many years, he died in the Charenton asylum.

Daumier received a typical lower middle class education, but he wanted to draw, and his studies did not interest him. His family therefore placed him with an old and fairly well-known artist, Alexandre Lenoir. Lenoir, a student and friend of Jacques-Louis David, a leading classicist painter, was more an aesthetician than a painter. He had a pronounced taste for Rubens, one of whose works he kept in his collection. A connoisseur of sculpture, he had saved the most beautiful medieval and contemporary sculptures from the Revolutionaries, which inspired a lasting interest in Daumier.

Daumier was then not at all the uncultured, self-taught genius that most art historians of the late 19th and early 20th centuries have depicted. He did not rise from an artistic void—he was the child of artists, however modest and unsuccessful they had been in making a name. Added to the advantage of this ancestry, he also benefited from a more interesting artistic education than his contemporaries.

Sources for his caricatures

At the age of 13 his father's breakdown forced Daumier to seek paying work. He first became a messenger boy for a bailiff and, from this experience, acquired his familiarity with the world of the lawcourts. He worked next as a bookseller's clerk at the Palais-Royal. The Palais-Royal, with its arcades surrounding the garden, was one of the busiest spots in Paris, and there Daumier saw, parading before his employer's window, all the characters of the *Comédie humaine*, about whom he would later talk with his friend Balzac: not only men and women of fashion, intellectuals, and artists but also "captains of industry," or swindlers, as they were commonly called—all of them picturesque personalities that lent themselves to caricature.

Daumier's development was thus complete at that moment when, about 1825–28, he decided to give up everything to embark on the artistic career of which he had dreamed so long. He was a young man of about 18 or 20, from a family of painters, who had had an opportunity to admire Rubens, had learned to analyze sculpture, and had been able to observe the appearance and behaviour of different classes of society.

Physically he was ugly, at least according to the taste of his time. Heavyset like Balzac, although probably smaller, he had small but lively eyes and a large nose. He always kept a pipe in his mouth, in order to mask his Provençal accent and the frequent lisp of his native region.

Daumier could not, of course, live from painting or from sculpture as he had set out to do. He therefore accepted commissions for lithographs: portraits and, at a

very early age, cartoons of morals and manners (*caricatures de moeurs*), the first of these dating from 1822, when he was scarcely 15 years old and was just beginning to produce lithographs. Although some of his first works were signed, many others were not: they were portraits of celebrities that were signed by another native of Marseilles, Zéphirin Belliard, Daumier's elder by ten years and the author of a lavish *Iconographie des contemporains*. For the most part these portraits were mediocre, modelled on another artist's style, but they constituted an excellent apprenticeship for someone interested in the human physiognomy.

His life, devoted entirely to his work, was to be divided into two parts: from 1830 to 1847, he was a lithographer, cartoonist, and sculptor; and, beginning in 1848 and lasting until 1871, he was an Impressionist painter whose art was reflected in the lithographs he continued to produce. Constant work was not a burden to him; while producing 4,000 lithographs and 4,000 illustrative drawings, he sang sentimental songs whose foolishness made him laugh, and, "unconcerned with his works, he was always out drinking cheap wine with barge captains."

Satirical lithographs. In 1830 Daumier began his satirical work: his busts lampooning certain contemporary types and his many lithographs. He enjoyed the company of grandiloquent men and mainly associated with men of the left. It was at this time that Charles Philipon, a liberal journalist who had founded the opposition journal *La Caricature*, invited him to become a contributor.

King Louis-Philippe generally tolerated jokes at his expense, but, when unduly provoked, rather than bring suit against a paper, he preferred to seize it, a procedure that meant ruin for its staff and financial backers. Only once during his reign did he deal severely with an offender—with Daumier in 1832, and then only after the second of the artist's most violent attacks. Sentenced to six months in prison, Daumier spent two of them in the state prison and four in a mental hospital, the King apparently wanting to show that one had to be mad to oppose and caricature him.

Imprisonment by King Louis-Philippe

After his release in February 1833, Daumier was never again indicted, even though in his cartoons he continued to attack a regime, a form of society, and a concept of life that he scorned, while at the same time creating unforgettable characters. Daumier's types were universal: businessmen, lawyers, doctors, professors, and petits bourgeois. His treatment of his lithographs was sculptural, leading Balzac to say about him that he had a bit of Michelangelo under his skin.

Daumier's sculptures have still not been sufficiently studied. The 15 or so small busts that he modelled in clay for the window of the satirical journal for which he worked and that remained there some 30 years occupy an important place in the history of sculpture. Scarcely differing from official busts, but with the accentuation of a detail that made them caricatures, they constitute an unforgettable gallery of the politicians of the July monarchy. The complete series has not been preserved: it included a Louis-Philippe, which Daumier hid, and other pieces that were broken in moving. A few copies of the busts were cast in bronze in the 20th century, and their originality is the more striking when they are compared with similar pieces by other sculptors of that period.

Daumier's sculptures

Daumier's only close friends were sculptors, all of them romantic, poor, and ardent left-wingers. Although intimate with these few friends, Daumier did not form part of any of the many artistic or literary coteries of the time. He did not consort with the painters Eugène Delacroix and Gustave Courbet and the writer Victor Hugo and was not inclined to frequent salons, or even saloons. When he went to a café, it was with his neighbours on the Île Saint-Louis, where he lived between 1833 and approximately 1850 in a studio on the quay d'Anjou. This old studio still exists, at the top of a house overlooking a world of roofs and open windows, behind which his models lived. Nor did his wife, to whom he was very attached, his dear "Didine" (Léopoldine), mix in artistic circles; she was a dressmaker.

In 1848 Daumier believed the era of social justice for which he had militantly fought for 20 years had arrived, and he took part in the official competition for the representation of the republic that was to replace the portrait of the King in all the municipal buildings of France. His rough sketch was beautiful, and, had he agreed to complete the painting, he would have received the prize.

Impressionist techniques. He did not do so, however, for he had become preoccupied with new technical studies; earlier than others, he had discovered Impressionism—faces and bodies devoured by the surrounding light and becoming one with the atmosphere. He painted a great deal, and the more so as his studies in the new technique did not interest the satirical journals to which he now submitted drawings devoid of humorous meaning. He was supported by Charles Baudelaire and by that poet's friends. The two men had met in 1845 and saw each other more frequently after 1848. Baudelaire, who "adored him," wrote in 1857 the only significant article on Daumier to appear in the painter's lifetime.

Daumier was indeed the first of the Impressionists. As early as 1848, his lithographs show contours effaced by light. Two factors, however, prevented this fact from being noticed: the lack of interest in his lithographs shown by historians of painting and the lack of research to establish the exact dates of the lithographs. These dates were not necessarily those of their appearance. It has been shown, for instance, that the Impressionist lithographs, which appeared around 1860, had been rejected by journals in 1848 as being too bold, too modern. Because of this lack of demand, Daumier's Impressionist lithographs are not very numerous, but his painting shows that he was won over to Impressionism. The paintings, also, were too few in number to assign him his true place in the history of Impressionism—before Manet and Monet, both of whom admired him greatly.

Impressionist lithographs

The dating of Daumier's paintings has given rise to controversy. Nineteenth-century art historians and those prior to 1938 dated them largely in relation to the paintings of his contemporaries. But it now seems more reasonable to date them from his lithographs, for it is not likely for an artist who changes styles frequently to work differently when drawing on a lithographic stone than when painting on canvas. When Daumier imagined a form, he would fashion and refashion it numerous times in his lithography. It is not likely that he would wait four to six years after he had stopped dealing with it in his lithographs to treat it in his painting. The question may be studied again, but not on the basis of documents of the period, for Daumier's notebooks were too abbreviated, and critiques by his contemporaries were rare, since he seldom showed his works, and they went largely unnoticed.

Daumier's paintings are highly original, both in their style and in the subjects they present. He created the painting of morals and manners (*la peinture de moeurs*); he shows the everyday life on the Île Saint-Louis and its quays, such as children playing in the water, one of them brought back to its parents after an accident; horses leaving a water trough; washerwomen wearily returning up the stairs from the river or fighting against the wind, their bundles in their arms; drinkers in a pub; masons on a scaffold. He was stirred by the theatre, then by railroads, which he used as a means of showing galleries of faces as powerful in their impact as those of the "Ventre législatif" ("Legislative Belly" or "Vile Body of the Legislature"). Earlier, the Palais de Justice had provided him with the opportunity of drawing his dramatically impressive lawyers. Then he went to Valmondois, on the outskirts of Paris, and depicted rustic scenes. Much of his painting was devoted to artists' studios—not studios of a studied picturesqueness but those where artists were concerned with creating works of art.

Paintings of morals and manners

These subjects are found again in his lithographs, together with topical subjects, such as seaside resorts, hunting, and winter scenes, all of which, having been commissioned, seemed to inspire him less. Thus he transposed into his painting what had until then belonged to the domain of the caricatures of morals and manners.

Daumier was not often inspired by religious subjects, except for the image of Christ, insulted and ridiculed, in which there appears to be personal allusion—that of a man who each day hoped to draw his last cartoon so that he could devote himself to painting. On the other hand, mythology excited him during his Rubens period under Lenoir; but in this case, obviously, the subject was only a prop for the painting.

On several occasions Daumier painted historical subjects. He painted "Camille Desmoulins," the Revolutionary leader, rousing the crowd in 1789; and his "Emigrants" of 1857 is an allusion to the authoritarian empire of Napoleon III, a painting that echoes the words of the proscribed Victor Hugo: "It is not I who am proscribed, it is liberty; it is not I who am exiled, it is France."

The types that Daumier created did not always survive him. He created a Louis-Philippe, but above all Robert Macaire (the typical businessman of Louis-Philippe's reign). His "Bons Bourgeois" probably served as a reference for the French middle class up to the '40s. In any case, his "Lawyers" remain up-to-date.

Following tradition, Daumier took pupils who learned their craft by copying and imitating his works. Two of them are known by name: Boulard and Gill. Their works are known—incorrectly—as *faux Daumier*, for works are only false when one is unable to see that they are the exercises of pupils.

Less solitary than he is said to have been, and admired by the new school, notably by Manet, Daumier grew old: sadly, for he would have liked to give up his lithographic work in favour of painting, but he could not do so. But he was happy in the knowledge that his lithographs preserved their force, as Hugo said in 1870 when Daumier symbolized his great satirical poem *Les Châtiments* by a crucified eagle.

In 1871, Daumier, who had discreetly refused to be decorated by the empire, became a member of the leftist Paris Commune. After 1871, Daumier, now almost blind, lived on until Feburary 11, 1879. His drawings, like those of Edgar Degas, gained a magnificent wholeness. He had one last joy, an exhibition at the Durand-Ruel Gallery in 1878. People began to say that his paintings were at least as good as his lithographs, though it was more the republican than the artist that they wished to celebrate.

As a cartoonist, Daumier enjoyed a wide reputation, although as a painter he remained unknown. His fame was not based, any more than it is today, on critical appreciations but, rather, on the smiling or laughing admiration of those who read the satirical journals.

MAJOR WORKS

PAINTINGS: "Three Lawyers in Conversation" (*c.* 1843–46; Phillips Collection, Washington, D.C.); "Les Noctambules" (*c.* 1843–48; National Museum of Wales, Cardiff); "The Republic" (*c.* 1848; Louvre, Paris); "The Miller, His Son, and the Ass" (*c.* 1849; Museum and Art Gallery, Glasgow); "Nymphs Pursued by Satyrs" (*c.* 1849–50; Montreal Museum of Fine Arts); "The Drinkers" (*c.* 1856; Metropolitan Museum of Art, New York); "The Melodrama" (*c.* 1856–60; Neue Pinakothek, Munich); "The Print Collector" (*c.* 1857–60; Musée du Petit Palais, Paris); "Crispin and Scapin" (*c.* 1858–60; Louvre); "The Horsemen" (*c.* 1860–62; Museum of Fine Arts, Boston); "The Washerwoman" (*c.* 1860–62; Louvre); "The Third-Class Carriage" (*c.* 1862; Metropolitan Museum of Art, New York); "The Painter at His Easel" (*c.* 1866; Phillips Collection); "Don Quixote" (*c.* 1868; Neue Pinakothek).

LITHOGRAPHS: "The Legislative Belly" (1834); "Rue Transnonain, April 15, 1834" (1834); "Freedom of the Press" (1834); "Pygmalion" (1842); "Pardon Me, Sir, If I Bother You a Bit" (1844); "The Orchestra During the Performance of a Tragedy" (1852); "The Muse of the Beer Parlour" (1864).

DRAWINGS: "Two Men Looking Toward the Left" (*c.* 1840; Louvre); "Connoisseurs" (*c.* 1858; Cleveland Museum of Art); "Soup" (*c.* 1860–62; Louvre); "Clown" (*c.* 1868; Metropolitan Museum of Art, New York).

SCULPTURE: "The Refugees" (1848–49; National Gallery of Art, Washington, D.C.); "Ratapoil" (1850; National Gallery of Art, Washington, D.C.).

BIBLIOGRAPHY. There is no definitive work on Daumier. Nineteenth-century studies include DURANTY, "Daumier," in *Gazette des beaux-arts*, 17:429–433 (1878); EUGENE MONTROSIER, "Honoré Daumier," in *L'Art*, 2:25–32 (1878); GUSTAVE GEFFROY, "Daumier," in *Revue de l'Art*, 9:229–250 (1901); and the only nearly contemporary book, written thanks to some testimonies, including those of his widow: ARSENE ALEXANDRE, *Honoré Daumier: l'homme et l'oeuvre* (1888; reissued as *Daumier*, 1928). Twentieth-century works, particularly recommended, are PAUL VALERY, *Daumier* (1938); HENRI FOCILLON, "Visionnaires—Balzac et Daumier," in *Essays in Honor of Albert Feuillerat* (1943); JEAN ADHEMAR, *Honoré Daumier* (1954; Eng. trans., *Daumier: Drawings and Watercolors*, 1954), an exhaustive study; KARL E. MAISON, *Honoré Daumier: Catalogue raisonné of the Paintings, Watercolours, and Drawings*, 2 vol. (1968); HOWARD P. VINCENT, *Daumier and His World* (1968); and BERNARD LEMANN, "Daumier and the Republic," *Gazette des beaux-arts*, 27:105–120 (1945). On the sculpture of Daumier, see MAURICE GOBIN, *Daumier sculpteur: 1808–1879* (1952); and JEANNE L. WASSERMAN, assisted by JOAN M. LUKACH and ARTHUR BEALE, *Daumier Sculpture: A Critical and Comparative Study* (1969), a very good exhibition catalog and study from the Fogg Art Museum.

(J.Ad.)

David

David, second of the Israelite kings, established a united kingdom over all Israel with Jerusalem as its capital. Through his political and religious leadership, he founded a national identity and unity that survived later political divisions and disasters. In Jewish tradition he became the ideal king, the founder of an enduring dynasty, around whose figure and reign clustered messianic expectations of the restoration of the city and the people of Israel. Since he was a symbol of fulfillment in the future, the New Testament writers emphasized that Jesus was of the lineage of David. He is also held in high esteem in the Islāmic tradition. For ancient Israel, it has been said, his role was second only to that of Moses.

Political career. The youngest son of Jesse (grandson of Boaz and Ruth), David was born in Bethlehem. He began his career as an aide at the court of Saul, Israel's first king, and became a close friend of Saul's son and heir, Jonathan, and the husband of Saul's daughter Michal. He so distinguished himself as a warrior against the Philistines that his resultant popularity aroused Saul's jealousy and a plot was made to kill him. He fled into southern Judah and Philistia, on the coastal plain of Palestine, where, with great sagacity and foresight, he began to lay the foundations of his career.

Rise to power

Beginning as an outlaw, with a price on his head, he led the life of a Robin Hood on the desert frontier of his country (Judah). He became the leader and organizer of other outlaws and refugees; and, according to the Bible, ". . . everyone who was in distress, and everyone who was in debt, and everyone who was discontented, gathered to him; and he became captain over them." This group progressively ingratiated itself with the local population by protecting them from other bandits or, in case they had been raided, by pursuing the raiders and restoring the possessions that had been taken. Though sometimes dependent upon the Philistine kings of Gath for protection from the pursuit of King Saul, David managed to retain his status as a patriot in the eyes of his own people in Judah and, even as one who had, indeed, been an innocent and loyal servant of the demented Saul. He also won the favour of many Judaean elders by various politic gestures. Thus, by biding his time, he eventually had himself "invited" to become king, first by Judah in Hebron and later by all Israel, not as a rebel against Saul but as his true successor.

This opportunity emerged when Saul and Jonathan were slain in battle against the Philistines on Mount Gilboa. David entered Hebron, where he was proclaimed king. He had to struggle for a few years against the contending claim and forces of Ishbaal, Saul's surviving son, who had also been crowned king, but the civil war ended with the murder of Ishbaal by his own courtiers and the anointing of David as king over all Israel (including tribes beyond Judah). He proceeded to conquer the walled city of Jerusalem, held by the alien Jebusites, which he made the capital of the new united kingdom,

and to which he moved the sacred ark of the Covenant, the supreme symbol of Israelite religion. He defeated the Philistines so thoroughly that they were never again a serious threat to the Israelites' security, and he annexed the coastal region. He went on to establish an empire by becoming the overlord of many small kingdoms bordering on Israel, including Edom, Moab, and Ammon. Beginning about 1000 BCE (before the Common Era—BC), David's reign lasted for about 40 years (until 962 BCE).

Family problems and personal character

David's great success as a warrior and empire builder was marred by family dissensions and political revolts, which were interrelated. To tie together the various groups that constituted his kingdom, David took wives from them and created a harem. The resultant family was an extreme departure from the family in the consanguinal context, the traditional clan structure. David's wives were mostly completely alien to one another, and his children were without the directing support of established social patterns that provided precedents for the resolution of conflict or for establishing the rights of succession. Thus, David's third son, Absalom, murdered the eldest son, Amnon, for the latter's rape of Tamar, the former's sister and the latter's half sister. After a period of exile and then of reconciliation with King David, Absalom used the favour he had gained among the people and some courtiers to launch a rebellion that sent his father fleeing across the Jordan and that made him master of Jerusalem and the royal harem for a time. Eventually, Absalom's forces were defeated, and he was killed by Joab, David's general, and it was Solomon, born of David's union with Bathsheba, who became the King's eventual heir.

The authors of the biblical accounts (in books I and II Samuel) of David's political career display a deep insight into his character. David was a man who could make an indelible personal impression in a specific situation. His doubling of the bride price set by Saul for Michal illustrates this capacity for imaginative action and dramatic publicity. Coupled with this ability to exploit the immediate situation in the service of his momentary requirements, he possessed the knack of making his conduct in particular situations serve his persistent and long-range aims. For example, the two versions of his refusal to assassinate King Saul when he had it in his power to do so (in I Samuel 24 and 26) do not simply present an inspiring example of gallantry in a moment of dramatic confrontation; they also contribute to the enduring reputation of David as a man who, even in his years as an outlaw, had a deep respect for established institutions, especially for the sacred office of the king ("the Lord's anointed"). Later, after the death of Saul and Jonathan, David again confirmed this point at a moment in which it was crucially important for him to do so for the sake of his own career. A young Amalekite who came to report Saul's death to David intimated that he had had a share in it. He thought that as the bearer of good news he would be rewarded, but his miscalculation cost him his life. David sensed that in an hour of national disaster the differences between him and Saul were of no importance. He had the Amalekite slain for having laid hands on the Lord's anointed, and with his men he performed the mourning rites for Saul and Jonathan, memorializing them in a deeply moving elegy. Somewhat later, after David had become king in Hebron, he learned that the men of Jabesh-Gilead, a town across the Jordan that had been fanatically attached to King Saul, had recovered the bodies of Saul and Jonathan to give them honourable burial. David sent the town a message commending it for its act of reverent loyalty, which had been undertaken at great risk. His action in this episode, also, was both political and sincere; and it was eminently suited to the situation in which the conciliation of all Israelites was of the greatest importance for both the career of David and the survival of the nation.

In the case of Absalom's rebellion, a poignant conflict took place between parental love and political power. When the news of his son's slaying came to him he broke down into deep grief and lamented, "O my son Absalom, my son, my son Absalom! Would I had died instead of you . . .!" But he was rebuked by his general Joab (who had ignored the King's direct order and had the young rebel killed) as showing more concern for his enemies than his supporters and risking the loss of public esteem and so of his rule. Thereupon he returned with his old energy and wile to the task of uniting and reconciling the various factions in Israel, including putting down another revolt, this time by Sheba, the son of Bichri, of the tribe of Benjamin.

Nation builder: David's political achievement. David was Israel's first successful king. He united all of the Israelite tribes, became the effective ruler over all, and was the founder of an enduring dynasty. Thus, he succeeded where King Saul had failed and attained a unique place in Israel's history and tradition. II Samuel 9–20 and I Kings 11–22 provide the primary source for knowledge of his reign and of the succession. It is generally agreed that this "history" was written very soon after the reign of David; as such, it is perhaps the oldest piece of historiography in the Western world. Known as the "Family History of David," and also as "The Succession History," it is an especially clear mirror in which to study the problems David faced in displacing charismatic political leadership and authority with hereditary monarchy.

From tribalism to monarchism

For centuries before David's rise to kingship, Israelites had been held together in loose tribal confederacies. The northern confederacy, with its centre at Shechem, is the best known. It was dominated by the tribe of Ephraim. A tribe was a collection of clans; and a clan was simply an expanded family. The consanguinal and familial character of Israelite society is a basic feature of Semitic tradition and is today still intact in the Arab society of the peninsula of Arabia. There the founding of the Sa'ūdī dynasty in the present century offers close and instructive parallels to David's problems and accomplishments 3,000 years ago. For example, both the revolt of Absalom, which necessitated David's exile for a time, and the grasp of the eldest surviving son, Adonijah, for the succession, in competition with the sponsors of Solomon, very nearly succeeded because they appealed to traditions of local and tribal authority, winning the support of many who were disillusioned with the swift centralization of power that had accompanied the establishment of the Davidic empire. In II Samuel 15, for example, Absalom, in his bid for support, says that he would like to exercise judgment in the premonarchic manner, as an elder in the gate. Ironically, he was attempting to displace his father by the same means by which David had so successfully risen to power; *i.e.*, appealing to local clans. Later, after Solomon's reign had ended, the united kingdom broke up when these tribalistic traditions again reasserted themselves. The relentless movement of social evolution made impossible the re-establishment of a tribal society; but the vitality of the tribal heritage was still very strong, both in David's day and later. Thus, there was a basic instability in his position; he faced the problem of winning consent for and establishing the legitimacy of his office, for it was an imported novelty in the social structures and traditions of Israel, on the model of the ancient Near Eastern kingships.

David's position in the tribal units that made up Judah was secure, for he had united them and had risen to authority over Judah through his adroit use of the indigenous social and political instruments of its clan structures. Therefore, Judah accepted his legitimacy and never disowned his dynasty. He sought to win the consent of all Israel, first, by the decisively successful war against the Philistines, which made the whole land secure; and then by establishing the city of Jerusalem as the centre both of Israel's political power and of its worship. On the political level this effort was not enough, for the kingdom was divided after the death of Solomon; but on the religious and cultic level it did eventually succeed, for Jerusalem, the "city of David," became the Holy City for all Jews, and the messiah, "the anointed one" of the house of David, a sign of the relationship between the God of Israel and his people.

Messianic symbol: David's religious role and significance. In Israel's religious tradition the royal line, or "house," of David became a primary symbol of the bond between God and the nation; the king was the mediator between the deity and his people. As in many ancient traditions, the king was thought of as both divine and human. The English word messiah is derived from *hameshiach* ("the anointed one"), the title of the kings of the line of David. Thus, in later times of disaster, Israel began to wait for a messiah, a new mediator of the power of God that would redeem the people and its land. By designating Jesus as the son of David, Christianity dramatized its conviction that this hope had been fulfilled. David lived in the memory of his people in a double way: as the great founder of their political power and as the symbol of a central facet of their religious faith.

David as king-priest and messiah in Jerusalem the Holy City

The process by which David achieved this status for himself, his house, and his city may be traced in II Samuel 5–8. When David took Jerusalem, he assumed the rule over its inhabitants and their religious institutions with the cult centred on Mt. Zion. The previous (Jebusite) ruler had been both king and high priest, and played the role of mediator between the city and its deity. There was no precedent for such a mediative and priestly role of kings in Israelite religion, nor of walled cities as the seat of government and worship. Apparently, David simply took over the Jebusite cult on Zion and adapted it to his own (and Israelite) use. Beginning with David and throughout the entire period of the monarchy, for about four centuries, Israel's worship on Zion gave a central place to the king, not simply as officiant but substantively, as the figure who in his office and person embodied the relationship between God and the nation. In contrast, the premonarchic worship of Israel, at Shechem and elsewhere, had featured a covenant between God and the people, through their tribal heads, as the bond in the relationship. By taking over and adapting Jerusalem's ancient cult, David provided Israel with a new worship, one that featured his own status and its sacral significance.

Israel's God was named Yahweh. David made this name the supreme name for deity in Jerusalem (previously perhaps "Salem"), to indicate his conquest of the city. All former names and titles of deity became attributes or titles of Yahweh, the God of Israel, the conqueror; for example, El 'Elyon (God Most High). While the Israelite name for God displaced all others, the substance of the worship remained similar: Yahweh had created the world and ruled the nations; he had established kingship as the sign and means of his universal rule; and Zion was the seat of his chosen king, David, his anointed. Yahweh himself was enthroned on Zion, and his king sat at his right hand as his regent. David thus continued the line of king-priests that had reigned in Jerusalem from the founding of the city, and, according to a legend that may have developed in this context, the patriarch Abraham had been blessed by Melchizedek, an earlier representative of the line, when he had presented tithes to him.

Having adopted the ancient cult of Jerusalem as a means of giving sacral significance to his royal status and having renamed it the cult of Yahweh, by whose power he had conquered, David also made an important move to make the new shrine and its worship relate to the premonarchic experience of Israel. He brought the ark to Jerusalem and established it as the central object of the cult. According to tradition, it had travelled with Israel in the wilderness and led the way into the land. It was a rectangular wooden box, originally without a cover, that established and located the presence of Yahweh with the people of Israel. So close was the connection that the ark could be addressed as Yahweh. The ark was carried into battle to demonstrate that Yahweh fought for Israel; and it was carried in the wilderness, to show that he travelled with his people. In worship, it was apparently carried in procession in the pilgrimages that were features of the annual feasts. It was a sign and even the embodiment of Yahweh's presence. David could have chosen no better way of making premonarchic Israelites accept the royal cult on Zion than by incorporating the ark, with all its ancient associations, into the new ceremonial.

Historical consequences

David's adaptation of the Zion cult, with its understanding of kingship as the substance and means of the presence of God on earth, was to have momentous consequences for the religious history of mankind, notably for the experience of the entire Western world. Because of it Jerusalem became the Holy City, and David became the prototype of an awaited messiah. As symbol of the Messiah, the return of David, or the coming of David's "son" stood for the reassertion of the divine rule and presence in history: to judge it, to redeem it, to renew it. David thus became the symbol of a fulfilment in the future, final peace.

In the apocalyptic developments in Judaism that mark the last two pre-Christian centuries, the symbolic role of David stressed his status as divine mediator. The son of David became more emphatically a heavenly figure, the son of God enthroned to rule over the nations of the world. This was the matrix for the rise of Christianity. The new faith interpreted the career of Jesus by means of the titles and functions assigned to David in the mysticism of the Zion cult, in which he served as priest-king and in which he was the mediator between God and man.

BIBLIOGRAPHY. R.A. CARLSON, *David, the Chosen King* (Eng. trans. 1964), provides a most complete analysis of the development of the traditions about David now embedded in biblical materials. JUAN BOSCH, *David: Biografía de un rey* (1963; Eng. trans., *David: The Biography of a King*, 1966), is a realistic account of the actual dynamics of the political career of the King. Two works on the role of David as a mystical figure in the religious mythology and imagination of ancient Israel are AAGE BENTZEN, *Messiah-Moses Redivivus-Menschensohn* (1948; Eng. trans., *King and Messiah*, 1955); and HELMER RINGGREN, *The Messiah in the Old Testament* (1956). The former stresses that he was the subject of messianic expectations that clustered around hopes for a political deliverance and national re-establishment. The latter connects the King with the "Suffering Servant" of Isaiah 53, which is used to illustrate the theme of vicarious suffering in the New Testament's interpretation of the career of Jesus.

(J.C.Ry.)

David, Jacques-Louis

In compartmented histories of European art, the French painter Jacques-Louis David appears as a principal exponent of the late-18th-century Neoclassical reaction against the pretty, feminine, allegedly decadent Rococo style and in favour of a return to the supposedly more manly, more virtuous styles of ancient Rome and Athens. But David cannot be easily assigned to a single historical category. He was the most talented delineator of the aspirations and the tragedy of the French Revolution, and he then became the chief celebrator in paint of the pomp of the Napoleonic empire. In his work and his doctrines can be found not only Neoclassicism but also the seeds of later Romanticism, Realism, and Academicism. His reputation, after a sharp decline under charges of dryness, frigidity, and theatricality, has been rising steadily during the second half of the 20th century.

Formative years. David was born in Paris on August 30, 1748, a year when new excavations at the ash-buried ruins of Pompeii and Herculaneum were beginning to encourage a stylistic return to antiquity (without being, as was long supposed, a principal cause of that return). His father, a small but prosperous dealer in textiles, was killed in a duel in 1757, and the boy was subsequently raised, reportedly not very tenderly, by two uncles. After classical literary studies and a course in drawing, he was placed in the studio of Joseph-Marie Vien, a history painter who catered to the growing Greco-Roman taste without quite abandoning the light sentiment and the eroticism that had been fashionable earlier in the century. At 18 the obviously gifted budding artist was enrolled in the school of the Académie Royale de Peinture et de Sculpture. After four failures in the official competitions and years of discouragement that included an attempt at suicide (by the stoic method of avoiding food), he finally obtained, in 1774, the Prix de Rome, a government scholarship that not only provided a stay in Italy



David, self-portrait, oil painting, 1794. In the Louvre. 81 cm X 64 cm.
Alinari

but practically guaranteed lucrative commissions in France. His prize-winning work, "Antiochus and Stratonice," reveals that at this point he could still be influenced slightly by the Rococo charm of the painter François Boucher, who had been a family friend.

Italian influences on David

In Italy there were many influences, including those of the dark-toned 17th-century Bolognese school, the serenely classical Nicolas Poussin, and the dramatically realistic Caravaggio. David absorbed all three, with an evident preference for the strong light and shade of the followers of Caravaggio. For a while he seemed determined to fulfill a prediction he had made on leaving France: "The art of antiquity will not seduce me, for it lacks liveliness. . . ." But he became interested in the Neoclassical doctrines that had been developed in Rome by, among others, the German painter Anton Raphael Mengs and the art historian Johann Joachim Winckelmann. In the company of Quatremère de Quincy, a young French sculptor who was a strong partisan of the return to antiquity, he visited the ruins of Herculaneum, the Doric temples at Paestum, and the Pompeian collections at Naples. In front of the ancient vases and columns, he felt, he said later, that he had just been "operated on for cataract of the eye."

Rise to fame: 1780–94. Back in Paris in 1780, he completed and successfully exhibited "Bélisaire demandant l'aumône" ("Belisarius Asking Alms"), in which he combined a nobly sentimental approach to antiquity with a pictorial technique reminiscent of Poussin. In 1782 he married the rather plain but spirited Marguerite Pécoul, whose father was a wealthy building contractor and the superintendent of construction at the Louvre—a position that carried considerable influence. From this date David prospered rapidly. The pathos and painterly skill of "Andromache Mourning Hector" brought him election to the Académie Royale in 1784; and that same year, accompanied this time by his wife and studio assistants, he returned to Rome with a commission to complete a painting that appears to have been originally inspired by a Paris performance of Pierre Corneille's *Horace*. The result, finally not based on any of the incidents in the play, was the "Oath of the Horatii." The subject is the solemn moment, charged with stoicism and simple courage, when the three Horatii brothers face their father and offer their lives to assure victory for Rome in the war with Alba; the pictorial treatment—firm contours, bare cubic space, sober colour, frieze-like composition, and clear lighting—is as austerely non-Rococo as the subject. Exhibited first in David's studio in Rome and then, following his return to France, in the official Paris Salon of 1785, the picture created a sensation; it was regarded as a manifesto for an artistic revival (the term Neoclassicism was not yet in use) that would cure Europe of the lingering addiction to dainty curves and boudoir themes. Eventually, it came to be regarded, although such was almost certainly not the first intention, as a manifesto for an end to the corruption of an effete aristocracy and for a return to the stern, patriotic morals attributed to republican Rome.

David became a culture hero; he was even referred to in some quarters as a messiah. He added to his fame by producing in 1787 the morally uplifting "Death of Socrates," in 1788 the less uplifting but archaeologically interesting "Amours de Paris et d'Hélene" ("Paris and Helen"), and in 1789 another lesson in self-sacrifice "Les Licteurs rapportent à Brutus les corps de ses fils" ("The Lictors Bringing to Brutus the Bodies of His Sons"). By the time the "Brutus" was on view, the French Revolution had begun, and this picture of the patriotic Roman consul who condemned his traitorous sons to death had an unanticipated political significance. It also had, through its presumably accurate reconstitution of the details of everyday Roman life, an effect that was perhaps equally unexpected, for with it David began the long and extensive influence he was to have on French fashions. Up-to-date homes began to display imitations of his Roman furniture; men cut their hair short in the Roman style; and women adopted the dresses and the coiffures of Brutus' daughters. Later on, even the flimsy Sabine dress, which left the breasts exposed, was adopted by the ultramodern.

Political activities

In the early years of the Revolution, David was a member of the extremist Jacobin group led by Robespierre, and he became an energetic example of the politically committed artist. He was elected to the National Convention in 1792, in time to vote for the execution of Louis XVI. By 1793, as a member of the art commission, he was virtually the art dictator of France and was nicknamed "the Robespierre of the brush." He preached moral and aesthetic sermons to the Convention:

> The artist must be a philosopher. Socrates the skilled sculptor, Jean-Jacques [Rousseau] the good musician, and the immortal Poussin, tracing on the canvas the sublime lessons of philosophy, are so many proofs that an artistic genius should have no other guide except the torch of reason.

Guided supposedly by the torch of reason and perhaps also by bitter memories of his many unsuccessful attempts to win the Prix de Rome, he succeeded in abolishing the Académie Royale and with it much of the old regime's system for training artists and providing them with patronage. The Académie was replaced briefly by a body called the Commune des Arts, then by a group called the Popular and Republican Society of the Arts, and then, finally, in 1795, after David was out of power, by the beginning of the system—a combination of the Institut de France and the École des Beaux-Arts—that dominated French artistic life during most of the 19th century.

As an artist during these years of his dictatorship, David was frequently busy with revolutionary propaganda. He had commemorative medals struck, set up obelisks in the provinces, and staged national festivals and the grandiose funerals the new government gave its martyrs. Some of his projects for paintings at this time were never completely carried out: one of these is the unfinished "Joseph Bara," which is a tribute to a drummer boy shot by the royalists, and another is the sketched "Oath of the Tennis Court" (Louvre, Musée National de Versailles et des Trianons, and the Fogg Art Museum, Cambridge, Massachusetts), which was to commemorate the moment in 1789 when the Third Estate (the commoners) swore not to disband until a new constitution had been adopted. The "Death of Lepeletier de Saint-Fargeau," painted to honour a murdered deputy and regarded by David as one of his best pictures, was eventually destroyed. The result of all this is that the artist's Jacobin inspiration is represented principally by "The Dead Marat," painted in 1793 shortly after the murder of the revolutionary leader by Charlotte Corday. This "pietà of the Revolution," as it has been called, is generally considered David's masterpiece and an example of how, under the pressure of genuine emotion, Neoclassicism could turn into tragic Realism.

Later years: 1794–1825. In 1794, after his friend Robespierre had been sent to the guillotine, David was arrested. At his trial he is said to have defended himself badly, mumbling that in the future he intended to attach himself "to principles and not to men." He was imprisoned twice, for four months in 1794 and for two more the next year, apparently most of the time in the not uncomfortable Palais du Luxembourg in Paris. He was consoled by being allowed to paint and also by the fact that his wife, who had divorced him two years earlier for having voted for the death of the King, now loyally returned in his hour of trouble and remarried him, on this occasion for good. During his first period in prison, he painted from his window his only landscape, the "Vue du jardin du Luxembourg à Paris" ("View of the Luxembourg Gardens"). While he was held temporarily in another Paris building, he did an unfinished "Self-Portrait." At 46 he appears as a boyish young man with romantically disheveled hair, brown eyes, and a generally aggressive, if worried, look; a cheek tumor from which he suffered all of his adult life and which is said to have impeded his speech gives his face a slight twist.

Arrest and imprisonment

Even during his imprisonment, he had retained three studios in the Louvre, and, after the amnesty of 1795, he devoted to teaching the same energy he had been devoting to revolutionary politics. Eventually, between the "Oath of the Horatii" and the Battle of Waterloo, he was responsible for the training and indoctrination of hundreds of young painters from all over Europe, among them such future masters as Baron François Gérard, Antoine-Jean Gros, and Jean-Auguste-Dominique Ingres. The indoctrination began with the premise that the basis of art was the contour, and so it can be held partly responsible for the excessive emphasis on drawing that characterized European academic painting in the 19th century. But David himself, as his works show, was not always hostile to rich chromatic effects; as late as 1860 he could be called, by no less a colorist than Eugène Delacroix, "the father of the whole modern school."

Mastery of portraiture

Neoclassicism was in theory inclined to scorn portraiture, since a contemporary sitter normally lacked both the universality and the nudity of an ancient statue. David, however, had done portraits, remarkable for their psychological individuality and their look of solid flesh, since the beginning of his career: in 1782–83 his sitter had been Alphonse Leroy, a Paris medical professor; in 1784 Mme Pécoul, his mother-in-law; in 1788 the chemist Antoine-Laurent Lavoisier, with Mme Lavoisier. In 1795 the freed artist portrayed his pretty, elegant sister-in-law, Mme Sériziat, and her dandyish husband. In the last year of the century, he produced his famous period piece, "Portrait de Mme Récamier," which he left unfinished because the sitter, then at the start of her career as a reigning Paris beauty, proved unreliable about hours for posing.

But David was not a man for the life of a mere teacher and portraitist. In 1799 he made a spectacular re-entry into public notice with a new giant canvas, "Les Sabines" ("The Intervention of the Sabine Women"). The picture, often mistakenly referred to as "The Rape of the Sabines," represents the moment, a few years after the legendary abduction, when the women, now contented wives and mothers, halt a battle between their Roman husbands and the Sabine men who have come on an unwanted rescue mission; in the middle of the melee stands the lovely Sabine woman Hersilia, appealing with one arm toward the Roman Romulus and the other toward the bearded Sabine Tatius. The artist had said that his aim was to move away from the allegedly crude Roman manner of the "Oath of the Horatii" into a more graceful Greek manner, and he did win enthusiastic applause for the elegance of his figures. He also won some approval for his supposed intention to preach conciliation after 10 years of bloodletting in France. But he attracted perhaps the most attention with the nakedness of his ancient warriors; having ceased to be the Robespierre of the brush, he now became, in a popular jingle, "the Raphael of the sans-culottes" ("without breeches"; the radical Republicans).

Napoleon admired "The Intervention of the Sabine Women" and saw possibilities for self-aggrandizement in the talent displayed. Soon David, without acquiring political office, was again a government painter, first under the Consulate and then, after 1804, under the Empire. He was not, however, the only prominent Frenchman to move from the Jacobin left to the Bonapartist right, and he had evidently always been a worshipper of historical heroes. His most important Napoleonic work is the huge "Coronation" of 1805–08, sometimes called "Napoleon Crowning the Empress Josephine"; in it Neoclassicism gives way to a style that combines the official portraiture of the old French monarchy with overtones—and occasional straight imitation—of the masters of the Italian Renaissance. This picture was followed in 1810 by the large "Napoleon Distributing the Eagles" and in 1812 by "Napoleon in His Study," a sharply perceptive portrait notwithstanding its conspicuously propagandistic intention.

After the fall of Napoleon in 1815, David was exiled to Brussels. Cut off from the excitement and stimulus of the great events he had lived through, he lost much of his old energy. Toward the end of his life, he executed, probably with considerable help from a Belgian pupil, François-Joseph Navez, one more remarkably convincing portrait: the "Les Trois Dames dites de Gand" ("Three Women of Ghent"). He died in Brussels on December 29, 1825.

MAJOR WORKS

"Combat de Minerve contre Mars" (1771; Louvre, Paris); "The Death of Seneca" (1773; Musée du Petit Palais, Paris); "Antiochus and Stratonice" (1774; École des Beaux-Arts, Paris); "Bélisaire demandant l'aumône" ("Belisarius Asking Alms," 1780; Musée Wicar, Lille); "Portrait of Pierre Desmaisons" (1782; Albright-Knox Art Gallery, Buffalo); "Andromache Mourning Hector" (1783; Louvre); "Oath of the Horatii" (1784; Louvre); "Portrait de Charles-Pierre Pécoul" (1784; Louvre); "Portrait of Mme Pécoul" (1784; Louvre); "The Death of the Ugolino" (1786; Musée des Beaux-Arts et d'Histoire Naturelle, Valence); "The Death of Socrates" (1787; Metropolitan Museum of Art, New York); "Les Amours de Paris et d'Hélene" ("Paris and Helen," 1788; Louvre); "Portrait of Lavoisier and His Wife" (1788; Rockefeller University, New York); "Les Licteurs rapportent à Brutus les corps de ses fils" ("The Lictors Bringing to Brutus the Bodies of His Sons," 1789; Louvre); "Self-Portrait" (1790; Louvre); "Le Serment du Jeu de Paume" ("The Oath of the Tennis Court," 1791; Louvre); "Portrait of Devienne" (1792; Musée Royaux des Beaux-Arts de Belgique, Brussels); "Portrait de Mme Trudaine" ("Portrait de Mme Chalgrin," 1792; Louvre); "The Dead Marat" (1793; Musée Royaux des Beaux-Arts de Belgique); "Vue du jardin du Luxembourg à Paris" ("View of the Luxembourg Gardens," 1794; Louvre); "Joseph Bara" (1794; Musée Calvet, Avignon); "Self-Portrait" (1794; Louvre); "Les Sabines" ("The Intervention of the Sabine Women," 1794–99; Louvre); "Portrait de Mme Sériziat" (1795; Louvre); "Bonaparte Crossing Mount St. Bernard" (1800; Musée National de Versailles et des Trianons); "Portrait de Mme Récamier" (1800; Louvre); "Sacre de l'empereur Napoleon Ier et couronnement de l'impératrice Joséphine dans la cathédrale Notre-Dame de Paris, le 2 décembre 1804" ("Coronation," "Napoleon Crowning the Empress Josephine," 1805–07; Louvre); "Portrait du Pape Pie VII" (1805; Louvre); "Sappho and Phaon" (1809; Hermitage, Leningrad); "Napoleon Distributing the Eagles" (1810; Musée National de Versailles et des Trianons); "Napoleon in His Study" (1812; National Gallery of Art, Washington, D.C.); "Léonidas aux Thermopyles" (1814; Louvre); "Les Trois Dames dites de Gand" ("Three Women of Ghent," *c.* 1815; Louvre); "Cupid and Psyche" (1817; private collection, Paris); "Mars and Venus" (1824; Musée Royaux des Beaux-Arts de Belgique).

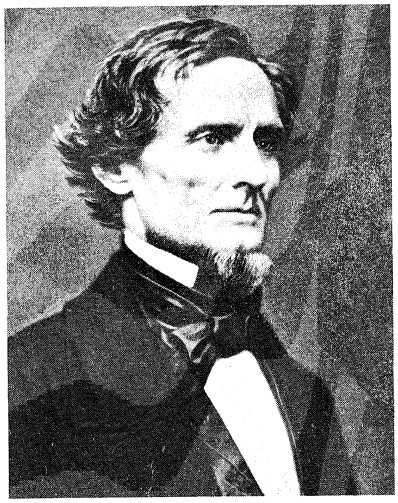
BIBLIOGRAPHY. E.J. DELECLUZE, *Louis David, son école et son temps* (1855), is a primary source of information by one of the artist's pupils. Standard modern biographies are D.L. DOWD, *Pageant-Master of the Republic: Jacques-Louis David and the French Revolution* (1948); and LOUIS HAUTECOEUR, *Louis David* (1954). General histories of style containing perceptive analyses of David's work are WALTER FRIEDLAENDER, *Von David bis Delacroix* (1930; Eng. trans., *David to Delacroix*, 1952); ROBERT ROSENBLUM, *Transformations in Late Eighteenth Century Art* (1967); and HUGH HONOUR, *Neo-Classicism* (1968).

(R.McMu.)

Davis, Jefferson

Jefferson Davis, whose truly unique place in world history is that of the first and only president of the Confederate States of America (1861–65), served in the United States Congress as representative and senator from Mississippi and was outstandingly successful as secretary of war under Pres. Franklin Pierce. After the death of John C. Calhoun in 1850, Davis was generally recognized as the foremost Southerner in Washington, and until 1861 he was admired by both the North and the South as a "virtuous and resolute man."

By courtesy of The Library of Congress, Washington, D.C.



Davis.

Jefferson Davis was born in western Kentucky, in what is now Fairview, on June 3, 1808, the tenth and last child of Samuel Emory Davis, a Georgia-born planter of Welsh ancestry who was also an American Revolutionary War veteran. When he was three his family settled on a plantation called Rosemont at Woodville, Mississippi. At seven he was sent for three years to a Dominican boys' school in Kentucky, and at 13 he entered Transylvania College. He later spent four years at the United States Military Academy, graduating in 1828.

Davis served as a lieutenant in the Wisconsin Territory and afterward in the Black Hawk War under the future president, then Col. Zachary Taylor, whose daughter Sarah Knox he married in 1835. According to a contemporary description, Davis in his mid-20s was "handsome, witty, sportful, and altogether captivating." He stood five feet eleven inches; his eyes were blue gray; his hair, gold.

In 1835 Davis resigned his commission and became a planter near Vicksburg, Mississippi, on land given him by his rich eldest brother, Joseph. Within three months his bride died of malarial fever. Grief-stricken, Davis stayed in virtual seclusion for seven years, creating a plantation out of a wilderness and reading prodigiously in constitutional law and world literature.

In 1845 Davis was elected to the United States Congress and, in the same year, married Varina Howell, a Natchez aristocrat who was 18 years his junior. He resigned his seat in Congress to serve in the war with Mexico and became a national hero for winning the Battle of Buena Vista with tactics that won plaudits even in the European press. After returning, severely wounded, he entered the Senate and shortly became chairman of the Military Affairs Committee. Pres. Franklin Pierce made him secretary of war in 1853. Davis enlarged the army, strengthened coastal defenses, and directed three surveys for railroads to the Pacific.

During the period of mounting intersectional strife, Davis spoke widely in both North and South, urging harmony between the sections. When South Carolina withdrew from the Union in December 1860, Davis still opposed secession, though he believed that the constitution gave a state the right to withdraw from the original compact of states. He was among those who believed that the newly elected Pres. Abraham Lincoln would coerce the South and that the result would be disastrous.

Twelve days after Mississippi seceded, Davis made a moving farewell speech in the Senate and pleaded eloquently for peace between brothers. He had hardly got back to his Brierfield plantation when the Confederate convention in Montgomery, Alabama, chose him provisional president. Inaugurated on February 18, 1861, his first act was to send a peace commission to Washington to prevent an armed conflict. Lincoln refused to see his emissaries and made secret preparations to send armed ships to Charleston, South Carolina, to "relieve" Ft. Sumter. When Lincoln called for 75,000 volunteers, four more states, including Virginia, promptly seceded.

President of the Confederacy

Davis faced a dire crisis. A president without precedent, he had to mold a brand-new nation in the midst of a war. With few resources except cotton and courage, with a white population only one-quarter that of the North, and with a small fraction of the North's industrial power, with inferior railroads, no powder mill, no navy, and no shipyard, the agricultural South was in no condition to stand an invasion. Davis dispatched agents to Europe to buy defensive arms and ammunition and sent representatives to try to secure recognition from England and France.

The Confederate capital was moved from Montgomery, Alabama, to Richmond, Virginia, in June 1861. On July 21 an invading Federal army was defeated at Manassas, Virginia, and retreated in disorder to the Potomac.

Davis had innumerable troubles, including a squabbling Congress, a dissident vice-president, two state governors who opposed the conscription law, and a few virulent newspapers, which, as a matter of principle, he refused to censor. Despite military defeats, unrelieved tensions, an appalling lack of manpower and armament, and skyrocketing inflation, Davis remained resolute, with Gen. Robert E. Lee in agreement by his side. In September 1864, a Northern writer wrote in *The Atlantic Monthly:*

> Our interview with him explained why, with no money and no commerce, with nearly every one of their important cities in our hands, and with an army greatly inferior in numbers and equipment to ours, the Rebels have held out so long. It is because of the sagacity, energy, and indomitable will of Jefferson Davis.

When General Lee surrendered to the North without Davis' approval, Davis and his Cabinet moved south, hoping to reach the trans-Mississippi area and continue the struggle until better terms could be secured from the North. At dawn on May 10, 1865, Davis was captured near Irwinville, Georgia. Imprisoned in a damp casemate on the moat at Ft. Monroe, Virginia, Davis was put in leg-irons. Two guards paced up and down his cell night and day, and a lamp was kept continually burning by his bedside. Though outraged Northern public opinion brought about his removal to healthier quarters, Davis remained a prisoner under guard for two more years. Finally, in May 1867, he was released on bail and went to Canada to regain his shattered health. Several notable Northern lawyers offered their free services to defend him in a treason trial, which Davis longed for. The government, however, never forced the issue, many believe because it feared that such a trial might establish that the original Constitution gave the states a right to secede. The case was finally dropped on December 25, 1868.

Capture and imprisonment

Davis made five trips to Europe in an effort to regain his health and to find employment as an agent for some British firm. In the United States, he was offered the presidency of three southern colleges, but the salaries were insufficient to support his wife and his four remaining children. For a few years he served as president of an insurance company in Memphis, Tennessee. In 1877, he retired to Beauvoir, a small Gulf-side estate near Biloxi, Mississippi, which a patriotic admirer provided for him. There he wrote his *Rise and Fall of the Confederate Government.* Though pressed to enter the United States Senate, he declined to "ask for amnesty," for he felt he had done nothing wrong, and he never regained his citizenship. He remained the acknowledged leader of the South.

Dedicated to the principles of democracy, Davis by nature was a benevolent aristocrat. Though diplomatic to a degree, he did not possess the pliancy of the professional politician. Austere in his dignity, he was really the warmest of men. Idolized by his family and relatives, he was said to be beloved even by onetime slaves. Through the fratricidal conflict he kept his old friends in the North. Many public men spoke and wrote unabashedly of their love for Jefferson Davis. Though he did not become a church member until he was past 50, when he joined the Episcopal Church, he was sustained throughout his life by a simple faith in the grace of God. Known from youth for his beautifully modulated voice, which profoundly influenced legislators and troops, in his old age he eloquently urged reconciliation with the North.

On December 6, 1889, at 81, Davis died in New Orleans of a complicated bronchial ailment. At his temporary interment he was accorded the greatest funeral the South has ever known. On May 31, 1893, he was buried permanently in Hollywood Cemetery in Richmond, Virginia.

BIBLIOGRAPHY. The first biography of Jefferson Davis was by one who knew and understood him, FRANK E. ALFRIEND, *The Life of Jefferson Davis* (1868). VARINA DAVIS, *Jefferson Davis, Ex-President of the Confederate States of America: A Memoir by His Wife*, 2 vol. (1890), is rich in family anecdotes as well as information about his public career. Two good short biographies are LANDON KNIGHT, *The Real Jefferson Davis* (1906); and MORRIS SCHAFF, *Jefferson Davis: His Life and Personality* (1922). A thoroughly sound appraisal is WILLIAM E. DODD, *Jefferson Davis* (1907). Of somewhat lesser value are H.G. ECKENRODE, *Jefferson Davis: President of the South* (1923); and ROBERT MCELROY, *Jefferson Davis: The Real and the Unreal*, 2 vol. (1937). In 1923, DUNBAR ROWLAND, after prodigious researches, published in ten volumes his invaluable compilation with notes: *Jefferson Davis, Constitutionalist: His Letters, Papers and Speeches.* R.W. PATRICK, *Jefferson Davis and His Cabinet* (1944), is sound throughout. HUDSON STRODE, *Jefferson Davis*, 4 vol. (1955–66), is the most up-to-date appraisal. The author had access to more than 1,000 hitherto unknown holograph letters owned by the Davis' grandson.

(Hu.S.)

Davy, Sir Humphry

A brilliant English chemist and exponent of the scientific method, inventor of the miner's safety lamp, and discoverer of the elements sodium and potassium, Humphry Davy was born in Penzance, Cornwall, on December 17, 1778, the elder son of middle class parents with an estate nearby. He was educated at the grammar school in Penzance, and, in 1793, at Truro. In 1795, a year after the death of his father, Robert, he was apprenticed to a surgeon and apothecary, and he hoped eventually to qualify in medicine. An exuberant, affectionate, and popular lad, of quick wit and lively imagination, he was fond of composing verses, sketching, making fireworks, fishing, shooting, and collecting minerals. He loved to wander, one pocket filled with fishing tackle and the other with rock specimens; he never lost his intense love of nature and, particularly, of mountain and water scenery.

While still a youth, ingenuous and somewhat impetuous, Davy had plans for a volume of poems, but he began the serious study of science in 1797, and these visions "fled before the voice of truth." He was befriended by Davies Giddy (later Gilbert; president of the Royal Society, 1827–30), who offered him the use of his library in Tradea and took him to a chemistry laboratory that was well equipped for that day. There he formed strongly independent views on topics of the moment, such as the nature of heat, light, and electricity and the chemical and physical doctrines of A.-L. Lavoisier. In his small private laboratory, he prepared and inhaled nitrous oxide (laughing gas), in order to test a claim that it was the "principle of contagion," that is, caused diseases. On Gilbert's recommendation, he was appointed (1798) chemical superintendent of the Pneumatic Institution, founded at Clifton to inquire into the possible therapeutic uses of various gases. Davy attacked the problem with characteristic enthusiasm, evincing an outstanding talent for experimental inquiry. He investigated the composition of the oxides and acids of nitrogen, as well as ammonia, and persuaded



Davy, oil painting after Sir Thomas Lawrence (1769–1830). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

his scientific and literary friends, including Samuel Taylor Coleridge, Robert Southey, and P.M. Roget, to report the effects of inhaling nitrous oxide. He nearly lost his own life inhaling water gas, a mixture of hydrogen and carbon monoxide sometimes used as fuel. The account of his work, published as *Researches, Chemical and Philosophical* (1800), immediately established his reputation, and he was invited to lecture at the newly founded Royal Institution of Great Britain in London, where he moved in 1801, with the promise of help from the British-American scientist Sir Benjamin Thompson (Count von Rumford), the British naturalist Sir Joseph Banks, and the English chemist and physicist Henry Cavendish in furthering his researches; *e.g.*, on voltaic cells, early forms of electric batteries. His carefully prepared and rehearsed lectures rapidly became important social functions and added greatly to the prestige of science and the institution. In 1802 he became professor of chemistry. His duties included a special study of tanning: he found catechu, the extract of a tropical plant, as effective as and cheaper than the usual oak extracts, and his published account was long used as a tanner's guide. In 1803, he was admitted a fellow of the Royal Society and an honorary member of the Dublin Society and delivered the first of an annual series of lectures before the board of agriculture. This led to his *Elements of Agricultural Chemistry* (1813), the only systematic work available for many years. For his researches on voltaic cells, tanning, and mineral analysis, he received the Copley Medal in 1805. He was elected secretary of the Royal Society in 1807.

Lectures at the Royal Institution

Davy early concluded that the production of electricity in simple electrolytic cells resulted from chemical action and that chemical combination occurred between substances of opposite charge. He therefore reasoned that electrolysis, the interactions of electric currents with chemical compounds, offered the most likely means of decomposing all substances to their elements. These views were explained in his Bakerian lecture of 1806, "On Some Chemical Agencies of Electricity," for which, despite the fact that England and France were at war, he received the Napoleon Prize from the Institut de France (1807). This work led directly to the isolation of sodium and potassium from their compounds (Bakerian lecture, 1807) and of the alkaline-earth metals from theirs (1808). He also discovered boron (by heating borax with potassium), hydrogen telluride, and hydrogen phosphide (phosphine). He showed the correct relation of chlorine to hydrochloric acid and the untenability of the earlier name (oxymuriatic acid) for chlorine; this negated Lavoisier's theory that all acids contained oxygen. He explained the bleaching action of chlorine (through its liberation of oxygen from water) and discovered two of its oxides (1811 and 1815), but his views on the nature of chlorine were disputed. He was not aware that chlorine is a chem-

Contributions to chemical theory

ical element, and experiments designed to reveal oxygen in chlorine failed.

In 1810 and 1811, he lectured to large audiences at Dublin (on agricultural chemistry, the elements of chemical philosophy, geology) and received £1,275 in fees, as well as the honorary degree of LL.D., from Trinity College. In 1812 he was knighted by the Prince Regent (April 8), delivered a farewell lecture to members of the Royal Institution (April 9), and married Jane Apreece, a wealthy widow well-known in social and literary circles in England and Scotland (April 11). He also published the first part of the *Elements of Chemical Philosophy*, which contained much of his own work; his plan was too ambitious, however, and nothing further appeared. Its completion, according to a Swedish chemist, J.J. Berzelius, would have "advanced the science of chemistry a full century."

His last important act at the Royal Institution, of which he remained honorary professor, was to interview the young Michael Faraday, later to become one of England's great scientists, who became laboratory assistant there in 1813 and accompanied the Davys on a European tour (1813–15). By permission of Napoleon, he travelled through France, meeting many prominent scientists, and was presented to the empress Marie Louise. With the aid of a small portable laboratory and of various institutions in France and Italy, he investigated the substance "X" (later called iodine), whose properties and similarity to chlorine he quickly discovered; further work on various compounds of iodine and chlorine was done before he reached Rome. He also analyzed many specimens of classical pigments and proved that diamond is a form of carbon.

Public service

Shortly after his return, he studied, for the Society for Preventing Accidents in Coal Mines, the conditions under which mixtures of firedamp and air explode. This led to the invention of the miner's safety lamp and to subsequent researches on flame, for which he received the Rumford medals (gold and silver) from the Royal Society and, from the northern mine owners, a service of plate (eventually sold to found the Davy Medal). After being created a baronet in 1818, he again went to Italy, inquiring into volcanic action and trying unsuccessfully to find a way of unrolling the papyri found at Herculaneum. In 1820 he became president of the Royal Society, a position he held until 1827. In 1823–25 he was associated with the politician and writer John Wilson Croker in founding the Athenaeum Club, of which he was an original trustee, and with the colonial governor Sir Thomas Stamford Raffles in founding the Zoological Society and in furthering the scheme for zoological gardens in Regent's Park, London (opened in 1828). During this period, he examined magnetic phenomena caused by electricity and electrochemical methods for preventing salt-water corrosion of copper sheathing on ships by means of iron and zinc plates. Though the protective principles were made clear, considerable fouling occurred, and the method's failure greatly vexed him. But he was, as he said, "burned out." His Bakerian lecture for 1826, "On the Relation of Electrical and Chemical Changes," contained his last known thoughts on electrochemistry and earned him the Royal Society's Royal Medal.

Davy's health was by then failing rapidly; in 1827 he departed for Europe and, in the summer, was forced to resign the presidency of the Royal Society, being succeeded by Davies Gilbert. Having to forgo business and field sports, Davy wrote *Salmonia: Or Days of Fly Fishing* (1828), a book on fishing (after the manner of Izaak Walton) that contained engravings from his own drawings. After a last, short visit to England, he returned to Italy, settling at Rome in February 1829—"a ruin amongst ruins." Though partly paralyzed through stroke, he spent his last months writing a series of dialogues, published posthumously as *Consolations in Travel, or the Last Days of a Philosopher* (1830). He suffered a further stroke and died in Geneva on May 29, 1829.

BIBLIOGRAPHY. J.A. PARIS, *The Life of Sir Humphry Davy* (1831), for more than a century the standard work on Davy's life; J. DAVY, *Memoirs of the Life of Sir Humphry Davy* (1836), a biographical work with emphasis on social conditions in early 19th-century Britain; T.E. THORPE, *Humphry Davy: Poet and Philosopher* (1896), primarily an analysis of the scientist's literary work; A. TRENEER, *The Mercurial Chemist: A Life of Sir Humphry Davy* (1963), chiefly concerned with the development of Davy as a scientist; SIR H. HARTLEY, *Humphry Davy* (1966), an extremely useful account of Davy's life that complements Paris' biography, but with changed emphases resulting from the passage of 135 years.

(F.W.G.)

Dead Sea

The Dead Sea, actually not a sea at all but a landlocked lake between Israel and Jordan, is the lowest body of water on earth, at approximately 1,296 feet (395 metres) below sea level. Its northern half belongs to Jordan; its southern half is divided between Jordan and Israel. After the 1967 Arab–Israeli War, however, the Israeli Army remained in occupation of the entire west shore. The Dead Sea, Hebrew Yam ha-Melaḥ ("Salt Sea"), Arabic al-Baḥr al-Mayyitt ("Sea of Death") or Buḥayrat Lūṭ ("Sea of Lot"), lies between the hills of Judaea to the west and the Transjordanic plateaus to the east. The Jordan River flows from the north into the Dead Sea, which is 50 miles long and attains a width of 11 miles. Its surface area is 405 square miles (1,049 square kilometres), and its maximum depth is 1,300 feet (400 metres). The peninsula of al-Lisān (the "tongue") divides the lake on its eastern side into two unequal basins; the northern basin measures 294 square miles and reaches a depth of 1,250 feet; the southern basin is smaller (99 square miles) and shallower (20 feet on the average). During biblical times and up to the 8th century AD, only the northern basin was inhabited, and the lake was 130 feet below its level of the early 1970s. It then rose to reach its highest level (34 to 36 feet above the 1970s level) around 1890 to 1900. In the 20th century its level has fallen; *e.g.*, 12 to 13 feet between 1935 and 1963. (For an associated physical feature, see JORDAN RIVER; for further historical aspects, see NEAR EAST, ANCIENT; SYRIA AND PALESTINE, HISTORY OF.)

Changes in surface levels

Historical and religious associations. The name Dead Sea can be traced back at least to the Hellenistic epoch (323 to 30 BC). It has been associated with biblical history since the time of Abraham (progenitor of the Hebrews) and the destruction of Sodom and Gomorrah (the two cities, according to the Old Testament, that were destroyed by fire from heaven because of their wickedness; the city sites are now submerged in the southern part of the Dead Sea). The desolate rivers of the lake offered refuge to David (king of Israel) and later to Herod I the Great, king of Judaea, who at the time of the siege of Jerusalem by the Parthians in 40 BC barricaded himself in a fortress at Masada. Masada was the scene of a three-year siege that culminated in the mass suicide of its Jewish Zealot defenders and the destruction of the fortress by the Romans in AD 73. The Jewish sect that left the biblical manuscripts known as the Dead Sea Scrolls, took shelter in caves to the northwest of the lake. (P.Sa.)

Physiography and geology. The Dead Sea occupies the lowest part of the 350-mile-long Jordan–Dead Sea trench, which is a northern extension of East Africa's rift valley. It is a sunken block confined by two parallel geological faults. The eastern fault, along the edge of the Moab Plateau, is more readily visible from the lake than is the western fault, which marks the gentler Judaean upfold.

In the Cretaceous and Jurassic periods (from 65,000,000 to 190,000,000 years ago), before the creation of the trench, an extended Mediterranean Sea covered Syria and Palestine. During the Miocene Epoch (from 7,000,000 to 26,000,000 years ago), upheaval of the sea bed produced the upfolded structures of the Transjordan highlands and the central range of Palestine, causing the fractures that formed the Dead Sea depression. At that time, the Dead Sea was probably the same size as now. During the Pleistocene Epoch (from about 10,000 to 2,500,000 years ago) it rose to a height of about 700 feet above its modern level, forming a vast inland sea, stretching 200 miles from the 'Emeq Hula (Huleh Valley) area

Formation of the Dead Sea

in the north to 40 miles beyond its present southern limits. The Dead Sea did not spill over into the Gulf of Aqaba because it was blocked by a 100-foot rise in the highest part of ha-ʿArava (Wādī al-ʿArabah), a seasonal watercourse that flows along an eastern extension of the central Negev highlands.

About 2,500,000 years ago or less, heavy stream flow into the lake deposited thick sediments of shale, clay, sandstone, rock salt, and gypsum. Later strata of clay, marl, soft chalk, and gypsum were dropped upon layers of sand and gravel. With the water evaporating faster than it was replenished by precipitation over the last 10,000 years, the lake gradually shrank to its present form. In so doing, it bared deposits that cover the Dead Sea Valley to a thickness of from one to four miles.

Al-Lisān and Mt. Sedom

Al-Lisān and Mt. Sedom are formations that resulted from movements of the earth's crust. Mt. Sedom's (historically Mt. Sodom) steep cliffs rise up from the southwestern shore. Al-Lisān is formed of strata of clay, marl, soft chalk, and gypsum interbedded with sand and gravel. Both al-Lisān and beds made of similar material on the western side of the Dead Sea Valley dip to the east. It is assumed that the uplifting of Mt. Sedom and al-Lisān formed a southern escarpment for the Dead Sea. Later, the sea broke through the western half of this escarpment to flood what is now the shallow southern end of the Dead Sea. (S.B.Co.)

Climate. The Dead Sea is situated in a desert. Rainfall is scanty and irregular. Al-Lisān has about 2.5 inches (six centimetres) of rain a year and Sedom (near historical Sodom) only about two inches. Because of the very low altitude and the sheltered location, winter temperatures are mild and pleasant, averaging 63° F (17° C) at Sedom and 58° F (14° C) at the northern end in January; freezing temperatures are unheard-of. Summer, on the other hand, is very hot, averaging 93° F (34° C) in August at Sedom, with a maximum temperature of 124° F (51° C). Evaporation of the lake's waters—estimated at 55 inches (140 centimetres) a year—often creates a thick mist above the lake. On the rivers the atmospheric humidity varies from 45 percent in May to 62 percent in October. Local winds, which are relatively constant, blow off the lake in all directions in the morning and toward the centre of the lake at night.

Hydrology. The inflow from the Jordan River, whose high waters occur in winter and spring, averages 19,000,000,000 cubic feet per year. Four modest but perennial streams descend from Jordan on the east through deep gorges: Wādī al-ʿUẓaymī, Wādī Zarqāʾ Māʿīn, Wādī al-Mawjib, and Wādī al-Ḥasā. Down numerous wadis (seasonal watercourses) streams flow spasmodically and briefly from the neighbouring heights as well as from the depression of ha-ʿArava. Thermal sulfur springs also feed the rivers. Evaporation in summer and the intake of water, especially in winter and spring, cause seasonal variations in the level of the lake of from 12 to 24 inches.

The water layers

Salinity. The waters of the Dead Sea are extremely saline, and the concentration of salt increases toward the bottom. In effect, two different masses of water exist in the lake. Down to a depth of 130 feet, the temperature varies from 66° to 98° F (19° to 37° C), the salinity is slightly less than 300 parts per thousand, and the water is particularly rich in sulfates and in bicarbonates. After a zone of transition located between 130 and 328 feet, the lower waters have a uniform temperature of about 72° F (22° C) and a higher degree of salinity (approximately 332 parts per thousand); they contain hydrogen sulfide and strong concentrations of magnesium, potassium, chlorine, and bromine. The deep waters are saturated with sodium chloride, which is precipitated to the bottom. The lower waters are fossilized (*i.e.*, being very salty and dense, they remain permanently on the bottom); the upper waters date from a few centuries after biblical times.

The water is saline and has a high density. Bathers float on it easily. The fresh water of the Jordan stays on the surface; in the spring its muddy colour can be traced across the lake as far as 30 miles south of the point where the river empties into the Dead Sea.

The extreme salinity excludes any animal or vegetable life except bacteria. Fish, which are carried in by the Jordan or by smaller streams when in flood, die instantly. Apart from the vegetation along the rivers, the only plant life is discontinuous and consists mainly of halophytes (plants that grow in salty or alkaline soil).

Resources. The Dead Sea constitutes an enormous salt reserve. In particular, salt deposits occur in the mountains of Sedom along the southwest shore. The salt has been exploited on a small scale since antiquity. In 1929 a potash factory was opened near the mouth of the Jordan at Kāliyā. Subsidiary installations were later built at Sedom. During the 1948-49 Arab–Israeli War, the factory at Kāliyā was destroyed. A factory was opened in Sedom in 1955 by the Dead Sea Works Ltd., producing potash, magnesium, and calcium chloride. Another plant produces bromine and other chemical products.

Production of chemicals at Sedom

Because its situation has been on the contested Jordan–Israeli frontier, the Dead Sea has not been used to a great extent for navigation. The river shores are deserted, and permanent establishments are rare. Exceptions are the factory at Sedom, a few hotels and spas at Kāliyā, and, in the west, a kibbutz (an Israeli agricultural community) in the region of ʿEn Gedi. Small cultivated plots are also occasionally found on the lakeshore.

Prospects. The establishment of the kibbutz at ʿEn Gedi has demonstrated that the soil of the region can be cultivated if it is desalted and certain precautions taken. The savage grandeur of the site and the extreme mildness of the winters favour the development of tourism and the establishment of health resorts. Mineral exploitation is expected to increase, with the factory at Sedom—which has been enlarged—producing more potash as well as other chemical products. For its part, Jordan envisages the eventual exploitation of potash and its shipment by rail to the Gulf of Aqaba. (P.Sa.)

BIBLIOGRAPHY

Preliminary exploration: A.G.C. MOLYNEUX, "Expedition to the Jordan and the Dead Sea," *Geogrl. J.*, 18:104–130 (1848); W.F. LYNCH, *Official Report of the U.S. Expedition to Explore the Dead Sea and the River Jordan* (1852); L. LARTET, *Exploration géologique de la mer Morte, de la Palestine et de l'Idumée* (1878).

Geology: L. PICARD, *Structure and Evolution of Palestine, with Comparative Notes on Neighboring Countries* (1943).

The Dead Sea trench: L. DUBERTRET, "Remarques sur le fossé de la mer Morte et ses prolongements au Nord jusqu'au Taurus," *Revue Géogr. Phys. Géol. Dyn.*, 9:9–16 (1967); A.M. QUENNEL, *Tectonics of the Dead Sea Rift, XXe Session Intern. Geol. Cong.*, pp. 385–405 (1959).

Hydrology and climatology: Y.K. BENTOR, "Some Geochemical Aspects of the Dead Sea and the Question of Its Age," *Geochim. Cosmochim. Acta*, 25:239–260 (1961); C. KLEIN, *On the Fluctuations of the Level of the Dead Sea Since the Beginning of the 19th Century*, Israel Ministry of Agriculture, Water Commission, Hydrological Service, Hydro. Paper, no. 7 (1961); D. NEEV and K.O. EMERY, *The Dead Sea: Depositional Processes and Environments of Evaporite*, Israel Geol. Surv. Bull., no. 41 (1967).

Botany: M. ZOHARY and G. ORSHANSKY, "Structure and Ecology of the Vegetation in the Dead Sea Region of Palestine," *Pal. J. Bot.*, 4:178–206 (1949).

Mineral extraction: D.H.K. AMIRAN and Y. KARMON, "The Expansion of the Dead Sea Works," *Tijdschr. Econ. Soc. Geogr.*, pp. 210–223 (1964).

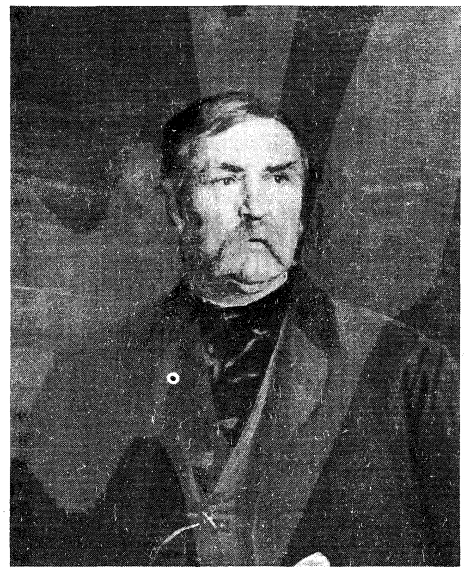
(P.Sa./S.B.Co.)

Deák, Ferenc

The wise guidance of the Hungarian statesman Ferenc Deák during the negotiations that led to the establishment of the dual monarchy of Austria-Hungary in 1867 earned for him the sobriquet "the sage of the country." He was born on October 17, 1803, at Söjtör, in West Hungary, where his father owned substantial estates. After graduating in law, he entered the administrative service of his county of Zala, which in 1833 sent him to represent it in the Hungarian Diet, in place of his brother, who had resigned his mandate. At that Diet and those of 1839 and 1841, Deák made his mark as a leader of the growing reform movement for the political emancipation and internal regeneration of Hungary. Although re-elected in

1843, he declined his mandate in protest against the scandalous manner in which the election had been conducted, but by this time his unimpeachable integrity, his unvarying good sense, and his encyclopaedic knowledge of the law had made him the most generally respected figure in his camp. In was he who, in 1847, drafted for the "national opposition" its program of reform for the Diet of 1847. Ill health prevented him from seeking election to that body, but when, in March 1848, the crown sanctioned the establishment of an independent Hungarian ministry, the new minister president, Count Lajos Batthyány, insisted on his taking the portfolio of justice.

By courtesy of the Hungarian National Museum, Budapest



Deák, detail of an oil painting by Bertalan Székely (1835–1910). In the Magyar Nemzeti Múzeum, Budapest.

Drafting of the April Laws, 1848

As minister of justice, Deák was mainly responsible for the drafting of the reforming "April Laws" sanctioned by the Monarch on April 11. During the next months, he took part in most of the negotiations between the Hungarian government and its opponents. He ceased to be minister when Batthyány resigned on September 28, and he refused reappointment, but in January 1849 he was a member of another mission that sought to intervene with Alfred, Fürst zu Windischgrätz, commander of the Austrian armies that had occupied Buda. Prevented from rejoining the Diet, he retired to his family estates, where he lived quietly for some years, unmolested by the Austrian authorities, who had decided that his activities had not been treasonable, but refusing all invitations to collaborate with a regime that he condemned as illegal. In 1854 he sold his estates—donating the bulk of the proceeds to his sister—and moved to Pest, taking up quarters in the Hotel Angol Királyné ("Queen of England"), which remained his home until his last illness.

In Pest, Deák became the oracle of all those who sought a practical and honourable way out of Hungary's political impasse. He maintained unwaveringly that the April Laws were legally valid, and that any regime that did not recognize this was itself illegal and was to be treated as such. His inflexible insistence on this principle in the Diet of 1861 caused the Monarch to dissolve the Diet and to reimpose an absolutist regime. But Deák also maintained that the laws were entirely compatible with the integrity of the monarchy and could be made the basis of a settlement satisfactory alike to the crown and the nation—further, that such a settlement was a vital interest of both parties. As his international difficulties gradually convinced Emperor Francis Joseph of the need of reconciliation with Hungary, Deák, in private conversations and public utterances, notably his famous "Easter Article" of April 16, 1865, put forward Hungary's conditions in terms that presently led to the "Compromise" (*Ausgleich*) of 1867, by which the dual monarchy was established.

Deák was beyond question the father of the Compromise. The machinery that it embodied was not all of his devising, but it was his faith that a constitutionally satisfied Hungary and a strong monarchy could and must coexist, that made the agreement possible; nor could agreement have been reached without his sagacity, resolution, and integrity, and the faith that these qualities inspired in both camps.

His adherents during the negotiations had styled themselves the "Deák Party," and they retained the name during the next years, when they formed the government. Deák, however, never looked on himself as a party man nor acted as a party leader. He left questions of party organization to the minister president of the day (he had again declined all ministerial office for himself), and while helping to complete the legislation deriving from the Compromise and defending it in Parliament when necessary, he thought and spoke only as guardian of the nation's interests, sometimes taking sides against the Deák Party itself on secondary issues. His interventions grew fewer as his health failed. His last public speech was made in November 1873. He died on the night of January 28–29, 1876.

Assessment of Deák's achievements

The Compromise was the crowning achievement of Deák's life work, but there was hardly a field of public life in which he did not exercise his influence and always on the side of humanity, reasoned reform, and good sense. He was a foremost fighter for the reform of the conditions of the Hungarian peasantry, setting the example on his own estates. Other fields in which he took keen interest included penal reform and interfaith relations. His moderation and sense of rectitude made him a regular choice when delicate questions had to be negotiated. He was a singularly modest man, avoiding all publicity and refusing all honours and offices, except in 1848, when his portfolio was pressed on him. His speeches were unemotional but carried conviction through their lucidity and their impeccable reasoning. His private life was entirely unpretentious. He never married.

BIBLIOGRAPHY. Deák destroyed all his papers before his death. His speeches and some of his letters were published (with an excellent introduction) by M. KONYI, *Deák Ferencz beszédei*, 6 vol. (1882–98). The only adequate biography of him is that by Z. FERENCZI, *Deák élete*, 3 vol. (1904).

(C.A.M.)

Death

Death, often defined simplistically as the absence of life, has always been viewed with mystery, superstition, and fascination by man. That definition is faulty, for it presents a negative: death is not life. What it is, however, can only be conjectured. It is the supreme puzzle of poets, the concern only of humankind, and the inevitable fate of all things living. This article attempts to summarize briefly and primarily in human terms the significance of death, the signs of death, and the responses to death. Statistics on death rate and other actuarial aspects of death are found in the article POPULATION.

THE CONCEPT OF DEATH

Man knows that things die. He witnesses their deaths and sometimes even executes them. These traumatic deaths are easy to understand. A complex living system is quickly and radically destroyed, and it ceases functioning. Natural death, the main subject of this article, is much more difficult to grasp.

Perhaps when man first discovered that certain slumbers were irreversible, he questioned the difference between that profound sleep and the daily slumbers he experienced. Manifest in ritual, symbol, and philosophical perspective are countless efforts throughout history to rationalize this final reality, the philosophical "why" of which still defies explanation, but the biomedical "how" of which is yielding to investigation.

Biological death. In the economy of nature one generation succeeds another, living its span, then yielding to the next. Those organisms survive that are best adapted to the prevailing environmental conditions. When those conditions change, the organisms least adapted may no

longer be capable of surviving and may migrate to more favourable surroundings or may die. But nature is not shortsighted. Organisms are often very productive before their death and in some cases appear to die in a frenzy of reproductive activity, as when plants, grown on poor soil or during a period of drought, flower and fruit, seemingly in a last effort to provide successors to carry on the line. Among those offspring may be many new sorts potentially capable of surviving under a variety of conditions hostile to the parental organisms. In this manner nature clears the way for innovation. Death moves exhausted organisms to another phase of the cycle of nature and allows potentially more successful combinations of genes to supplant them. Death, therefore, is one of nature's ways of improving life; it provides a changing stage on which natural selection can try new experiments in evolution.

The biological significance of death

Human death. Primitive man lived in a world beset by death, to the degree that fertility and other life symbols often were the focal point of his worship. He did not deny death's reality but interpreted death as a transition from one phase of life to another. Burial, then, became a rite of passage similar to puberty, from one mode of participation in life to another. The contents of early tombs bear witness to this view of death among prehistoric cultures.

In later times, the afterlife was envisioned in more ethereal and spiritual terms, as exemplified in early religions. The Judeo-Christian tradition distinguishes itself from others in terms of its historicity, emphasizing the entrance of God's revelation as events (*e.g.*, Passover, crucifixion, resurrection) that at one and the same time establish temporality and future possibility.

Much serious misunderstanding exists relative to the biblical understanding of death and resurrection, upon which much of the view of Christianity is said to rest, and with which modern natural science ostensibly is said to be in conflict. The most obvious misunderstanding is in the notion of physical resuscitation of the body, or belief in the physical conquest of death. Theologians have sought to reinterpret the concept of resurrection in language that simultaneously affirms and denies the exclusiveness of flesh or spirit as the biblical understanding of resurrection. Saint Paul, in the earliest writings on resurrection, uses the Greek term *sōma* (body), which would correspond much more closely to modern terms such as identity, ego, or even gestalt than to either flesh or spirit. The biblical witness stresses, therefore, a real but not a physical conquest over death. Later, as theology developed historically and particularly as folk religion struggled to come to terms with the agnostic and Eastern religions, the emphasis often shifted to either a purely physical phase change or a total spiritualizing of the death event. Contemporary theology seems to be developing a more profound yet realistic conception of death as the foreclosure of life, a point at which God calls man into the future. Thereby hope becomes the alternative to unempirical despair.

THE BIOMEDICAL ASPECTS OF DEATH

There is little argument about the determination of the natural death of a subhuman animal. Its heart fails or it ceases breathing and is pronounced dead. A loved pet may, of course, be given extraordinary medical care to keep it alive, but not for long. It is only the human value of the level of life that dictates the criteria for determining the moment that life ends. For this reason a precise definition of human death is required.

Although death as a concept continues to elude the philosopher and the theologian, its physical signs are yielding to codification by the biologist and the physician. Except for immediate death by overwhelming injury or accident, higher organisms die gradually, from their cells and tissues to their organs and vital systems, ending finally in collapse of the entire body.

Among lower organisms, the commonly accepted criteria of life, which include such functions as respiration, transfer of vital substances (*e.g.*, water, ions, metabolites), and reduplication, are not manifest at all times. Viruses outside of a host cell, for example, fail to show any of these criteria and can therefore be regarded as dormant proteinaceous particles containing ribonucleic acid (RNA) or deoxyribonucleic acid (DNA). When introduced into a host cell, the virus is capable of reproduction and renewed life. Under unfavourable environmental conditions, fungi and bacteria are able to assume an arrested stage wherein the criteria of the living cell are absent. These periods of dormancy offer no information about death as such; they simply suggest that the traditional criteria for determining death are not adequate.

Death versus dormancy

Death of cells and tissues. Death at the cellular level is as obscure as death of the whole organism, but arbitrary signs have been generally agreed upon to judge the limits of life.

The cell appears to be the smallest functioning unit of all many-celled organisms. It is a membrane-bound mass of cytoplasm containing a membrane-bound nucleus. It has the ability to grow, assimilate, and undergo division, reproducing itself under normal conditions. The cell can, therefore, be regarded as a two-phase system. The nucleus, directing cell growth, development, and division, represents one phase. The cytoplasmic mass, representing the second phase, controls the absorptive, secretory, and transport functions of the cell. With the exception of nerve cells, all cells of the body are constantly being replaced, some more rapidly than others. Death of certain tissues or localized areas, from a variety of causes, may lead to death of the body as a whole. But the point at which the cell's self-regulatory, or homeostatic, mechanism undergoes irreversible damage, with subsequent death of the cell, continues to evade discovery. What is injurious to one cell may be of little or no consequence to another cell in the same organ or in other organs. Chloroform, for example, affects the more centrally positioned cells of the lobes of the liver, while phosphorus affects the peripheral cells of those lobes. Nerve cells show a selective vulnerability to brief periods of oxygen deprivation (hypoxia). The nerve cells of a part of the hippocampus (Sommer sector of the Ammon's horn) and the Purkinje cells of the cerebellum swell initially when deprived of oxygen but undergo irreversible shrinkage degeneration if hypoxia is prolonged. The nerve cells of the medulla or spinal cord will show little or no changes unless the hypoxic episode is very severe and sustained. The point at which the cell undergoes the irreversible change and begins to die is unknown.

In certain tissue culture preparations, injured cells lose protein from both the cytoplasm and nucleus. The degree of protein loss often appears to be proportional to the degree of injury. The mechanism controlling this loss remains unknown. The injured cells also exhibit a reduction in cell division but show little initial change in structure. These findings suggest that one pathway leading to cell death would be interference with protein synthesis, resulting in eventual protein deficiency.

At the electron microscopic level, initial cell damage is seen early, in contrast with light microscopic observation in which several hours may pass before structural changes can be recognized. The fine structures of the cell swell and become disrupted, leading to death via several processes. The cell, after swelling, may rupture, extruding its contents into the surrounding tissue (cytolysis). The nucleus alone may swell and rupture (karyolysis), fragment (karyorrhexis), or shrink (pyknosis).

Death of the entire organism. From the earliest days it was known that death of all parts of a multicellular organism did not occur simultaneously. Since the heart was considered to be the central organ, its cessation of function was used to mark the beginning of death for other vital organs. Historically, clinical death has been based on the absence of the heartbeat (and therefore the peripheral pulse) and the absence of breathing (with resultant bluing of the extremities, mouth, and lips). A lack of certain reflexes of the eye is also noted when death is further advanced, with absolute signs, such as algor, rigor, and livor mortis, finally becoming apparent. Algor mortis represents the fall of body temperature to that of the environment. Rigor mortis is the rigidity of skeletal

The traditional definition of clinical death

muscles. Finally, livor mortis is the purple-red discoloration of parts of the body as a result of the settling of blood.

The essential components of most mammalian organisms include nervous, circulatory, respiratory, gastrointestinal, excretory, endocrine, and supportive organ systems. In general, parts of one or more of these systems may be injured for variable periods of time without causing irreversible changes. There is a hierarchy, however, *i.e.*, some of these systems are much more essential than others. Large portions of bone, for example, may be lost forever, and a kidney may cease functioning for several hours; but if organs such as the heart or the lungs stop working for only a few minutes, irreparable damage may occur. Cardiac or even respiratory standstill does not, however, automatically mean the onset of clinical death. Through the use of respirators, pacemakers, and other sophisticated apparatus life can be extended greatly. It is therefore possible to sustain the function of both the heart and the lungs even in the face of total loss of brain activity. Life of a sort may continue, but it is not life as the term is normally understood. It would appear that the definition of death should be more closely related to the determination of brain (cerebral) death, or at least to the condition of irreversible coma (see *The determination of death*, below).

THE PROSPECT OF DYING

The awareness of death is difficult to study objectively. The reality of death for the individual is not dwelt upon but is usually—and thankfully—of momentary concern. Although each person may be reminded of death frequently enough, the personal reality of it is often remote. Studies on patients with terminal illnesses who face death imminently have revealed much of interest and psychosocial value regarding the personal response to the approach of death.

The stages of dying. Terminally ill patients often go through a series of five stages in accepting the reality of their own finiteness and oncoming death. Most patients initially respond with temporary shock and a mental denial to the first awareness of a fatal, or potentially fatal, illness. Informed about the diagnosis, a patient may not be able to "hear" it and may go from hospital to hospital in an attempt to find someone who will reassure him that his condition is not serious. Eventually, he will face the reality, but often only for brief periods of time, to provide a last will and perhaps make arrangements for dependents.

Sooner or later, however, he will drop his denial and enter the second stage, the stage of anger, during which he complains about any service anyone renders and is difficult to deal with; thus he is often avoided, which only increases his rage. This is the stage at which he no longer says "no, not me" but instead "why me?" His anger is displaced to the nurse, the physician, the family members, in fact to all the people around him who represent life, energy, and functioning—the very qualities the patient is in the process of losing. The anger also includes God, who "is not listening" to the patient's pleas for help and good health. Patients who are not made to feel guilty or ungrateful during this difficult time—who are encouraged to express their anger—will soon proceed to the third stage, the stage of bargaining.

Bargaining for life

With the realization that he is to die soon, the patient offers a token in exchange for prolongation of life, usually in form of prayers: "if you give me one more year of life, I will become a good Christian," or "I will go to the Synagogue every day," or "I will make a pilgrimage. . . ." This is the time of truce, psychologically speaking. The patient is not at peace, though he is more comfortable emotionally and believes that he has a certain extension of life, no matter how brief. He will say "yes, it is me, but . . .," and only when the bargaining time is up will he enter the fourth stage, one of depression. He now faces that certainty that it is truly his life that is coming to an end. He has responded already to the losses he has experienced (an organ perhaps, a job, being home with the family) and gradually descends into deep and silent depression called the preparatory grief. This is the time when he faces the impending loss of everybody and everything he has ever loved. He begins to separate from the living and finally wants only the next of kin by his side. When he has finished all his business, has ventilated his anger and expressed his grief, he will be able to reach the final stage, of acceptance.

A patient who has no moral support and help during this crucial period of his life or who is isolated in a hospital or nursing home will often remain merely resigned to death. He is bitter, unhappy, and often complains, frequently to the very end of life. By contrast a patient in a true stage of acceptance is almost void of feelings; he may say simply: "My time comes very close and it is all right." If no extraordinary means are employed to prolong his life at this time, the patient is prepared to die with peace and dignity.

Three basic requirements are necessary to assist chronically ill patients to reach this final stage of equanimity: (1) excellent nursing care with adequate pain relief, (2) a flexible management in order to gratify simple needs (lifting severe restrictions against food, drink, and cigarettes), and (3) a hospital staff and a family who themselves accept the reality that this patient is dying and who can openly talk about it, if and when the patient expresses a desire to talk about it.

The patient's hope, originally associated with cure, treatment, or prolongation of life, is then diverted to hope for God's acceptance, for maintaining a sound mind to the end, or simply for the well-being of the children or family left behind.

A patient with a genuine faith (of whatever religion, including those that do not believe in an afterlife) dies with less turmoil than one who lacks belief. A person raised in a rural environment or in a non-death-denying society faces death with more equanimity than a person who grew up in an urban environment, where death occurs in hospitals, where children are "protected" from sharing this experience in a community, and where mortuaries attempt to disguise the reality of death with heavy makeup and other forms of denial.

Family members and attending staff also proceed through these five stages, when personally involved in the care of a dying patient. Parents of dying children and anyone who loses a loved one suddenly and unexpectedly may go through this adjustment after the death occurs—often in a prolonged manner and requiring more time to reach a true stage of acceptance. Widows and widowers often die soon after the death of the spouse, during their own grief process, which apparently renders them more vulnerable to death, often without reaching the stage of acceptance.

Awareness of impending death

A patient is often able to convey his awareness of impending death and will communicate this knowledge to those around him if they are listening. He may use verbal or nonverbal communication. He may comment simply: "My time is very close now and I am ready." He may use symbolic language: "I had this fantasy of a train running rapidly down the hill toward the end of the tracks and I had an argument with the trainmaster. I demanded that he stop the train 1/10th of an inch short of the end." This patient's request for a tenth of an inch represented his bargaining for a very short extension of time to complete some unfinished business. When he accomplished his goal, he died very peacefully a few days later. It is not known exactly how patients realize the moment of impending death, but it is a fact that many of them are able to postpone death psychologically or to hasten it (without committing suicide) when they no longer desire to struggle for life.

The fear of death. The fear of death is universal and consists of two parts. Consciously, people may be afraid of the pain and suffering associated with dying, the separation from the living, or simply the state of nonbeing or nothingness. Fear of the unknown is frequently mentioned as the most outstanding fear. Much more significant, however, is a fear heavily repressed and often revealed through dream material or, in the case of children through play or in drawings. The often-repressed

fear of death is seen as a fear of catastrophic destructive force bearing down and leaving the victim with a sense of utter helplessness, impotence, and rage. It is the fear of a traumatic annihilation that can be seen in children even before they have an empirical knowledge of death. Perhaps this fear originates in attitudes formed during the infantile stage that precede those attitudes imposed by culture or religion. Whether anxiety of death is an inherent property or not, it seems clear from clinical evidence that the early mother–child relationship is an essential source of the death complex in its catastrophic aspects.

Historically, a catastrophic death was also seen in the form of epidemics, which men attempted to conquer. Since death is still popularly conceived of not as a natural ending here on earth but simply as being killed, man learned to kill rather than be killed. He conquered epidemic diseases, he added years to the life-span of man, he learned to transplant organs, and he is now on the verge of actually creating life. In his fear of death, however, man also created weapons of mass destruction (the atomic bomb, chemical and bacteriological warfare) and, for the first time in the history of mankind, he created a concrete reality of what man associates with the fear of death—a catastrophic destructive force that hurls from nowhere and against which there is no defense. This is most likely a contributing factor to the extreme death denial of the current society.

Individuals as well as entire societies believe that "death will happen to you and to you, but not to me." Thus man is able to function and to plan for the future despite the fact that his own weapons of mass destruction may be used against him. With an increased availability of such destructive devices, the anxiety increases, the denial of his own finiteness becomes more difficult to maintain, and the psychological and financial expenses multiply in order to maintain a sense of security (distant wars, antiballistic missile systems). The "generation gap" may be based not upon father–son differences on matters of progress in science, technology, and medicine, but on weapons of mass destruction and death denial.

THE DETERMINATION OF DEATH

The medical aspects. Is a patient alive who has only his vegetative functions and no consciousness of his being? Is not a patient virtually dead who has no brain activity whatsoever? And if he is in irreversible coma, can his organs be removed for transplantation into a conscious body? These questions and even more vexing ones (regarding possible future brain transplants) have raised a host of legal, ethical, moral, and religious considerations that require resolution.

The revised definition of death

A standardized diagnosis of clinical death is vitally needed. One such definition holds that death is equated with irreversible coma. In the late 1960s several medical conferences met with the primary purpose to define irreversible coma as a new criterion for death. It was recommended by these groups that a comatose patient be regarded as dead, if he has a permanently nonfunctioning brain, nonreceptive to external stimuli and showing no spontaneous respirations, muscular movements, or cephalic reflexes. An examination that reveals no brain waves (a flat electroencephalographic recording) lends further confirmatory support. Subsequent investigators have added the requirement of the presence of the so-called agonal angiogram, another examination that reveals a markedly diminished blood circulation. The Ad Hoc Committee of the Harvard Medical School further noted that the conditions previously described must be present for 24 hours. Despite the growing list of signs that must be present before a person can be judged dead, a danger still remains, as noted by the committees that seek to define death. In certain cases these criteria of death do not apply, as in patients that exhibit marked hypothermia (body temperatures below 90° F or 32.2° C) or marked central nervous system depression as a result of drug overdose. Furthermore, the notion of irreversibility is not easily agreed upon. Conditions thought to be irreversible in the 19th century are not invariably so today. Will an irreversible coma become reversible sometime in the future because of biomedical advances?

The legal aspects. Since individual tissues die at different times, and even cerebral death has to be observed over a defined period of time, as dealt with earlier, it is often difficult to speak of a moment of death. An exact time of death is sometimes necessary, however, in order to ensure that an organ is indeed being removed from a donor patient who is clinically dead. Aside from these medical implications, an unambiguous moment of death is often important to establish for legal purposes. In the death of a couple, it may be necessary to know whose death occurred first. It may become important to know the exact time of death in civil and criminal liability as well as in inheritance of property, insurance, or survivor's benefits. Unfortunately, current legal concepts appear inconsistent with the revised definition of death, which is gaining wide medical acceptance.

With technological advance and the increasing demand for organs for use in transplants, it becomes mandatory that more than one physician verify the death of a potential donor; it is especially recommended that a verifying physician not be involved in any later effort to transplant organs or tissues from the deceased individual. This precaution lessens the appearance of self-interest by the physicians involved and obviates later legal disputes. A declaration of death is generally made before disconnecting a life-sustaining machine such as a respirator, thus relieving physicians and hospitals of possible future legal entanglements. A physician—and a hospital—otherwise may become liable for suit and accused of euthanasia, or "mercy killing."

Laws relating to organ donations upon death

Laws concerned with gifts or bequests of organs are becoming more common. Many of them include the principle of the "separation of powers," meaning a complete separation of authority and responsibility between the physicians who care for the prospective donor and those who care for the potential recipient of the organs or tissues of the donor. Many localities with anatomical and tissue gift acts have specific forms and provide for procedures before death to make transplantation more effective. A patient can also bequeath his body to a medical school, but such provisions are not binding on the family.

A series of steps ensue upon the death of a person. The death is usually registered with a governmental department established for maintaining records of births and deaths. In many cases the body cannot be removed or disposed of until such certification of death has been made with the local authority.

A person legally can be presumed dead in many countries after an unexplained absence of seven years. Specific laws relating to the presumption of death and to the disposition of property of a person presumed dead, however, vary widely and cannot easily be generalized. Similarly, legal regulations concerned with the disposition of human remains are too diverse to consider here. In every age, however, and in every culture, such disposition complies with prevailing public interest, moral values, and cultural tradition.

BIBLIOGRAPHY

Philosophical and theological aspects: H. FEIFEL (ed.), *The Meaning of Death* (1959); R.C.W. ETTINGER, *The Prospect of Immortality* (1964); M.M. GATCH, *Death: Meaning and Mortality in Christian Thought and Contemporary Culture* (1969).

Biomedical aspects: D.W. KING *et al.*, "Cell Death, pt. 4, The Effect of Injury on the Entrance of Vital Dye in Ehrlich Tumor Cells," *Amer. J. Path.*, 35:1067–1079 (1959); W.E.D. EVANS, *Chemistry of Death* (1963); AD HOC COMMITTEE OF THE HARVARD MEDICAL SCHOOL, "A Definition of Irreversible Coma," *J.A.M.A.*, 205:337–340 (1968); DONALD CUTLER (ed.), *Updating Life and Death: Essays in Ethics and Medicine* (1969).

Psycho-social aspects: E. KUBLER-ROSS, *On Death and Dying* (1969); J. RHEINGOLD, *The Mother, Anxiety and Death* (1967).

Forensic aspects: M.M. HALLEY and W.F. HARVEY, "Medical vs. Legal Definitions of Death, *J.A.M.A.*, 204:423–425 (1968).

(E.K.-R.)

Death and Gift Taxes

Death and gift taxes are levies imposed on gratuitous transfers of property—that is, transfers made without compensation in either money or its equivalent. In this respect they differ from sales taxes imposed on transfers made in exchange for something of value and from property taxes and capital levies that are based on the mere ownership or possession of property and from income taxes levied on earnings.

Death taxes. Death taxes are of two kinds: those imposed on the property left at death are known as "estate" taxes and those imposed on the acquisition of property from a person who has died are known as "inheritance" taxes. The two kinds of death taxes are sometimes both used in the same system.

Discrimination on basis of relationship

Inheritance taxes generally discriminate on the basis of relationship: there may, for example, be larger exemptions or lower rates for spouses and children. Inheritance taxes are in the nature of "accessions" taxes, a tax on receiving rather than giving. Even the United States estate tax, which is primarily a tax on the estate rather than its acquisition, takes some account of who benefits; anything left to a spouse is tax-free up to one half of the estate, and anything left to charity is entirely exempt.

Gift taxes. Levies on gratuitous transfers between living persons are known as "gift" taxes. These may be integrated with death taxes to make a single structure of taxation applicable to gratuitous transfers. Death and gift taxes are closely related in France, Germany, and Sweden. A complete "accessions" tax would cumulatively combine both inheritance and gifts received. Japan had such a tax in the 1950–53 period. Italy and Colombia have partial accessions taxes.

Origin and history. Death in most countries is considered to be a "taxable event," one that prompts the imposition of a levy by the state. Such imposts at death predate general income and sales taxes. Even the Romans imposed a tax on inheritances. Though a death tax was levied in the United Kingdom in 1694, the present structure dates back to 1779–80, and the prevailing estate duty was introduced in 1894. Death taxes are widely employed in both developed and developing countries.

In the United States there were several short-lived death taxes before the present tax was introduced at the federal level in 1916. Many of the states already had death taxes. In an attempt to reduce the lack of uniformity in the inheritance taxes of the various states and to protect them from federal incursion, the federal government of the United States gives credit for state inheritance taxes against the estate tax up to a certain point. But the lack of uniformity persists. A comparison of death duties in Great Britain, Canada, Australia, and the United States discloses some interesting differences in federal–state relationships. In all these nations, except Great Britain, death taxes are imposed by subnational units (state or provincial governments), as well as by the central government. Canada and the United States have a generally higher federal tax and allow substantial credits for the subnational taxes. In Australia, however, the state taxes are significant in themselves.

Legal and social basis

The controversy about death and gift taxes. *Arguments for.* Death taxes can be supported on both legal and social grounds. Legally, the tax may be considered a fee for the privilege of passing property on after death. A traditional argument is that the state is permitting and facilitating the transfer of property. Even where the property is transferred by operation of law—as in an insurance policy or a joint bank account with survivorship rights—it is the legal system and its enforcement power that makes the transfer possible. Socially, the tax tends to reduce inequalities in the distribution of wealth and provides an opportunity to break up large estates. The heavy death taxation in England has had such a result. Substantial economic benefits may derive from the resulting reduction in the concentration of economic power.

Arguments against. The main argument against large death taxes rests on their possible negative effects on incentives. In this regard, the argument is that the person building up an estate will be less inclined to do so if he knows a large part of it will never go to his heirs. An economic limit to the usefulness of death taxes may be said to exist when the accumulation of wealth is discouraged to the point of hampering the growth of the economy. For the heirs, hardship may also result if forced sales must be made to obtain liquid resources with which to pay taxes. In the case of a family business there is a danger that the business will have to be sold quickly below its actual value or that strangers will be brought in to supply funds, with a resulting loss of family control.

The negative arguments may be countered to some degree. It may be that a person who seeks to provide a certain estate for his heirs will have an incentive to work harder, since he will want to leave a large enough estate to absorb the taxes. The problem of cash requirements may be taken care of in various ways, such as provisions allowing the tax to be paid over a period of years or allowing various insurance arrangements that make liquidation of a business unnecessary.

Economic importance. Death and gift taxes are of greater symbolic than practical significance. Of all taxes, they are among the least productive of revenue in both absolute and percentage terms, and their relative importance has dwindled with the growth of income, sales, and excise taxes. The tax-rate structure alone is misleading, since the many deductions, exemptions, exclusions, and allowances generally reduce the yield drastically. Moreover, most people have no taxable estate whatever and are never subject to the gift tax. The exact yield varies from year to year, but some generalizations may be made. In the United States the federal estate tax generally has produced less than 2 percent of federal tax revenues. In the Irish Republic the percentage is about the same. In Canada the figure has been closer to 1 percent, with the federal government paying 75 percent of the federal estate tax revenue to those provinces that do not levy succession duties. The Australian national estate duty has also produced less than 1 percent of the total national revenue. A similarly small percentage has held in Sweden. Even in Great Britain, where death duties are high, the yield has been only about 3 percent of total tax revenues.

Burden on the taxpayer

A comparison of the actual burden of the tax on the taxpayer is difficult because of the innumerable differences in provisions, but it has been attempted for Great Britain, Canada, Australia, and the United States. Where there is no applicable family relationship, Australian rates on the small estates are heavier than on the large ones. In the middle ranges, Australian and British rates are loosely comparable, as are those of Canada and the United States; but the former two are roughly double those of the latter. In the higher ranges, Great Britain and the United States both take more than Australia and Canada, the key difference being that the top rates are higher and are reached sooner in Great Britain than in the United States. These differences are accentuated when closeness of family relationship is taken into account. In general, it may be said that Great Britain's tax is substantially heavier than the others in the middle and higher ranges. The maximum rates that have prevailed in several countries in recent years are suggestive of the burden at the top (figures are percent): United States, 77; Canada, 54; Netherlands, 54; Germany, 60; Japan, 70; Belgium, 72.6. Italy and the United Kingdom have had a maximum effective rate of 80 percent, and Austria and France have had a maximum effective rate of 60 percent. These are variously "marginal," "bracket," or "overall" rates.

Economic effects. The economic effects of death taxes are, in a word, unimportant by the usual criteria. There is little reason to believe that a person will alter his consumption patterns because of the prospect of death taxes. Those who might consume more quickly to avoid leaving an estate and paying an estate tax are likely to be offset by those who consume less and save more in order to leave enough to their heirs after taxes. The main effect of death and gift taxes is to encourage estate planning and the employment of legal devices to minimize taxes rather than to change behaviour.

The person receiving property after estate taxes will

have less to consume, but the inheritance will generally be of a windfall nature and may have few enduring effects on the heir's lifetime pattern of consumption. In any case, the effect on consumption in the economy as a whole is trivial. Death and gift taxes do tend to achieve a more equal distribution of wealth, but the magnitude of the effect is small in most countries.

Effect on business structure and practices

A significant effect that is often overlooked is that on business structure and practices. This stems from two problems that exist in death taxation. One of these is that of setting a valuation on the estate for tax purposes. Another is the problem of liquidity; *i.e.*, getting the cash with which to pay the taxes.

The valuation of an estate suffers from the same problems as the valuation of any property. There may not be an active market in the various assets involved. The assets may be unique in some sense. There may be a problem of "blockage": a large block of stock may actually be worth more than the apparent market value of the individual shares. A family-owned business may be particularly difficult to evaluate. Appraisers may be called in to set a value on the property, but their evaluations may vary widely. These problems introduce an element of uncertainty into estate planning inasmuch as an individual cannot foresee what value will be placed on various assets he might be holding, especially if they are assets for which there is no active market. It is therefore difficult for him to make adequate preparation for the taxes that may have to be paid. To some extent it also complicates the problem of deciding how much to leave to whom.

Even if there were certainty in the valuation of an estate, there remains the problem of paying the taxes. The gravity of this problem depends partly on the nature of the assets in the estate. If they are shares of stock that are traded on a national stock exchange there is relatively little difficulty. If a family-owned business or a major portion of any business is involved, serious losses, including a change of control of the business, may result from attempts to liquidate the ownership interest in order to obtain cash. Thus the problem of liquidity presents a major consideration in estate planning.

In spite of all efforts to mitigate the impact of estate taxes, there is a strong temptation for an owner to convert his assets into a highly marketable, easily evaluated form while he is alive. This may require their sale to another company, either for cash or for nationally traded securities. The result is a merger or absorption of the smaller company and an increase in the concentration of economic power.

Avoidance of death taxes. On the books, death taxes look highly progressive: the rates of taxation approach 100 percent in the highest brackets in some countries. In fact, however, the opportunities for avoidance are so great that death taxes often have little impact on the enjoyment and use of property passing from one generation to the next.

It is significant that most people do not incur any death or gift taxes at all. But those who are required to pay are liable to pay substantial taxes if advance plans are not made to prevent it. As a result the avoidance of death taxes is a major preoccupation of lawyers and accountants. Refined economic calculation and modern systems analysis techniques may be employed in estate and gift tax planning to minimize the total tax by selecting the optimal estate or will. Efforts at avoidance are highly successful and present a serious problem of equity.

Techniques used to avoid the tax

Family bequests. The best way to reduce death and gift taxes generally is to leave property to a spouse. Under the prevailing United States laws, all that goes to a husband or wife, up to half the total estate, is tax-free; and half of any lifetime gift to a spouse is similarly tax-free. Under inheritance taxes, bequests to a spouse and children are also given preferential treatment. In Canada, under a bill passed in 1969, all gifts and bequests to a spouse are exempt.

Trusts and life estates. A common technique of avoidance is to skip a generation. Rather than leave everything to one's wife, the device is to leave the property to children, with the wife to enjoy the income from the property during her lifetime. The total tax is then generally less than if the estate passed entirely to the wife and from her to the children. Probate of a will is also avoided, since the property passes by virtue of the trust deed—though avoidance of probate does not in itself provide any assurance of avoidance of death taxes.

A wife may have title to the property for life—*i.e.*, a "life estate," which gives her the right to use and enjoy the property for her lifetime. A trust deed is simply a document that provides for putting the property in the hands of trusted persons, such as a bank, a friend, or a lawyer, with instructions to use the property to benefit certain persons (the "beneficiaries"). The trust device is peculiarly an Anglo-American instrument, but comparable devices have arisen under other systems of law.

Gifts between living persons. An obvious way to avoid death taxes is to give one's property away before death. Precisely for this reason, death taxes sometimes take account of "gifts made in contemplation of death" in totalling up the estate for death-tax purposes. Gifts made during life may also be subject to tax at the time they are made. In the United States the gift tax has large exemptions and carries lower rates than the estate tax. Each person has a federal lifetime exemption of $30,000 and can give $3,000 tax-free each year to as many individuals as he wishes. If a spouse joins in the gifts, the annual exemptions are effectively doubled. Those gifts that are taxable are subject to rates that are three-quarters of the estate-tax rates. A few other cases may be noted. The United Kingdom has no gift tax. Canada's gift tax was first introduced in 1935, prior to the federal succession duties of 1941, primarily to preserve the progressivity of income taxes. The Swedish gift and succession taxes both impose their liability on the recipient, and at the same rates. The German tax generally treats all gratuitous transfers alike. France imposes a death and gift duty on the value received by the beneficiary.

Charitable and other contributions. Gifts to charitable, religious, and educational institutions are generally exempt from death and gift taxes. This fact is sometimes used, not only to reduce taxes but to ensure maintenance of family control of a business that would pass to strangers if a large portion of the stock must be sold to pay death taxes. Since charitable contributions are exempt from death and gift taxes, a gift of stock to a charitable foundation can reduce or eliminate the gift or estate tax. The donated stock may be a nonvoting variety, leaving control in the hands of the owners of a small amount of voting stock. The controlling trustees of the foundation may even be members of the family or their agents, hence the family has lost little if anything by this manoeuvre. It has presumably retained sufficient assets after taxes to yield the income it actually needs; for the rest, it has retained family control of the assets and has ensured perpetuation of family control through trustees. It has lost some of the incidents of ownership, but none that really affect its current standard of living or its economic power.

Change of residence. An effective way to avoid death taxes is to move to a "tax haven," a jurisdiction that has no death taxes. Nevada is the only state in the United States without death taxes; but a resident of Nevada is still subject to federal death taxes. There is always a problem of establishing residence, however, and in the case of a wealthy decedent several jurisdictions are likely to claim a tax.

Present trends. Despite their relative unimportance for revenue, death and gift taxes are under constant review for improvement, the main purpose being to achieve equity. The assumption is that glaring inequity jeopardizes the tax system as a whole.

Modification of the estate tax. The death levy has long been regarded as a unique type of impost, to be studied apart from the rest of the tax structure. The current trend is to consider this levy in relation to other taxes, especially those imposed on gifts, income, and capital gains. In this view, any tax that is imposed at death is only one of a series of taxes imposed on the decedent during his life and later to be imposed on the inheritor with respect to the same property.

Solving problems of valuation and liquidity

The problems mentioned above (uncertainty of valuation and need for liquid resources) have led to numerous legal provisions and private arrangements that may have significant economic consequences. The problem of liquidity can be solved by giving an estate a period of time, say ten years, to pay the taxes. Even long-term loans, with an appropriate interest charge, may be desirable. As for valuation, various expedients are possible. The estate may be given some choice in the date of valuation, thereby avoiding a particularly unfavourable market condition. There could be an arrangement whereby the estate-tax authorities commit themselves to particular property valuations on a year to year basis, while the owner is living. In partnerships, fixed prices of ownership interests may be set by agreement. These prices are persuasive, even if not necessarily binding on the authorities. Life insurance can be provided to generate funds to buy out the interests of the decedent. This will give his estate enough cash for taxes and prevent ownership interest from passing into the hands of a stranger. A possible modification is to acknowledge some of these private arrangements more fully in the administration of death taxes.

Proposals to reduce avoidance. The first step in reducing avoidance is a fuller integration of death and gift taxes. The present requirement in the United States is that only "gifts made in contemplation of death" be included in the estate for tax purposes. This involves knowing what was in the mind of a dead person for a specified period before death. Complete integration of death and gift taxation, whether or not further integrated with income taxation, would eliminate such areas of dispute and litigation. A single schedule applying to all gifts, whether during life or after death, would cover the problem. A plan of this sort was enacted in Canada in 1969. Some other instances of integration may be noted. In Italy all gifts that are made after a certain date are included in the estate. Taxes that have been paid on the gifts are credited against the estate tax, the amounts paid as gift tax being cumulative. In Colombia, the taxable estate includes all gifts, but no credit is allowed for any gift tax that has been paid. Germany has a tax on the recipient of gifts and inheritances that allows for a limited degree of cumulation.

It is also necessary to limit the use of the trust device and of gifts to controlled charitable foundations. The latter abuse may be curtailed by stricter surveillance of foundations, including a requirement that the donor retain no control whatever, directly or indirectly. The trust instrument could still be used as a legal method of separating legal ownership and control from a beneficiary. For tax purposes, however, the treatment might be the same as if no trust existed. There would then be a tax both at the death of the original decedent and at the death of the person who enjoys the income for life.

Treatment of capital gains at death

Integration with other taxes. A serious problem in death taxation is the treatment of capital gains at death. In the United States, gains on appreciated property that remains in an estate are forever free of the capital gains tax. The estate does not pay a capital gains tax if it sells the property at the same value it is given for estate tax purposes. Anyone who inherits the property takes it at the appreciated basis for all future capital gains taxes. If the decedent had sold the appreciated property just before he died, he would have had to pay a tax on the gain; the proceeds would then have appeared in his estate and been subject to the estate tax. This appears to represent a pointless inequality of treatment, despite the fact that the full value of the appreciated property will be subject to the estate tax. The large exemptions and deductions in the estate tax may still leave the appreciation free of tax, and the capital-gains-tax rates are very different from the estate-tax rates. Countries that do not have a capital gains tax are not faced with this problem.

One argument against taxing capital gains at death is essentially the one that is used against taxing any capital gains; such gains result from price fluctuations and not from the production of income. Another argument against doing so is that an even greater cash burden would be imposed at death. The introduction of a capital gains tax in Great Britain has created new problems for the family company because it has to make provision for postponed tax liabilities in addition to the provision for death duties. A solution to this lies in making allowance for the capital gains tax liability in computing the amount of death tax due.

The gift tax provisions solve a similar problem by requiring the person who receives the gift to take the donor's initial value as his own. Any appreciation that occurs from the time the original donor acquired the property is subject to the capital gains tax when the property is first sold. A similar idea has been proposed as an alternative to the taxation of capital gains at death. This proposal, currently under discussion in the United States, would impose a capital gains tax on the sale (rather than the acquisition) of inherited property.

An ultimate modification would be the complete integration of income, capital gains, death, and gift taxes. The accretion in one's economic power (*i.e.*, the increase in the value of one's property, including income) would be evaluated and taxed each year. In the extreme form of this plan, the year of death would just be another year, and the disposition of property would itself have no tax consequences or benefits. But the practical problems of valuation throw doubt on the wisdom of any such approach. A recent proposal of the Canadian government would require an evaluation of property every five years for purposes of a capital gains tax.

Exemptions and rate structure. The existing exemptions and rate structure are an expression of social policy toward (1) institutions that are to be favoured, including marriage, schools, charities, and churches, and (2) redistribution of wealth and income. Exemptions or low rates on gifts and bequests to spouse and children represent a social policy in favour of the family unit. This notion is carried to its limit in a proposal recently considered in Canada that all transfers among members of a family unit be tax-free; only when a person leaves the family unit, as by marriage or setting up housekeeping on his or her own, would a tax be assessed. All gifts and bequests would be treated as increasing the economic power of the recipient. Taxation would thus be rationalized in the same way as income taxes. Existing estate and gift taxes would be repealed and all gifts would be brought into the recipient's comprehensive income tax base and taxed at full progressive rates in the same way as wages and other income. All gifts and bequests within the "family tax unit" —husband, wife, and dependent children—would be exempt. Essentially the tax would be imposed on gifts and bequests into and out of the family unit. Provisions for averaging would prevent an excessive tax in a particular year.

Canada in 1969 actually exempted from tax the transfer of property by gift or inheritance between a husband and his wife but raised taxes significantly on transfers to children and others.

Provisions, in tax laws that favour educational, religious, and charitable organizations reflect a desire to encourage private support of these institutions. The social cost of this policy is sometimes very high, and in some instances it is possible for a taxpayer to make a tax-free profit through the donation of appreciated property.

Tax exemptions and rate structure also serve as an instrument for redistributing wealth and income. But, as in the case of the income tax, the methods of avoidance must be considered before any conclusion is drawn concerning the actual impact of death and gift taxes on the distribution of wealth and income.

BIBLIOGRAPHY. Death and gift taxes in various countries are covered in the HARVARD LAW SCHOOL, INTERNATIONAL PROGRAM IN TAXATION, "World Tax Series" (Australia, 1958; Brazil, 1957; Colombia, 1964; France, 1966; Federal Republic of Germany, 1969; India, 1960; Italy, 1964; Mexico, 1957; Sweden, 1959; United Kingdom, 1957; United States, 1963). A comparison of practices in the United Kingdom, Canada, and Australia is G.S.A. WHEATCROFT (ed.), *Estate and Gift Taxation: A Comparative Study* (1965). The subject has been much discussed in Canada; particularly useful are vol. 3 of the *Report of the Royal Commission on Taxation* (1966); and *Studies of the Royal Commission on Taxation*, especially

no. 11, *Death Taxes*, by JOHN G. SMITH, D.B. FIELDS, and E.J. MOCKLER (1967); and no. 13, *Gift Tax*, by E.J. MOCKLER and D.B. FIELDS (1966).

Comprehensive studies include: C. LOWELL HARRISS, *Gift Taxation in the United States* (1940); GERALD R. JANTSCHLER, *Trusts and Estate Taxation* (1967); CARL S. SHOUP, *Federal Estate and Gift Taxes* (1966); HAROLD M. SOMERS, *Capital Gains, Death and Gift Taxation* (1965); UNITED STATES ADVISORY COMMISSION ON INTERGOVERNMENTAL RELATIONS, *Coordination of State and Federal Inheritance, Estate, and Gift Taxes* (1961); WILLIAM S. VICKREY, *Agenda for Progressive Taxation* (1947); and DOUGLAS A. KAHN, EARL M. COLSON, and GEORGE CRAVEN, *Federal Taxation of Estates, Gifts, and Trusts* (1970).

(H.M.S.)

Death Rites and Customs

Man is the only creature known to bury his dead. The fact is of fundamental significance. For the practice was not originally motivated by hygienic considerations but by ideas entertained by primitive peoples concerning human nature and destiny. This conclusion is clearly evident from the fact that the disposal of the dead from the earliest times was of a ritual kind. The Paleolithic peoples not only buried their dead but they provided them with food and other equipment, thereby implying a belief that the dead still needed such things in the grave. In such provision for the dead, Paleolithic man had been anticipated, inevitably in a cruder manner, by his predecessor, the so-called Neanderthal or Mousterian man, so that this very significant practice can be traced back to an even greater antiquity, possibly to about 50,000 BC.

Early practices

The ritual burial of the dead, which is thus attested from the very dawn of human culture and which has been practiced in most parts of the world, stems from an instinctive inability or refusal on the part of man to accept death as the definitive end of human life. Despite the horrifying evidence of the physical decomposition caused by death, the belief has persisted that something of the individual person continues to survive the experience of dying. In contrast, the idea of personal extinction through death is a sophisticated concept that was unknown until the 6th century BC, when it appeared in the metaphysical thought of Indian Buddhism; it did not find expression in the ancient Mediterranean world before its exposition by the Greek philosopher Epicurus (341–270 BC).

The belief that human beings survive death in some form has profoundly influenced the thoughts, emotions, and actions of mankind. The belief occurs in all religions, past and present, and decisively conditions their evaluations of man and his place in the universe. Mortuary rituals and funerary customs reflect these evaluations; they represent also the practical measures taken to assist the dead to achieve their destiny and sometimes to save the living from the dreaded molestation of those whom death had transformed into a different state of being.

RELEVANT CONCEPTS AND DOCTRINES

Life and death. The evidence of Paleolithic burials shows that already, in that remote age, various ideas were held about death and the state of the dead. The provision of food, ornaments, and tools in the graves implies a general belief that the dead continued to exist, with the same needs as in this life. Other customs, however, indicate the currency of a variety of notions about the postmortem existence, particularly about the potentialities and destiny of the dead. Thus, the presence of red ochre in some burials suggests the practice of contagious magic: the corpse had possibly been stained with the colour of blood in order to revitalize it. The fact that, in Paleolithic burials, the skeleton has often been found lying on its side, in a crouched position, has been interpreted by some prehistorians as evidence of belief in rebirth, in that the posture of the corpse imitated the position of the child in the womb. In some crouched burials, however, there is reason for suspecting a more sinister motive; for the limbs are sometimes so tightly flexed that the bodies must have been bound in that position before rigor mortis set in. Such treatment of the corpse was doubtless prompted by fear of the dead, for similar customs have been found among later peoples. Preventive action of this kind has a further significance, for it implies a belief that the dead might be malevolent and had power to harm the living.

That death was sometimes regarded as transforming those who experienced it into a state of being balefully different from that of those living in this world is evident in later mortuary rites and customs. Indeed, the proper performance of funerary rites was deemed essential by many peoples, to enable the dead to depart to the place and condition to which they properly belonged. Failure to expedite their departure could have dangerous consequences. Many ancient Mesopotamian divinatory texts reveal a belief that disease and other misfortunes could be caused by dead persons deprived of proper burial. The fate of the unburied dead finds expression in Greek and Roman literature. The idea that the dead had to cross some barrier that divided the land of the living from that of the dead also occurs in many religions: the Greeks and Romans believed that the dead were ferried across an infernal river, the Acheron or Styx, by a demonic boatman called Charon, for whose payment a coin was placed in the mouth of the deceased; in Zoroastrianism, the dead cross the Bridge of the Requiter (Činvato Paratu); bridges figure also in Muslim and Scandinavian eschatologies (speculations concerning the end of the world and the afterlife)—the Ṣirāṭ bridge and the bridge over the Gjöll River (Gjallarbrú)—and Christian folklore knew of a Brig o' Dread, or Brig o' Death.

It is significant that in few religions has death been regarded as a natural event. Generally, it has been viewed as resulting from the attack of some demonic power or death god: in Etruscan sepulchral art a fearsome being called Charun strikes the deathblow, and medieval Christian art depicted the skeletal figure of Death with a dart. In many mythologies death is represented as resulting from some primordial mischance. According to Christian theology, death entered the world through the original sin committed by Adam and Eve, the progenitors of mankind.

Human substance and nature. The conception of death in most religions is closely related to the particular view held about the constitution of human nature. Two major traditions of interpretation have provided the basic assumptions of religious eschatologies and have often found expression in mortuary rituals and funerary practice. The more primitive of these interpretations has been based on an integralistic evaluation of human nature. Thus, the individual person has been conceived as a psychophysical organism, of which both the material and nonmaterial constituents are all essential for a properly integrated personal existence. From such an evaluation it has followed that death is the fatal shattering of personal existence. Although some constituent element of the living person has been deemed to survive this disintegration, it has not been regarded as conserving the essential self or personality. The consequences of this estimate of human nature can be seen in the eschatologies of many religions. The ancient Mesopotamians, Hebrews, and Greeks, for example, thought that after death only a shadowy wraith descended to the realm of the dead, where it existed miserably in dust and darkness. Such a conception of man, in turn, has meant that, where the possibility of an effective afterlife has been envisaged, as in ancient Egyptian religion, Judaism, Zoroastrianism, Christianity, and Islām, the idea of a reconstitution or resurrection of the body has also been involved; for it has been deemed essential to restore the psychophysical complex of personality. In Egypt, most notably, provision was made for the reconstitution in an elaborate mortuary ritual, which included the mummification of the corpse to preserve it from disintegration.

Two dominant traditions concerning human nature

The alternative view of human nature may be termed dualistic. It conceives of the individual person as comprising an inner essential self or soul, which is nonmaterial, and a physical body. In many religions based on this view of human nature, the soul is regarded as being essentially immortal and as existing before the body was formed. Its incarnation in the body is interpreted as a penalty incurred for some primordial sin or error. At

death, the soul leaves the body, and its subsequent fate is determined by the manner in which it has fulfilled what the particular religion concerned has prescribed for the achievement of salvation. This view of human nature and destiny finds most notable expression in Hinduism and, in a subtly qualified sense, in Buddhism; it was also taught in such mystical cults and philosophies of the Greco-Roman world as Orphism (an ancient Greek mystical movement with a significant emphasis on death), Gnosticism (an early system of thought that viewed spirit as good and matter as evil), Hermeticism (a Hellenistic esoteric, occultic movement), and Manichaeism (a system of thought founded by Mani in ancient Iran).

Forms of survival. The conception of human nature held in any religion has, accordingly, determined the manner or mode in which postmortem survival has been envisaged. Where the body has been regarded as an essential constituent of personal existence, belief in a significant afterlife has inevitably entailed the idea of the reconstitution of the decomposed corpse and its resurrection to life. In turn, a dualistic conception of human nature, which regards the soul as intrinsically nonmaterial and immortal, envisages postmortem life in terms of the disembodied existence of the soul. This dualistic conception, in many religions, has also involved the idea of rebirth or reincarnation. In Hinduism, Buddhism, and Orphism this idea has inspired a cyclical view of the time process and produced esoteric explanations of how the soul becomes reborn into a physical body, whether human or animal.

The ultimate destiny of the dead. Belief in postmortem survival has been productive also of a variety of images concerning the destiny of the dead. This imagery is closely related to the conception of man that is held in each religion. Thus, the magical resuscitation of the dead in ancient Egypt was designed to enable them to live forever in their well-furnished tombs; according to Christian and Islāmic belief, God will ultimately raise the dead with their physical bodies and assess their merits for eternal bliss in heaven or everlasting torment in hell; the Buddhist concept of Nirvāṇa (Enlightenment) is achieved only when the individual has eradicated all desire for existence in the empirical world.

PATTERNS OF MYTH AND SYMBOL

Geography of the afterlife. Inhumation naturally prompted the idea that the dead lived beneath the ground. The mortuary cults of many peoples indicate that the dead were imagined as actually residing in their tombs and able to receive the offerings of food and drink made to them; *e.g.*, some graves in ancient Crete and Ugarit (Ras Shamra) were equipped with pottery conduits, from the surface, for libations. Often, however, the grave has been thought of as an entrance to a vast, subterranean abode of the dead. In some religions this underworld has been conceived as an immense pit or cavern, dark and grim (*e.g.*, the Mesopotamian *kur-nu-gi-a* ["land of no return"], the Hebrew Sheol, the Greek Hades, and the Scandinavian Hel). Sometimes it is ruled by an awful monarch, such as the Mesopotamian god Nergal or the Greek god Hades, or Pluto, or the Yama of Hindu and Buddhist eschatology. According to the view of man's nature and destiny held in a particular religion, this underworld may be a gloomy, joyless place where the shades of all the dead merely survive, or it may be pictured as a place of awful torments where the damned suffer for their misdeeds. In those religions in which the underworld has been conceived as a place of postmortem retribution, the idea of a separate abode of the blessed dead became necessary. Such an abode has various locations. In most religions it is imagined as being in the sky or in a divine realm beyond the sky (*e.g.*, in Christianity, Gnosticism, Hinduism, and Buddhism); sometimes it has been conceived as the "Isles of the Blessed" (*e.g.*, in Greek and Celtic mythology) or as a beautiful garden or paradise, such as the *al-firdaws* of Islām. Christian eschatology, which came to conceive of both an immediate judgment and a final judgment, developed the idea of a purgatory, where the dead expiated their venial sins in readiness for the final judgment. Although the dead suffered there in a disembodied state, because their bodies would not be resurrected until the last day, the purifying flames of purgatory were usually regarded as burning in a physical sense, as Dante's *Purgatorio* vividly shows. The idea of a postmortem purgatory had been adumbrated in the 1st and 2nd centuries BC in Jewish apocalyptic literature (*I Enoch* 22:9–13). The ten hells of Chinese Buddhist eschatology may be considered as purgatories, for in them the dead expiated their sins before being incarnated once more in this world.

The underworld and heaven

Means of approach to the afterworld. The idea that the dead had to make a journey to the otherworld, to which they belonged, finds expression in many religions. The oldest evidence occurs in the Egyptian Pyramid Texts (*c.* 2375–*c.* 2200 BC). The journey is conceived under various images. The dead pharaoh flies up to heaven to join the sun-god Re, in his solar boat, on his unceasing voyage across the sky, or he joins the circumpolar stars, known as the "Imperishable Ones," or he ascends a ladder to joins the gods in heaven. Later Egyptian funerary texts depict the way to the next world as beset by awful perils: fearsome monsters, lakes of fire, gates that cannot be passed except by the use of magical formulas, and a sinister ferryman whose evil intent must be thwarted by magic. The idea of crossing water en route to the otherworld, which first appears in Egyptian eschatology, occurs in the eschatological topography of other religions, as was noted above. Many mythologies describe journeys to the underworld; they invariably reflect the fear felt for the grim experience that was believed to await the dead. Ancient Mesopotamian literature records the visit of the goddess Ishtar to the realm of the dead, the way to which was barred by gates. At each gate the goddess was deprived of some article of her attire, so that she was naked when she finally came before Ereshkigal, the queen of the underworld. It is possible that this successive stripping of the celestial goddess was meant to symbolize the stripping away of the attributes of life that the dead experienced as they descended into the "land of no return." An 8th-century Japanese text, the *Koji-ki*, tells of the first contact with death experienced by the primordial pair, Izanagi and Izanami. When his wife died, Izanagi descended to Yomi, the underworld of darkness, to bring her back. His request was granted by the gods of Yomi, on condition that he did not look at her in the underworld. Impatiently he struck a light and was horrified to see her as a decomposed corpse. He fled in terror and disgust. Blocking the entrance to Yomi with a great rock, he then sought desperately to purify himself from the contagion of death.

The journey after death

Such myths doubtless reflect an instinctive feeling that death works an awful change in those who experience it. The dead cease to belong to the world of the living and become uncanny and dangerous: hence, their departure to the world of the dead must be expedited. To assist that grim journey, various aids have been provided. Thus, on some Egyptian coffins of the 11th dynasty, a plan of the "Two Ways" to the underworld was painted, and from the New Kingdom period (*c.* 1567–1085 BC), copies of the Book of the Dead, containing spells for dealing with perils encountered en route, were placed in the tombs. Orphic communities in southern Italy and Crete provided their dead with directions about the next world by inscribing them on gold laminae deposited in the graves. Advice about dying was given to medieval Christians in a book entitled *Ars moriendi* ("The Art of Dying") and to Tibetan Buddhists in the *Bardo Thödol* ("Book of the Dead"). Chinese Buddhists were informed in popular prints of what to expect as they passed after death through the ten hells to their next incarnation. More practical equipment for the journey to the next world was provided for the Greek and Roman dead: in addition to the money to pay Charon for their passage across the Styx, they were provided with honey cakes for Cerberus, the fearsome dog that guarded the entrance to Hades.

Forms of final determination. Those religions that have taught the possibility of a happy afterlife have also

devised forms of postmortem testing of merit for eternal bliss. Ancient Egypt has the distinction of conceiving of a judgment of the dead of an essentially moral kind. This conception finds graphic expression in the vignettes that illustrate the Book of the Dead. The heart of the deceased is represented as being weighed against the symbol of Maat (Truth) in the presence of Osiris, the god of the dead. A monster named Am-mut (Eater of the Dead) awaits an adverse verdict. The judgment of the dead as conceived in other religions (*e.g.*, Christianity, Islām, Zoroastrianism, Orphism) is basically a test of orthodoxy or ritual status, although moral qualities were included to varying degrees. The Last Judgment, as presented in Jewish apocalyptic literature, was essentially a vindication of Israel against its Gentile oppressors. Religions that held no promise of a significant afterlife (*e.g.*, those of ancient Mesopotamia and classical Greece) had no place for a judgment of the dead.

Judgment of the dead

DEATH AND FUNERARY RITES AND CUSTOMS

Before and at death. The process of dying and the moment of death have been regarded as occasions of the gravest crisis in many religions. The dying must be especially prepared for the awful experience. In China, for example, the head of a dying person was shaved, his body was washed and his nails pared, and he was placed in a sitting position to facilitate the exit of the soul. After the death, relatives and friends called the soul to return, possibly to make certain whether its departure from the body was definitive. Muslim custom decrees that the dying be placed facing the holy city of Mecca. In Catholic Christianity, great care is devoted to preparing for a "good death." The dying person makes his last confession to a priest and receives absolution; then he is anointed with consecrated oil: the rite is known as "anointing of the sick" (formerly called extreme unction). According to medieval Christian belief, the last moments of life were the most critical, for demons lurked about the deathbed ready to seize the unprepared soul as it emerged with the last breath.

By courtesy of the Bibliotheque Nationale, Paris



Muslim gentleman's funeral. Relatives, wearing mourning bands, look on as the body, wrapped in a seamless shroud, is entombed on its side facing Mecca. Illustration from *Maqāmāt* of al-Harīrī, painted by Yahya ibn Mahmūd al-Wāsitī, Baghdad, 1237. In the Bibliothèque Nationale, Paris (MS. Arabe 5847).

Modes of preparation of the corpse and attendant rites. After death, it has been the universal custom to prepare the corpse for final disposal. Generally, this preparation has included its washing and dressing in special garments and sometimes its public exposure. In some religions this preparation is accompanied by rites designed to protect the deceased from demonic attack; sometimes the purpose of the rites has been to guard the living from the contagion of death or the malice of the dead; for it has often been believed that the soul continues to remain about the body until burial or cremation. The most elaborate known preparation of the dead took place in ancient Egypt. Because the Egyptians believed that the body was essential for a proper afterlife, a complex process of ritual embalmment was established. This process was intended not only to preserve the corpse from physical disintegration but also to reanimate it. The rites were based upon the belief that, because the dead body of the god Osiris had been preserved from decomposition and raised to life again by the gods, the magical assimilation of a dead person to Osiris and the ritual enacting of what the gods had done would achieve a similar miracle of resurrection. One of the most significant of these ritual transactions was the "opening of the mouth," which was designed to restore to the mummified body its ability to see, breathe, and take nourishment.

Mummification in cruder forms has been practiced elsewhere (notably in Peru), but not with the same complex motives as in Egypt. The preparation of the corpse has also frequently included the placing on or in it of magical amulets; these were variously intended to protect or vitalize the corpse. Evidence found in tombs of the Shang dynasty (*c.* 1766–*c.* 1122 BC) suggests that the Chinese placed life-prolonging substances, such as jade, in the orifices of the corpse. Crosses or crucifixes are frequently placed upon the Christian dead, and sometimes in the Middle Ages the consecrated bread of the Eucharist (the Lord's Supper) was buried with the body. It has also been a Christian custom to furnish a dead priest with a chalice and paten, the instruments of his sacerdotal office.

Modes of disposal of the corpse and attendant rites. The form of the disposal of the dead most generally used throughout the world in both the past and present has been burial in the ground. The practice of inhumation (burial) started in the Paleolithic era, doubtless as the most natural and simplest way of disposal. Whether it was then prompted by any esoteric motive, such as the return to the womb of Mother Earth, as has been suggested, cannot be proved. Among some later peoples, who have believed that primordial man was formed out of earth, it may have been deemed appropriate that the dead should be buried—the idea found classical expression in the divine pronouncement to Adam, recorded in Genesis 3:19: "You are dust, and to dust you shall return." There is evidence that in ancient Crete the dead were believed to serve a great goddess, who was the source of fertility and life in the world above and who nourished and protected the dead in the earth beneath.

Inhumation

The mode of burial has varied greatly. Sometimes the body has been laid directly in the earth, with or without clothes and funerary equipment. It may be placed in either an extended or crouched position: the latter posture seems to have been more usual in prehistoric burials. Sometimes evidence of a traditional orientation of the corpse in the grave can be distinguished, which may relate to the direction in which the land of the dead was thought to lie. The use of coffins of various substances dates from the early 3rd millennium BC in Sumer and Egypt. Intended probably at first to protect and add dignity to the corpse, coffins became important adjuncts in the mortuary rituals of many religions. Their ritual use is most notable in ancient Egypt, where the mummies of important persons were often enclosed in several human-shaped coffins and then deposited in large, rectangular wooden coffins or stone sarcophagi. The interiors and exteriors of these coffins were used for the inscription of magical texts and symbols. Sarcophagi, elaborately carved with mythological scenes of mortuary significance, became fashionable among the wealthier classes of Greco-Roman society. Similar sarcophagi, carved with Christian scenes, came into use among Christians in the 4th and 5th centuries and afford rich iconographic evidence of the contemporary Christian attitude to death.

In the ancient Near East, the construction of stone tombs began in the 3rd millennium BC and inaugurated a tradition of funerary architecture that has produced such

diverse monuments as the pyramids of Egypt, the Tāj Mahal, and the mausoleum of Lenin in Red Square, Moscow. The tomb was originally intended to house and protect the dead. In Egypt it was furnished to meet the needs of its magically resuscitated inmate, sometimes even to the provision of toilet facilities. Among many peoples, the belief that the dead actually dwelt in their tombs has caused the tombs of certain holy persons to become shrines, which thousands visit to seek for miracles of healing or to earn religious merit; notable examples of such centres of pilgrimage are the tombs of St. Peter in Rome, of Muḥammad at Medinah, and, in ancient times, the tomb of Imhotep at Ṣaqqārah, in Egypt.

Ewing Krainin—Stockpile



Hindu-animist cremation in Bali, Indonesia. Bodies are hidden inside gilded papier-mâché cattle to confuse evil spirits.

The funeral

The disposal of the corpse has been, universally, a ritual occasion of varying degrees of complexity and religious concern. Basically, the funeral consists of conveying the deceased from his home to the place of burial or cremation. This act of transportation has generally been made into a procession of mourners who lament the deceased, and it has often afforded an opportunity of advertising his wealth, status, or achievements. Many depictions of ancient Egyptian funerary processions graphically portray the basic pattern: the embalmed body of the deceased is borne on an ornate sledge, on which sit two mourning women. A priest precedes the bier, pouring libations and burning incense. In the cortege are groups of male mourners and lamenting women, and servants carry the funerary furniture, which indicates the wealth of the dead man. Ancient Roman funerary processions were notable for the parade of ancestors' death masks. In Islāmic countries, friends carry the corpse on an open bier, generally followed by women relatives, lamenting with dishevelled hair, and hired mourners. After a service in the mosque, the body is interred with its right side toward Mecca. In Hinduism the funeral procession is made to the place of cremation. It is preceded by a man carrying a firebrand kindled at the domestic hearth; a goat is sometimes sacrificed en route, and the mourners circumambulate the corpse, which is carried on a bier. Cremation is a ritual act, governed by careful prescriptions. The widow crouches by the pyre, on which in ancient times she sometimes died. After cremation, the remains are gathered and often deposited in sacred rivers.

Christian funerary ritual reached its fullest development in medieval Catholicism and was closely related to doctrinal belief, especially that concerning purgatory. Hence, the funerary ceremonies were invested with a sombre character that found visible expression in the use of black vestments and candles of unbleached wax and the solemn tolling of the church bell. The rites consisted of five distinctive episodes. The corpse was carried to the church in a doleful cortege of clergy and mourners, with the intoning of psalms and the purificatory use of incense. The coffin was deposited in the church, covered with a black pall, and the Office of the Dead was recited or sung, with the constant repetition of the petition: "Eternal rest grant unto him, O Lord: and let perpetual light shine upon him." Next, requiem mass was said or sung, with the sacrifice especially offered for the repose of the soul of the deceased. After the mass followed the "Absolution" of the dead person, in which the coffin was solemnly perfumed with incense and sprinkled with holy water. The corpse was then carried to consecrated ground and buried, while appropriate prayers were recited by the officiating priest. Changes in these rites, including the use of white vestments and the recitation of prayers emphasizing the notions of hope and joy, began to be introduced into the Catholic liturgy only following the second Vatican Council (1962–65).

In some societies the burial of the dead has been accompanied by human sacrifice, with the intention either to propitiate the spirit of the deceased or to provide him with companions or servants in the next world. A classic instance of such propitiatory sacrifice occurs in Homer's *Iliad* (xxiii:175–177): 12 young Trojans were slaughtered and burnt on the funeral pyre of the Greek hero Patroclus. The royal graves excavated at the Sumerian city of Ur, dating *c.* 2700 BC, revealed that retinues of servants and soldiers had been buried with their royal masters. Evidence of a similar Chinese practice has been found in Shang dynasty graves (12th to 11th centuries BC), at An-yang. In ancient Egypt, models of servants, placed in tombs, were designed to be magically animated to serve their masters in the afterlife. A particular type of these models, known as an *ushabti* ("answerer"), was inscribed with chapter VI of the Book of the Dead, commanding it to answer for the deceased owner if he were required to do service in the next world.

Dismemberment of the body

The custom has also existed among some peoples of dismembering the body for burial or subsequently disinterring the bones for storage in some form. There is Paleolithic evidence of a cult of skulls, which suggests that the rest of the body was not ritually buried. The Egyptians removed the viscera, which were preserved separately in four canopic jars. The Romans observed the curious rite of the *os resectum:* after cremation a severed finger joint was buried, probably as a symbol of an earlier custom of inhumation. In medieval Europe, the heart and sometimes the intestines of important persons were buried in separate places: *e.g.*, the body of William the Conqueror was buried in St. Étienne at Caen, but his heart was left to Rouen Cathedral and his entrails for interment in the church of Chalus. To be noted also is the Zoroastrian and Parsi custom of exposing corpses on dakhmas ("towers of silence") to be devoured by birds of prey, thus to avoid polluting earth or air by burial or cremation.

The alternative use of inhumation or cremation for the disposal of the corpse cannot be interpreted as generally denoting a difference of view about the fate of the dead. In India, cremation was indeed connected with the fire god Agni, but cremation does not necessarily indicate that the soul was thus freed to ascend to the sky. Burial has been the more general practice, whether the abode of the dead be located under the earth or in the heavens.

Post-funerary rites and customs. Funerary rites do not usually terminate with the disposal of the corpse either by burial or cremation. Post-funerary ceremonies and customs may continue for varying periods; they have generally had two not necessarily mutually exclusive motives: to mourn the dead or to purify the mourners. The mourning of the dead, especially by near relatives, has taken many forms. The wearing of old or colourless dress, either black or white, the shaving of the hair or letting it grow long and unkempt, and abstention from

amusements have all been common practice. The meaning of such action seems evident: grief felt for the loss of a dear relative or friend naturally expresses itself in forms of self-denial. But the purpose may sometimes have been intended to divert the ill humour of the dead from those who still enjoyed life in this world.

The purification of the mourners has been the other powerful motive operative in much post-funerary action. Death being regarded as baleful, all who came in contact with the corpse were contaminated thereby. Consequently, among many peoples, various forms of purification have been prescribed, chiefly by bathing and fumigation. Parsis are especially intent also on cleansing the room in which the death occurred and all articles that had contact with the dead body.

In some post-funerary rituals, dancing and athletic contests have had a place. The dancing seems to have been inspired by various but generally obscure motives. There is some evidence that Egyptian mortuary dances were intended to generate a vitalizing potency that would benefit the dead. Dances among other peoples suggest the purpose of warding off the (evil) spirits of the dead. Funeral games would seem to have been, in essence, prophylactic assertions of vitalizing energy in the presence of death. It has been suggested that the funeral games of the Etruscans, which involved the shedding of blood, had also a sacrificial significance.

Another widespread funerary custom has been the funeral banquet, which might be held in the presence of the corpse before burial or in the tomb-chapel (in ancient Rome) or on the return of the mourners to the home of the deceased. The purpose behind these meals is not clear, but they seem originally to have been of a ritual character. Two curious instances of mortuary eating may be mentioned in this connection. There was an old Welsh custom of "sin eating": food and drink were handed across the corpse to a man who undertook thereby to ingest the sins of the deceased. In Bavaria, *Leichennudeln*, or "corpse cakes," were placed upon the dead body before baking. By consuming these cakes, the kinsmen were supposed to absorb the virtues and abilities of their deceased relatives.

Chastisement of the Tomb

A remarkable post-funerary custom has been observed in Islām; it is known as the Chastisement of the Tomb. It is believed that, on the night following the burial, two angels, Munkar and Nakīr, enter the tomb. They question the deceased about his faith. If his answers are correct, the angels open a door in the side of the tomb for him to pass to repose in paradise. If the deceased fails his grisly interrogation, he is terribly beaten by the angels, and his torment continues until the end of the world and the final judgment. In preparation for this awful examination the roof of the tomb is constructed to enable the deceased to sit up; and, immediately after burial, a man known as a *fiqī* (*or faqih*) is employed to instruct the dead in the right answers.

CULTS AND MEMORIALS OF THE DEAD

Commemorative rites and services. The attitude of the living toward the dead has also been conditioned by the particular belief held about the human nature and destiny. Where death is regarded as the virtual extinction of the personality, the dead should logically have no more importance beyond that which their memory might stir in those who knew them. Even in the negative eschatologies of ancient Mesopotamia and Greece, however, the dead were thought of as still existent and capable of malevolent action if food offerings were not made to them. In those religions that have envisaged a more positive afterlife, the tendance of the dead has been developed in varying ways. In Egypt, it led to the building and endowment of mortuary temples or chapels, in which portrait images preserved the memory of the dead and offerings of food and drink were regularly made. In China, an elaborate ancestor cult flourished. The ancestral shrine contained tablets, inscribed with the names of ancestors, which were revered and before which offerings were made. The number of tablets displayed in the shrine was determined by the social status of the family. When the tablet of a newly deceased member was added to the collection, the oldest tablet was deposited in a chest containing still older ones: offerings to the remoter ancestors were made collectively at longer intervals. In India, three generations of deceased ancestors are venerated at the monthly *śrāddha* festival, at which mortuary offerings were made.

The Christian cult of the dead found early expression in the catacombs, where mural paintings and inscriptions record the names of those buried there and the hopes of eternal peace and felicity that inspired them. Special chapels were made where the bodies of martyrs were entombed, and the anniversaries of their martyrdoms were commemorated by the celebration of the Eucharist (the Lord's Supper). The development of cults of martyrs and other saints in the medieval church centred on the veneration of their relics, which were often divided among several churches. The introduction of the doctrine of purgatory profoundly affected the postmortem care devoted to the ordinary dead. It was believed that the offering of the sacrifice of the mass could alleviate the sufferings of departed souls in purgatory. Consequently, the celebration of masses for the dead proliferated, and wealthy Christians endowed monasteries or chantry chapels where masses were said regularly for the repose of their own souls or those of their relatives. Prayers for the well-being of the dead have an important place in Mahāyāna Buddhism, and so-called "masses for the dead" were celebrated by Chinese Buddhists, influenced originally perhaps by the practice of the Nestorian Christians, who entered China in the 7th century AD.

Periodic commemorations

In many religions, in addition to private cults of the dead, periodic commemorations of the dead have been kept. The oldest of the Hindu sacred texts, the Ṛgveda, records the practice of the ancient Aryan invaders of India. The sacred beverage called *soma* was set out on "the sacred grass," and the ancestors were invited to ascend from their subterranean abode to partake of it and to bless their pious descendants. A similar ceremony, called the Anthesteria, was held in ancient Athens. On the day concerned, the souls of the dead (*kēres*) were believed to leave their tombs and revisit their former homes, where food was prepared for them. At sundown they were solemnly dismissed to the underworld with the formula: "out, *kēres*, the Anthesteria is ended." Buddhist China kept a Feast of Wandering Souls each year, designed to help unfortunate souls suffering in the next world. The Christian All Souls Day, on November 2, which follows directly after All Saints Day, commemorates all the ordinary dead: requiem masses are celebrated for their repose, and in many Catholic countries relatives visit the graves and place lighted candles on them. After World War I the public commemoration of the fallen was instituted on November 11, the day of the armistice in 1918, in many of the countries concerned: the memory of the dead was solemnly recalled in a two-minute silence during the ceremony. The body of an unknown soldier, killed in the fighting, was also buried in the capital cities of many countries and has become the accepted focus of national reverence and devotion.

Cult of the dead. Among many peoples it has been the custom to preserve the memory of the dead by images of them placed upon their graves or tombs, usually with some accompanying inscription recording their names and often their achievements. This sepulchral iconography began in Egypt, the portrait statue of King Djoser (second king of the 3rd dynasty [*c.* 2686–*c.* 2613 BC]), found in the *serdab* (worship chamber; from the Arabic word for cellar) of the Step Pyramid being the oldest known example. The Egyptian images, however, had a magical purpose: they not only recorded the features of the deceased but also provided a locus for his *ka*, the mysterious entity that constituted an essential element of the personality. The sculptured gravestones of classical Athens deserve special notice, for they are among the noblest products of funerary art. They are expressive of a restrained grief for those who had departed to the virtual extinction of Hades. The deceased are often shown performing some familiar act for the

Images of the dead

last time. The inscriptions are very brief and usually record only the name and parentage; sometimes the word farewell is added. Etruscan mortuary art is characterized by the effigy of the deceased, sometimes with his wife, represented as reclining on the cover of the funerary casket. These images are obviously careful portraits, but whether they had some magical use as substitute bodies or are only commemorative is unknown. Roman funerary images seem to have been essentially commemorative, as were those of Palmyra.

Christianity has provided the richest legacy of funerary monuments. In the catacomb art of the 4th and 5th centuries, the deceased was sometimes depicted on the plaster covering of the niche in which his body was laid. From the early Middle Ages onward, the more affluent dead were represented in sculptured effigy or engraved in outline on stone or brass. In this tomb iconography, they are shown in a variety of postures: lying, kneeling, seated, standing, and sometimes on horseback. They are generally presented in the dress appropriate to their office or social standing: kings wear crowns, knights their armour; bishops are in copes and mitres and ladies in the fashionable attire of the day. This iconography is patently commemorative of the appearance in life, the achievements, and the status of the persons concerned. In the later Middle Ages, however, there was a remarkable innovation in this funerary art, which was designed to emphasize the horror and degradation of death. In what are known as memento mori tombs, below the effigies of the deceased as they were in life, there were placed effigies of their naked decaying corpses or skeletons. Such tomb sculpture reflected a contemporary obsession with the corruption of death.

PSYCHOLOGICAL AND SOCIOLOGICAL ASPECTS OF DEATH

The Paleolithic burials reveal that the pattern of man's reaction to the fact and phenomena of death has been set from the dawn of culture. Unlike the other animals, man has been unable to ignore the mysterious cessation of activity and lapse of consciousness that cause his body to decay and befall members of his own kind. Death has, accordingly, constituted a problem for man, and he has felt impelled to take special action to cope with it. The pattern of his reaction has been twofold: confronted with the deaths of his companions, he has recognized an obligation to attend to their needs as he has conceived them, believing that they continued to exist in some form, either in the grave or in an underworld to which the grave gave access. But man's concern with death has not been confined to his tendance of the dead; for in the deaths of his fellows he has seen a presage of his own demise. This anticipation on the part of the living of the experience of dying has been a factor of immense psychological and social import. It is essentially a human characteristic; it stems from a consciousness of time, of which the immense cultural significance is only now beginning to be properly evaluated.

Awareness of time and anticipation of death

Awareness of time in its three categories of past, present, and future has decisively contributed to man's success in the struggle for existence. For it has enabled him to draw upon past experience in the present to anticipate future needs. Thus, from the making of the first stone tools to the complex structure of his modern technological civilization, man has sought by planning to render himself economically secure and to improve the standard of his living. But his time consciousness, which has made this immense achievement possible, is an ambivalent endowment. For, although it has enabled man to win economic security, it has also made him acutely aware of his own mortality and the inevitability of his own demise. Hence, his anticipation of death presents him with a profound emotional challenge, unknown to other species. The repercussions of this challenge can be traced in almost every aspect of his social and cultural life; but it is in his religions that man's reaction to death finds its most significant expression. All religion is concerned with postmortem security—with linking mortal man to an eternal realm—whether it be achieved by ritual magic, divine assistance, or mystic enlightenment.

MODERN NOTIONS OF DEATH

Continuation of traditional responses. Religious rites and customs continue to be practiced, because of conservatism, long after the ideas and beliefs that originally inspired them may be forgotten or abandoned. This is particularly true with regard to rites and customs pertaining to death. It is difficult to assess to what extent in the more sophisticated societies of the modern world the traditional eschatologies are still effectively held. Although a general skepticism obviously manifests itself toward the medieval imagery of death and judgment, of purgatory, heaven, and hell, modern modes of thinking have not lessened the mystery of death and its impact on the emotions. Indeed, in modern society, where expectation of life has been prolonged and standards of living raised, the negation of death is probably felt more keenly and also more hopelessly than in any other age.

Avowed secular inattention and unconcern. The reaction to death most apparent today among those having no effective religious faith is that of seeking to treat it as a disagreeable happening that must be dealt with as quickly and unobtrusively as possible. Funerals are no longer elaborately organized, mourning attire is rarely worn, and graveyards are landscaped, thus discreetly removing the earlier memorials of death. The increasing use of cremation facilitates this disposition to reduce the social intrusion of death and banish the traditional grave as a reminder of human mortality.

Rites and customs among secular materialists. It is significant, however, that, even where secularist principles are consciously professed, the dead are rarely disposed of without some semblance of ceremony. A deeply rooted feeling prompts most people to treat a dead human body with a respect that is not felt for a dead animal. It is significant that Communists make pilgrimages to the graves of Lenin and Marx; and, in the modern State of Israel, great effort is being made to record in the shrine of Yad va-Shem the names of those who died in the persecution of the Jews in Germany during the Nazi regime of Adolf Hitler in the 1930s and '40s and, if possible, to bring their ashes there. In America, morticians strive to preserve the features of the dead as did the embalmers of ancient Egypt, though for somewhat different motives. Finally, as further evidence of modern preoccupation with death, it may be noted that, in Western society, Spiritualism witnesses to a widespread desire to have communication with the dead, and recently, in England, there has even been a recrudescence of necromancy.

BIBLIOGRAPHY. E. BENDANN, *Death Customs: An Analytical Study of Burial Rites* (1930), a useful account of relevant ethnological material available at the time; S.G.F. BRANDON, *Man and His Destiny in the Great Religions* (1962), extensive bibliographies and documentation; *The Judgment of the Dead* (1967), the only comprehensive study of the subject in English, with extensive bibliographies and documentation; *Man and God in Art and Ritual* (1972), a profusely illustrated study that deals with mortuary rituals, conceptions of tomb, and funerary iconography, with bibliographies and documentation; and "The Personification of Death in Some Ancient Religions," *Bull. John Rylands Library*, 43:317–385 (1961); J. MARINGER, *Vorgeschichtliche Religion* (1956; Eng. trans., *The Gods of Prehistoric Man*, 1960), deals with Paleolithic and Neolithic burial practices; E.A.W. BUDGE, *The Mummy*, 2nd ed. rev. and enl. (1925), a valuable handbook on Egyptian funerary archaeology; G.E. SMITH and W.R. DAWSON, *Egyptian Mummies* (1924), fullest study of subject; J. ZANDEE, *Death As an Enemy, According to Ancient Egyptian Conceptions* (1960), deals also with Coptic evidence; A.R. JOHNSON, *The Vitality of the Individual in the Thought of Ancient Israel* (1949), an analysis of Hebrew ideas about human nature and destiny; J. JEREMIAS, *Heiligengräber in Jesu Umwelt (Mt. 23, 29; Lk. 11, 47)* (1958), a valuable account of Jewish mortuary beliefs; A. PARROT, *Le "Refrigerium" dans l'au-delà* (1937), a comparative study of ancient ideas about the condition of the dead; E. ROHDE, *Psyche: Seelenkult und Unsterblichkeitsglaube der Griechen*, 8th ed. (1921; Eng. trans., *Psyche: The Cult of Souls and Belief in Immortality Among the Greeks*, 1925, reprinted 1966), the fundamental book on the subject; F.V.A. CUMONT, *After Life in Roman Paganism* (1922); J.M.C. TOYNBEE, *Death and Burial in the Roman World* (1971); R. EKLUND, *Life Between*

Death and Resurrection According to Islam (1941); J.D.C. PAVRY, *The Zoroastrian Doctrine of a Future Life: From Death to the Individual Judgment*, 2nd ed. (1929); J.J. MODI, *The Religious Ceremonies and Customs of the Parsees*, 2nd ed. (1937); J.A. DUBOIS, *Hindu Manners, Customs and Ceremonies*, 3rd ed. (1906), an old book, but invaluable for its descriptions; M. GRANET, *La Civilisation Chinoise* (1929; Eng. trans., *Chinese Civilization*, 1930, reprinted 1957); *La Religion des Chinois*, 2nd ed. (1951); L. HERTLING and E. KIRSCHBAUM, *Die römischen Katakomben und ihre Martyrer*, 2nd ed. (1950; Eng. trans., *The Roman Catacombs and Their Martyrs*, rev. ed., 1960); W.K.L. CLARKE (ed.), *Liturgy and Worship: A Companion to the Prayer Books of the Anglican Communion* (1932); T.S.R. BOASE, "King Death: Morality, Judgment and Remembrance," in J. EVANS (ed.), *The Flowering of the Middle Ages* (1966); A.K. FORTESCUE, *The Ceremonies of the Roman Rite Described*, 6th ed. rev. by J.B. O'CONNELL (1937); E. PANOFSKY, *Tomb Sculpture* (1964), an illustrated survey from ancient Egypt to the Renaissance.

(S.G.F.B.)

Death Valley

One of the most remarkable natural features of North America, Death Valley is not only the driest and hottest area of that continent but also contains the lowest point in the Western Hemisphere. In some years there is no measurable rainfall, and the summer daytime temperature often exceeds 120° F (49° C). Its extreme environment is of considerable scientific interest.

Death Valley lies in southeastern California near the Nevada border, is bounded on the west by the Panamint Range, and on the east by the Black, Funeral, and Grapevine mountains of the Amargosa Range. The valley extends roughly north–south, with a length of nearly 140 miles (225 kilometres) and a width of between five and 15 miles (eight and 24 kilometres). Geologically it is an elongated structural depression forming part of the southwestern portion of the Basin and Range Province—a physiographic subdivision of the western United States. It is similar in area and overall form to other structural basins of the region but unique in its depth. Parts of the great salt pan that forms the floor of part of the valley form the lowest dry land areas of the Western Hemisphere. The lowest point is 282 feet (86 metres) below sea level and is found 4.75 miles (7.64 kilometres) west of Badwater, California.

The physical environment. For several decades after its christening in 1849 by a hapless party of emigrants who endured intense suffering while crossing it, Death Valley was little known except to the Indians of the area and to prospectors searching the surrounding mountains.

History of geological study. The first scientific notice of the valley seems to have been a brief note published in 1868 by the California state geologist. Mining activities, especially after the borax discoveries of the 1880s, drew the attention of other geologists. Reconnaissance reports began to appear around the turn of the century, but little detailed geological work was done until after the establishment of Death Valley National Monument in 1933. Since that time, under the auspices of the National Park Service, the U.S. Geological Survey, and the California Division of Mines and Geology, the central and southern parts of the valley and the adjacent mountains have been studied in detail. Information on the northern part of the valley, especially the area outside the National Monument, is still relatively scanty.

Establishment of National Monument

Geology. The mountain walls of Death Valley are composed largely of marine sedimentary rocks, deposited over a span of time from Late Precambrian, more than 600,000,000 years ago, to Triassic, about 200,000,000 years ago, in a great trough, the Cordilleran Geosyncline, which occupied much of what is now western North America. In places the sedimentary rocks can be seen to rest upon a base of much older, metamorphic (heat-altered) and granitic rocks. In Middle Mesozoic time (180,000,000 years ago) compressive forces deformed and uplifted the rocks that had accumulated in the geosyncline, converting much of the former seaway into a mountain belt.

In Late Mesozoic or Early Cenozoic time, about 100,000,000 years ago, the earth's crust in the region of Death Valley was broken by a series of fractures, some of them forming flat-lying surfaces along which great plates of rock moved over one another, sometimes for many miles. This process of thrust faulting, accompanied by folding, seems to have continued for many millions of years. The complexity of the results is suggested by the use of the term chaos to describe the structure of some of the thrust plates, which form a jumble of blocks of rock of various ages, with dimensions ranging from a few feet to thousands of feet. In places the fault planes have been exposed, either by erosion or by later sliding of the thrust plates from their tilted surfaces. Some of these form domed hills, referred to as turtle backs. Many questions about the mechanism of thrust faults in general, and about their part in the history of Death Valley, remain unanswered.

Faulting and volcanic activity

Magma from deep within the earth followed some of the fractures toward the surface. In two areas in the Panamint Range the magma domed up the overlying rocks and then cooled and crystallized at depth to form masses of granite. Elsewhere it broke through the surface to form volcanoes. Volcanic activity continued into recent geological times in the northern end of the valley, where explosive eruptions blasted out the Ubehebe Craters and scattered volcanic ash that still blankets the land surface.

Another type of fault activity, block faulting, in which the movement is dominantly vertical, began to form the contemporary Death Valley in Middle Tertiary time (40,000,000 years ago). Remains of early Oligocene mammals are preserved in sediments that accumulated in one of the basins thus formed, now exposed in Titus Canyon on the east side of Death Valley. Sinking of a series of crustal blocks to form the great trough of Death Valley and uplift of other blocks to form the adjacent mountain ranges progressed gradually through the rest of the Cenozoic Era. As the valley sank, sediments eroded from the surrounding hills accumulated, along with the products of the intermittent volcanic activity. Later fault movements tilted and exposed some of the older Cenozoic sediments along the east side of the valley; their subsequent erosion has formed the spectacular badlands near Furnace Creek. In the central part of the valley the bedrock floor is buried beneath as much as 9,000 feet (2,750 metres) of sediment.

Right lateral displacement, a third type of fault activity, has moved the west side of the valley to the northwest. The magnitude of this movement is still a matter of controversy: it may be as much as fifty miles.

Tilting and sinking of the valley floor have continued to the present time. Recent faults have affected alluvial fans in many parts of the valley perimeter. The shoreline of a lake that existed 2,000 years ago is now 20 feet (6 metres) lower on the east side than on the west, while changes in surveyed elevations suggest that parts of the valley floor have subsided several inches during the present century.

Climate and hydrology. The floor of Death Valley is noted for its extremes of temperature and aridity. A record North American high shaded air temperature of 134° F (57° C) was recorded in 1913, and summer temperatures often exceed 120° F (49° C). Ground temperatures as high as 190° F (88° C) have been reported.

Winter minimum temperatures rarely fall to the freezing point. For a 50 year period in the 20th century the average annual rainfall at Furnace Creek was only 1.66 inches (4.22 centimetres); the maximum annual rainfall was 4.5 inches (11.4 centimetres), and two years passed with no measurable rainfall. High temperatures, low humidity, and fairly constant air movement contribute to an exceptionally high evaporation rate, averaging about 150 inches (381 centimetres) per year.

Most of the surface water in Death Valley is in the saline ponds and marshes around the salt pan, but freshwater springs and seeps are less rare than in adjacent desert areas. The Amargosa River brings some water into the southern end of the valley from desert areas to the east, but most of its flow is underground. Salt Creek, draining the northern arm of the valley, also has only short stretches of perennial surface flow. Water emerging

at major springs in Furnace Creek Wash may follow fault zones from a source far to the east.

At times in the past much more water reached Death Valley. During the Wisconsin Glacial Age of the Pleistocene Epoch, perhaps about 50,000 years ago, a body of water (Lake Manly) filled the valley to a depth of as much as 600 feet (180 metres). Wave-cut terraces on Shoreline Butte near the south end of the valley record its fluctuations. The source of this water is uncertain: drainage from much of the eastern slopes of the Sierra Nevada may have flowed into Death Valley by way of Owens and Panamint valleys, but some evidence points to the Mojave Desert area as a more probable source. More recently, perhaps 2,000 to 5,000 years ago, a shallow lake occupied the floor of the valley, its evaporation producing the present salt pan.

Plant and animal life. Lack of water makes Death Valley a desert, but it is by no means devoid of life. Plant life above the microscopic level is absent from the salt pan, but salt-tolerant pickleweed, saltgrass, and rushes grow around the springs and marshes at its edges. Introduced tamarisks provide welcome shade around some of the springs and in the inhabited areas at Furnace Creek. Mesquite flourishes where less saline water is available. Creosote bush dominates the gravel fan surfaces around most of the valley, giving way to desert holly at the lowest elevations. Cactus is rare in the lowest part of the valley but abundant on the fans farther north. Spring rains bring out a great variety of desert wildflowers.

Wildlife in the valley

Animal life is varied, although nocturnal habits conceal many of the animals from visitors to the valley. Rabbits and several types of rodents, including antelope squirrels, kangaroo rats, and desert wood rats, are present and are preyed upon by coyotes, kit foxes, and bobcats. The largest native mammal in the area, and perhaps the best studied member of the fauna, is the desert bighorn sheep. Small herds of sheep are most commonly found in the mountains surrounding Death Valley but at least occasionally visit the valley floor. Wild burros, descendants of animals lost or abandoned by prospectors and miners, have become so numerous as to threaten, through overgrazing, the natural vegetation and other animals dependent upon it.

Visitors may gain the impression that the only birds present are the raucous and numerous ravens, but the first biological survey of the valley, in the 1890s, reported 78 species of birds, and nearly three times that number are now known to inhabit or visit the area. Lizards are numerous, snakes comparatively rare. Even native fish are to be found in Death Valley—several forms of desert pupfish of the genus *Cyprinodon* live in Salt Creek and other permanent bodies of water. Their ancestors apparently entered Death Valley during one of the past periods of high rainfall and permanent streamflow into the valley.

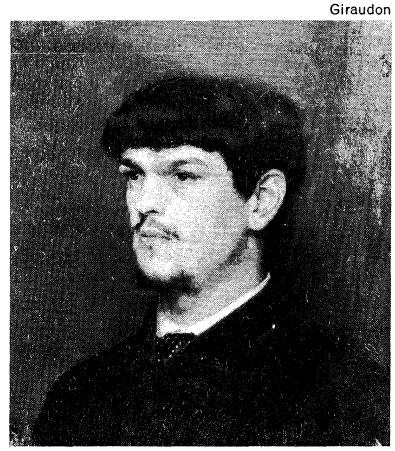
BIBLIOGRAPHY. C.B. HUNT and D.R. MABEY, "Stratigraphy and Structure, Death Valley, California," *Prof. Pap. U.S. Geol. Surv. 494-A* (1966), C.B. HUNT, *et al.*, "Hydrologic Basin, Death Valley, California," *ibid. 494-B* (1966), together, the most detailed geologic study available of central and southern Death Valley; C.B. HUNT, "Plant Ecology of Death Valley, California," *ibid.* 509 (1966), a detailed survey of plants of the lower part of Death Valley, emphasizing distribution in relation to geology and hydrology; R.E. and F.B. WELLES, *The Bighorn of Death Valley*, National Park Service, Fauna Series No. 6 (1961), a study of bighorn ecology; W.A. CHALFANT, *Death Valley: The Facts*, 3rd ed. (1936), dated, but still interesting; E.C. JAEGER, *A Naturalist's Death Valley* (1968), a brief survey of plants and animals of the valley.

(Jo.E.M.)

Debussy, Claude

The works of Achille Claude Debussy have been a seminal force in the music of the 20th century. He developed a highly original system of harmony and musical structure that expressed in many respects the ideals to which the Impressionist and Symbolist painters and writers of his time aspired. As a man he was enigmatical, saturnine, and unsociable. He shunned the limelight and resented having to make his living by writing music. Debussy hated everything grandiloquent and rejected all the trappings, pleasant and unpleasant, that eventually came to him through fame. Born of humble parents on August 22, 1862, at Saint-Germain-en-Laye, near Paris, Claude Debussy showed a gift as a pianist by the age of nine. He was encouraged by Madame Mauté de Fleurville, who was associated with the Polish composer Frédéric Chopin, and in 1873 he entered the Paris Conservatoire, where he studied the piano and composition, eventually winning in 1884 the Grand Prix de Rome with his cantata *L'Enfant prodigue* (*The Prodigal Child*).

Giraudon



Debussy, painting by Marcel-André Baschet, 1884. In the Musée de Versailles.

Debussy's youthful years were spent in circumstances of great turbulence. He was almost overwhelmed by situations of great extremes, both material and emotional. While living with his parents in a poverty-stricken suburb of Paris, he unexpectedly came under the patronage of a Russian millionairess, Nadezhda Filaretovna von Meck, who engaged him to play duets with her and her children. He travelled with her to her palatial residences throughout Europe during the long summer vacations at the Conservatoire. In Paris during this time he fell in love with a singer, Blanche Vasnier, the beautiful young wife of an architect; she inspired many of his early works. It is clear that he was torn by influences from many directions; these stormy years, however, contributed to the sensitivity of his early style.

Early style

This early style is well illustrated in one of Debussy's best known compositions, *Clair de lune* (*Moonlight*), from his *Suite bergamasque*. The title refers to a folk song that was the conventional accompaniment of scenes of the love-sick Pierrot in the French pantomime; and indeed the many Pierrot-like associations in Debussy's music, notably in the orchestral work *Images* (1912) and the *Sonata for Cello and Piano* (1915; originally titled *Pierrot fâché avec la lune*), show his connections with the circus spirit that also appeared in works by other composers, notably the ballet *Petrushka* (1911) by Igor Stravinsky and *Pierrot Lunaire* by Arnold Schoenberg.

As a holder of the Grand Prix de Rome, Debussy was given a three-year stay at the Villa Medici, in Rome, where, under what were supposed to be ideal conditions, he was to pursue his creative work. Most composers who were granted this state scholarship, however, found life in this magnificent Renaissance palace irksome and longed to return to simpler and more familiar surroundings. Debussy himself eventually fled from the Villa Medici after two years and returned to Blanche Vasnier in Paris. Several other women, some of doubtful reputation, were also associated with him in his early years. At this time Debussy lived a life of extreme indulgence. Once one of his mistresses, Gabrielle ("Gaby") Dupont, threatened suicide. His first wife, Rosalie ("Lily") Texier, a dressmaker, whom he married in 1899, did in fact shoot herself, though not fatally; and, as is sometimes the case with artists of passionate intensity, Debussy himself was haunted by thoughts of suicide.

The main musical influence in Debussy's work was the

work of Richard Wagner and the Russian composers Aleksandr Borodin and Modeste Moussorgsky. Wagner fulfilled the sensuous ambitions not only of composers but also of the Symbolist poets and the Impressionist painters. Wagner's conception of *Gesamtkunstwerk* ("total art work") encouraged artists to refine upon their emotional responses and to exteriorize their hidden dream states, often in a shadowy, incomplete form; hence the more tenuous nature of the work of Wagner's French disciples. It was in this spirit that Debussy wrote the symphonic poem *Prélude à l'après-midi d'un faune*, (1894; *Prelude to the Afternoon of a Faun*). Other early works by Debussy show his affinity with the English Pre-Raphaelite painters; the most notable of these works is *La Damoiselle élue* (1888), based on "The Blessed Damozel" (1850), a poem by the English poet and painter Dante Gabriel Rossetti. In the course of his career, however, which covered only 25 years, Debussy was constantly breaking new ground. Explorations, he maintained, were the essence of music; they were his musical bread and wine. His single completed opera, *Pelléas et Mélisande* (first performed 1902) demonstrates how the Wagnerian technique could be adapted to portray subjects like the dreamy nightmarish figures of this opera who were doomed to self-destruction. Debussy and his librettist, Maurice Maeterlinck, declared that they were haunted in this work by the terrifying nightmare tale of Edgar Allan Poe, "The Fall of the House of Usher." The style of *Pelléas* was to be replaced by a bolder, more highly coloured manner. In his seascape *La Mer* (1905; *The Sea*) he was inspired by the ideas of the English painter J.M.W. Turner and the French painter Claude Monet. In his work, as in his personal life, he was anxious to gather experience from every region that the imaginative mind could explore.

Influence of Wagner on Debussy

In 1905 Debussy's illegitimate daughter, Claude-Emma, was born. He had divorced Lily Texier in 1904 and subsequently married his daughter's mother, Emma Bardac. Repelled by the gossip and scandal arising from this situation, he sought refuge for a time at Eastbourne, on the south coast of England. For his daughter, nicknamed Chouchou, he wrote the piano suite *Children's Corner* (1908). Debussy's spontaneity and the sensitive nature of his perception facilitated his acute insight into the child mind, an insight noticeable particularly in *Children's Corner*, a French counterpart to Moussorgsky's song cycle *The Nursery*; in the *Douze Préludes* (first book, 1910; second book 1913) for piano; and in the ballet *La Boîte à joujoux* (*The Box of Toys;* first staged 1919).

In his later years, it is the pursuit of illusion that marks Debussy's instrumental writing, especially the strange, other-worldly *Cello Sonata*. This noble bass instrument takes on, in chameleon fashion, the character of a violin, a flute, and even a mandolin. Debussy was developing in this work ideas of an earlier period, those expressed in a youthful play he had written, *Frères en art* ("Brothers in Art"), where his challenging, indeed anarchical, ideas are discussed among musicians, painters, and poets. (He had in fact published in one of the anarchist journals poems that he had written and that he later set to music in the song cycle *Proses lyriques* [1893].)

Unique quality of Debussy's music

Debussy's music marks the first of a series of attacks on the traditional language of the 19th century. He did not believe in the stereotyped harmonic procedures of the 19th century, and indeed it becomes clear from a study of mid-20th-century music that the earlier harmonic methods were being followed in an arbitrary, academic manner. Hence his formulation of the "21-note scale" designed to "drown" the sense of tonality, though this system was never adhered to in the inflexible manner of the 12-note system of Schoenberg. Debussy's enquiring mind similarly challenged the traditional orchestral usage of instruments. He rejected the traditional dictum that string instruments should be predominantly lyrical. The pizzicato scherzo from his *String Quartet* (1893) and the symbolic writing for the violins in *La Mer*, conveying the rising storm waves, show a new conception of string colour. Similarly, he saw that woodwinds need not be employed for fireworks displays; they provide, like the human voice, wide varieties of colour. Debussy also used the brass in original colour transformations. In fact, in his music, the conventional orchestral construction, with its rigid woodwind, brass, and string departments, finds itself undermined or split up in the manner of the Impressionist painters. Ultimately, each instrument becomes almost a soloist, as in a vast chamber-music ensemble. Finally, Debussy applied an exploratory approach to the piano, the evocative instrument *par excellence* since notes struck at the keyboard are, by the nature of the piano mechanism, neither eighth notes, quarter notes, nor half notes, but merely illusions of these notes.

During the latter part of his life Debussy created an alter ego, "Monsieur Croche," with whom he carried on imaginary conversations on the nature of art and music. "What is the use of your almost incomprehensible art?" Monsieur Croche asks. "Is it not more profitable to see the sun rise than to listen to the Pastoral Symphony of Beethoven?" Elsewhere Monsieur Croche supports the cause of the musical explorer: "I am less interested in what I possess than in what I shall need tomorrow."

Debussy died in Paris of cancer on March 25, 1918, while the city was being bombarded by German guns. In his last works, the piano pieces *En blanc et noir* (1915; *In Black and White*) and in the *Douze Études* (1915), Debussy had branched out into modes of composition later to be developed in the styles of Stravinsky and the Hungarian composer Béla Bartók. It is certain that he would have taken part in the leading movements in composition of the years following World War I had his life not been so tragically cut short.

Debussy's influence

In a sense Debussy's influence on 20th-century music has increased since his death; and yet, paradoxically, it has also diminished. Wagner, said Debussy, was a wonderful sunset that had been mistaken for a dawn. As one looks back on the music of the last century this seems a remarkably shrewd observation. It was true of Wagner, of course, but it is now seen to be more true of Debussy himself. The fact is that there comes a time when the peak, the zenith of a civilization is reached. Critics have frequently noted this evolutionary stage in the music of Wagner, Debussy, or in one of their followers. A quintessential spirit is presented by these composers; and it seemed at the time that they could never be surpassed. But of course it is at this very stage that a decline in musical values sets in. Hence the paradoxical element in Debussy's stature. Undoubtedly, he was aware of this duality in his achievement, as may be gathered from his searching, hesitant letters. Sensitive to sham in every sphere and also a child of his environment, he not only perceived this dual aspect of his work but also realized the extent to which he himself was caught up in this vast evolutionary transformation.

Debussy's work cannot be judged on the musical level alone. "One must seek the poetry in his work," said his friend the French composer Paul Dukas. There is not only poetry in his music; there is often an inspiration from painting. "I love painting [*les images*, a generic term that might apply to the whole of Debussy's work] almost as much as music itself," he told the Franco-American composer Edgard Varèse. This association of the arts is a theme that runs through the whole of the 19th century—it originated with the theories of the German short-story writer E.T.A. Hoffmann—but for Debussy it was a theory more sensitively expressed in the tales of Poe. Throughout his life Debussy planned to set "The Fall of the House of Usher" in the form of an opera —the shadow of the tale never having been realized in *Pelléas et Mélisande*—and actually signed a contract for the production of this work at the Metropolitan Opera in New York, but it was never completed. The fact is that the hero of the tale, Roderick Usher, was a hypersensitive being like Debussy himself—a poet, a painter, and a musician. Moreover, the reputation of Poe was, during Debussy's life and after, almost entirely a French reputation. The French poets translated his works, and the French painters appreciated his genius;

and it was therefore only natural that a French musician should similarly have reflected the nature of his appeal. Debussy's adaptation of Poe's work, published only in the form of an opera libretto, is a clear indication of the cast of his artistic mind.

MAJOR WORKS

Piano music

SOLO PIANO: *Deux Arabesques* (1888); *Suite bergamasque* (1890–1905); *Estampes* (1903); *Images*, two sets (1905 and 1907); *Children's Corner* (1908); *Douze Préludes*, two books (1910 and 1913); *Douze Études*, two books (1915).

TWO PIANOS: *En blanc et noir* (1915).

Orchestral music

Prélude à l'après-midi d'un faune (completed 1894); *Nocturnes* (1899); *La Mer* (1905); *Images* (1912).

Ballet

Jeux (first performed 1913).

Vocal music

OPERA: *Pelléas et Mélisande*, on a text by Maeterlinck (first performed 1902).

CANTATA: *L'Enfant prodigue* (1884), awarded Grand Prix de Rome; *La Damoiselle élue* (1888), text by Dante Gabriel Rosetti, trans. by Gabriel Sarrazin.

INCIDENTAL MUSIC: *Le Martyre de Saint-Sébastien* (first performed 1911), to the mystery play by Gabriele D' Annunzio.

UNACCOMPANIED CHOIR: *Trois Chansons de Charles d'Orléans* (1908).

SONGS: Over 50 songs, including: *Ariettes oubliées* (1888), text by Verlaine; *Fêtes galantes*, two sets (1892 and 1904), texts by Verlaine; *Proses lyriques* (1892–93); *Chansons de Bilitis* (1897), text by Pierre Louÿs; *Le Promenoir des deux amants* (1904–10); *Trois Ballades de François Villon* (1910); *Trois Poèmes de Stéphane Mallarmé* (1913).

Chamber music

String Quartet (1893); *Syrinx* (1912), for unaccompanied flute; *Sonata for Cello and Piano* (1915); *Sonata for Flute, Viola and Harp* (1915); *Sonata for Violin and Piano* (1917).

BIBLIOGRAPHY. EDWARD LOCKSPEISER, *Debussy: His Life and Mind*, 2 vol. (1962–65), a biographical, aesthetic, and psychological investigation, contains details of the ten publications of Debussy's correspondence and a discussion of Debussy's critical articles and librettos. LEON VALLAS, *Claude Debussy et son temps* (1932; Eng. trans., *Claude Debussy: His Life and Works*, 1933), is the first and most reliable of the early source books on Debussy. On individual works, see ALFRED D. CORTOT, "La Musique pour piano de Claude Debussy," *La Revue musicale* (1920; Eng. trans., *The Piano Music of Claude Debussy*, 1922); and MAURICE EMMANUEL, *Pelléas et Mélisande de Claude Debussy* (1926, reprinted 1950).

(E.L.)

Decapoda

Decapods comprise a moderately large group of specialized, usually aquatic crustaceans commonly known as shrimp, lobsters, crayfish, hermit crabs, and crabs. More than 8,500 Recent species are known, the total number approaching one-third of all known crustaceans. The name decapod refers to the presence of five pairs of thoracic legs (pereiopods), usually modified for grasping, digging, walking, or swimming; the first pair of legs is usually modified into a pincher, or chela (cheliped).

Some decapods are among the more common animals of the seashore, and, because of their abundance and occurrence in many habitats, they have been known to man since earliest times. A few species of decapods, including many lobsters, some shrimps, and a few crabs, form the basis for economically important commercial fisheries. Many other species are sought for food locally but have limited economic value. Most of the known decapods are too small or rare to have any potential economic importance, but some of the smaller species are important as food items of higher animals. A few decapods serve as secondary hosts for parasites of man and other animals.

Size range of decapods

Decapods exhibit an enormous range of size as adults, from shrimps and crabs one centimetre long (less than one-half inch) to the giant spider crab of Japan (*Macrocheira kaempferi*), which measures almost 400 cm (12 ft) between the tips of its outstretched claws and may measure 45 cm (18 in.) across its body, making it the largest living arthropod (the phylum including invertebrates such as insects, arachnids, and crustaceans). An Australian crab, *Pseudocarcinus gigas*, may span 41 cm (16 in.) across the body and may weigh more than 13.5 kilograms (about 30 pounds). The American lobster (*Homarus americanus*) may attain a weight of 20 kg (44 lb), although individuals that size are rare.

By courtesy of (B, left) National Marine Fisheries Service; from (A) *Invertebrate Identification Manual* by Richard A. Pimental © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company; (B, right) *U.S. National Museum Bulletin* ((1930) in A. Kaestner, *Invertebrate Zoology*, © 1970 by John Wiley & Sons, Inc.; (C) *Natural History of Economic Crustaceans of the United States* (1893) in A. Kaestner, *Invertebrate Zoology;* (E) A.B. Williams, U.S. Fish and Wildlife Service, *Fishery Bulletin # 65* (1965) in A. Kaestner, *Invertebrate Zoology;* (D) M.E. Christiansen, *Decapoda Brachyura: Marine Invertebrates of Scandinavia*, Universitetsforlaget, Oslo; (F) after Calman from Schellenberg in F. Dahl, *Die Tierwelt Deutschlands*, vol. 10, Gustav Fischer Verlag

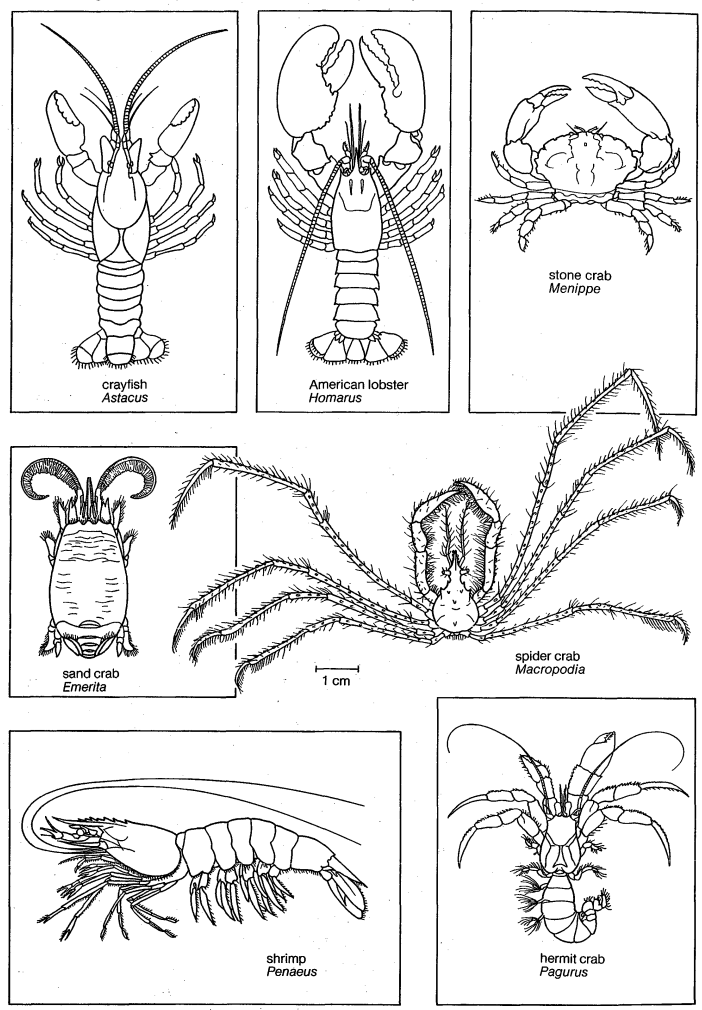


Figure 1: Diversity among decapods.

Members of this order exhibit great diversity in structure. Two basic body types can be recognized, the macrurous and the brachyurous. In the macrurous, or shrimplike, type of body organization, the body is elongate, often compressed laterally, with a long abdomen and a well-developed tail fan. The legs, usually two or more pairs of the anteriors, may be chelate and are often long and slender. In the brachyurous, or crablike, body organization, found in a few decapod groups, the body is flattened and laterally expanded, the abdomen folded beneath the carapace, and the tail fan reduced. The legs are often stout and short, and only the anteriors have claws.

Decapods are primarily marine animals, occurring in all oceans, from the intertidal zone to a depth of about 5,500 metres. In terms of numbers of species, they are most abundant in shallow waters and in warm tropical environments. Although the majority of decapods are marine, perhaps as many as 10 percent of the known species occur in freshwater or terrestrial habitats. Members of the caridean families (crustaceans in which the lateral plate

of the second abdominal segment overlaps the first) Atyidae and Palaemonidae, the macruran families (crustaceans with well-developed abdomens) Astacidae, Parastacidae, and a few relatives, the anomuran family (crustaceans with abdomens more or less reduced) Aeglidae, and the brachyuran family (crustaceans with greatly reduced abdomens) Potamidae, occur in fresh water. Many of these forms are restricted to freshwater habitats. Decapods utilizing terrestrial habitats are less numerous and include members of the anomuran family Coenobitidae as well as crabs of the families Ocypodidae, Gecarcinidae, and Grapsidae.

IMPORTANCE TO MAN

Some decapods form the basis for economically important commercial fisheries in areas around the world.

Species of shrimps and crabs, and many of the lobsters of the families Nephropidae and Palinuridae are caught in commercial quantities in both the Atlantic and Pacific oceans. The American lobster supports a small but valuable fishery off the northeastern United States, and the related *Homarus gammarus* and *Nephrops norvegicus*, are fished extensively in European waters. *Nephrops norvegicus*, known as the dublin prawn, scampi, and langostino, is a well-known commercial lobster.

Locally, burrowing decapods may have deleterious effects on farming and fishing. The land crab (*Cardisoma guanhumi*), from southern Florida and the Caribbean Sea, will eat young plants and is particularly destructive during spawning migrations, when local populations move as a unit. In Asia the mud shrimp *Thalassina anomala*, as well as a number of freshwater crabs, can seriously affect farming activities, not only by their burrowing activities but also by feeding on crop plants. The large swimming crab *Scylla serrata*, of the Pacific Ocean, is known to cause extensive damage to oysters but may itself be a locally important commercial species.

Although most decapods are edible, at least when thoroughly cooked, there are a few scattered references to poisonous crabs. In Hawaii, S.W. Tinker reports that the crab *Dromidiopsis dormia*, is regarded as poisonous by some people, but others are known to eat it. The 18th-century naturalist G.E. Rumpf (Rumphius) reported that a xanthid crab, *Eriphia sebana*, from Indonesia had caused two deaths. In a review of crab poisoning, L.B. Holthuis suggests that no species is poisonous, but that individual crabs might be toxic after ingesting toxic substances; he attributes some reports of poisonous crabs to other factors, including appearance, poor flavour, and allergic reactions by some people. Scientific evidence indicates that there are at least three species of poisonous crabs in the Ryukyu and Amani islands, Japan.

Crabs as carriers of human parasites

Although most decapods are not harmful when properly cooked, several of the Asian freshwater crabs serve as intermediate hosts for such parasites as the lung fluke. Although the flukes can be transmitted to a person eating the crab, they can be destroyed by cooking. In Africa, an eye disease, onchocerciasis, is caused by a nematode that can be transmitted by a fly, *Simulium*. Control of the disease was thwarted for many years until it was learned that fly larvae which had eluded medical entomologists for years, lived attached to the backs of river crabs.

NATURAL HISTORY

Life cycle. Most decapods brood their eggs. Only the shrimp of the section Penaeidea shed the eggs directly into the water. In the remainder of the decapods, the eggs are cemented to the abdominal appendages and carried until hatched. In crabs the abdomen, normally tightly appressed to the underside of the thorax, serves as a brood pouch. In shrimps a more open pouch is formed by the lateral extensions of the sides of the abdomen.

Most decapods hatch as free-swimming larvae, of which several basic kinds can be recognized. Although many different names have been applied to larval forms of different groups within the decapods, there are four basic larval types, each characterized partly by its mode of locomotion: the nauplius, with three pairs of appendages (antennules, antennae, and mandibles); the protozoea (from "stage preceding the zoea"), with three pairs of anterior appendages, as well as first and second maxillae and first and second maxillipeds; the zoea, with the remainder of the thoracic appendages; and the postlarva, in which the abdominal appendages are developed. Several molts can take place during each of these stages, with the appendages appearing in succession at each molt. The nauplius is believed to be the most primitive larval form, occurring in the section Penaeidea and in some other crustacean groups. The postlarva is a transitional stage between the pelagic larval stages and the juvenile stages in the life cycle. In benthic decapods, the postlarva is the first benthic stage; earlier stages are free-swimming.

In development, larvae undergo fundamental changes in methods of propulsion. In the nauplius and protozoea stages, propulsion is provided by the antennae, whereas in the zoeal stages it is by the thoracic appendages, as in some adult decapods of the open ocean. Only in the postlarval stage are the abdominal appendages developed enough to be used for locomotion.

In most groups the larva hatches in an advanced form, the zoea or an equivalent stage (mysis or phyllosoma). In relatively few species, the larval forms are completely suppressed or greatly reduced in number and the larvae pass through their metamorphic stages before hatching.

Shrimp life cycle

The commercially important pink shrimp (*Penaeus duorarum*) of the southeastern United States has a complicated life history, which may serve as an example of decapod development. Off the southwestern coast of Florida, adult shrimps are common in a large area with a depth of around 15–35 m (50 to 120 ft). Following copulation, the shrimp release their eggs in the open sea. The larvae hatch 13–14 hours later and make their way to the shallow, protected waters of the estuary in Everglades National Park. During their journey, the larvae pass through five naupliar, three protozoeal, three mysis, and three postlarval stages. Young shrimps spend up to seven months in the estuary and grow to a length of 7–10 cm (3–4 in.). At this size they begin a return migration to the spawning grounds, which may take as long as two months. In general, the shrimps travel and are active only at night. During the day they burrow into the bottom. The remainder of their life is spent in the area of the spawning grounds.

Behaviour. Complicated behaviour patterns are known in many decapods but perhaps are best known in some of

Types of Larval Development in Decapoda

group	larval forms
Suborder Natantia	
Family Penaeidae	Nauplius→protozoea→mysis*→mastigopus†
Family Sergestidae	Nauplius→elaphocaris‡→acanthosoma*→mastigopus†
Section Caridea	Protozoea→zoea→parva†
Section Stenopodidea	Protozoea→zoea→postlarva
Suborder Reptantia	
Section Macrura	
Superfamily Scyllaridea	Phyllosoma*→puerulus†, nisto†, or pseudibacus†
Superfamily Nephropidea	Mysis*→postlarva
Section Anomura	Zoea→glaucothoe† or grimothea†
Section Brachyura	Zoea→megalopa†

*Zoea. †Postlarva. ‡Protozoea.
Source: T.H. Waterman and F.A. Chace (1960).

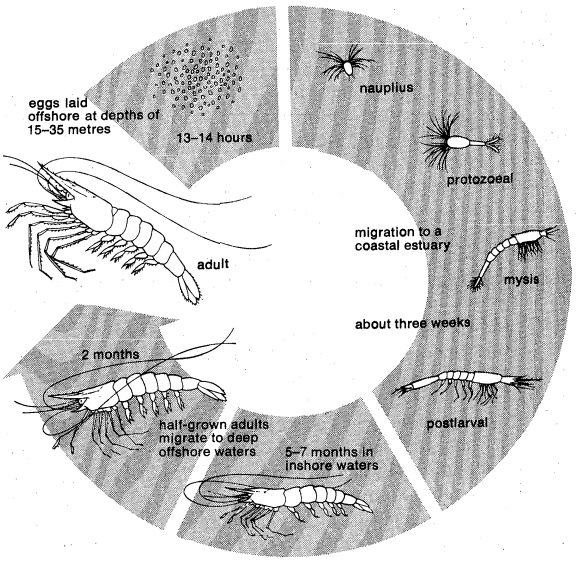


Figure 2: Life cycle of the pink shrimp.

the fiddler crabs of the genus *Uca*, which generally inhabit areas on the shore exposed at low tide. In *U. maracoani*, a series of six phases of activity patterns have been observed. In the underground phase, during high tide, the animals remain in their burrows. As the tide recedes, the maintenance activity phase begins, devoted primarily to feeding and burrow repair. Next, in the nonaggressive wandering phase, the animals appear to move at random and feed, exhibiting, as in the previous phase, no defensive or aggressive behaviour. This phase may be transformed into an aggressive wandering phase in which males may attack other males and pursue females. Preliminary courtship behaviour occurs during this phase. Then, in the territorial phase, the male occupies and defends a burrow. Finally, in the display phase, the male displays in the vicinity of his burrow. Display behaviour (beckoning) by males of *Uca* is intensified by the approach of a female. During this behaviour, the large claw is extended slowly to the side, opened and raised slowly, and then retracted rapidly. In *Uca*, the display pattern known as beckoning is distinctive in each species. All of the known fiddler crabs can be classified into two large groups based on their beckoning patterns. These patterns are correlated with the shape of the carapace (dorsal shell). Species with a broad front on the carapace share similar display patterns that distinguish them from species with the narrow front. In *Uca*, perhaps more than in any other group of decapods, studies on behaviour, ecology, and morphology have combined to yield a great deal of information on the natural affinities and the classification of the animals.

Display behaviour of fiddler crabs

A variety of activities relating to cleaning the body have been observed in decapods. In the shrimps, the second and third pereiopods are used to clean other appendages and the body; occasionally other legs are used also. The cleaning claws are often equipped with brushlike tufts of setae (bristles). The pleopods (abdominal appendages) are thoroughly cleaned before egg laying. For cleaning the eyes, fiddler crabs may use the small claw, the large claw, and the third maxillipeds, which may also clean the antennules and antennae.

Various forms of masking or camouflage behaviour are known in hermit crabs and in several groups of brachyuran crabs as well. The former are specially adapted to inhabit empty snail shells. The hermit crabs *Pagurus prideauxii* and *Dardanus arrosor* find a sea anemone, detach it from its normal substratum, and attach it to the gastropod shell that is used to protect the soft abdomen of the animal. As the crabs grow they change shells, selecting a larger one with each molt, and the anemone is transferred to the new shell. Many brachyuran crabs practice camouflaging behaviour. Some spider crabs select pieces of algae or colonies of hydroids or bryozoans and attach these to the dorsal surface of the carapace. Often the crabs are ornamented with small, hooked hairs to which algae and other foreign objects can be attached. The sponge crab *Dromia*, however, shapes a piece of sponge, which it carries on its back using the fifth pereiopods. This crab has been observed to use a variety of mechanisms for placing the sponge on its back. Other crabs carry half of the shell of a bivalve mollusk instead of a sponge.

Locomotion. Adult decapods utilize two basic types of locomotion, swimming and crawling, and different appendages may be used for each type. Most pelagic shrimps swim by means of long, fringed exopods on the thoracic appendages or by the pleopods. Crabs of the family Portunidae have the fifth pair of pereiopods modified into swimming paddles. In benthic decapods, all or most of the pereiopods, including the chelipeds, may be used in walking.

In the various stages of the life history of a given species, the individual may use different types of locomotion, involving different appendages, body orientation, and neuromuscular control. A swimming portunid crab hatches as a zoea and swims with its anterior end down and its dorsal surface forward, using the thoracic appendages for propulsion. As a postlarva, or megalops, it swims forward with the pleopods and may also walk sideways with four pairs of legs. As an adult it may walk sideways, utilizing three pairs of legs, or swim sideways utilizing the paddle-shaped fifth legs.

The macruran decapods share one mode of locomotion that they use as an escape mechanism. By strongly flexing the abdomen under the thorax, these animals can dart backward rapidly. Many of the thoracic appendages may be extended anteriorly, achieving a streamlined effect. The animals may use a few rapid flips of the abdomen to move short distances, but some, including lobsters of the genus *Panulirus*, may flex the abdomen many times and move for considerable distances. *Panulirus* has been reported to move 0.9–1.2 metres per second utilizing this mechanism.

Burrowing techniques

Decapods use a variety of techniques for burrowing. Burrows may be excavated by digging motions with the thoracic legs. Some shrimps burrow by rapidly beating the pleopods, and the mole crab *Emerita* uses its uropods. To transport material from the burrow the pereiopods may be used, but the mud shrimp *Callianassa* uses the third maxillipeds. *Uca* carries out mud or sand utilizing pereiopods on the side with the small chela, and these same legs may be used to push the mud pellet away from the mouth of the burrow. Several forms, including the mole crab, have basic modifications of the body, in addition to those of the appendages, as adaptations to a burrowing existence.

Reproduction. The sexes are separate in the majority of decapods. Protandrous hermaphroditism, in which the animal spends a period of its life as a male then reverses its sex to become a female, is known among a few species of shrimps of the families Hippolytidae and Pandalidae. One species of thalassinid shrimp, *Calocaris macandreae*, is a functional hermaphrodite. In other decapods, paired testes or ovaries usually lie in the thorax but may extend into the abdomen; in hermit crabs the gonads are abdominal and may be located on the left side only. Paired male genital openings are on the ventral surface of the eighth thoracic somite; paired female openings are on the sixth thoracic somite.

Sperm is usually passed from the male to the female in spermatophores formed in the terminal portion of the male genital ducts (vasa deferentia). Each vas deferens is modified into an ejaculatory duct that opens on the coxa, or between the coxa and the basis, of the fifth leg. Rarely, the male genital openings are sternal in position. The first, and often the second, pleopods are modified in males as gonopods for transmission of spermatophores. In many crabs and in the crayfishes, the form and ornamentation of the gonopod is unique for each species. The female is grasped by the male, and spermatophores are attached to the ventral surface of the female or, more often, are

placed by the male in a seminal receptacle or spermatheca of the female.

In most macrurous decapods, as well as in some crab groups (gymnopleurans and dromiaceans), the oviducts open on the coxae of the third pereiopods and the spermathecae are sternal; in these groups fertilization is external. In the other crab groups, the female genital openings are on the thoracic sternum, the spermathecae are enlargements of the oviducts, and fertilization is internal.

Timing of mating

There are several different patterns of mating activity, many of which apparently are linked to the molting cycle. Crabs of the genus *Maia* undergo a molt of puberty at a relatively large size and do not molt again after becoming sexually mature. In *Carcinus*, which may molt and grow indefinitely, the copulatory appendages appear and the animal reaches sexual maturity after the molt of puberty. In both *Carcinus* and *Maia* there may be a definite prepuberal molt. In the shrimp *Lysmata seticaudata*, many individuals of which are protandrous hermaphrodites, a molt of prepuberty precedes that of puberty by about six months. At the molt of puberty each individual becomes a male, but 18 months later the animal undergoes a critical molt during which the male appendages are shed. The shrimp may then be a sterile intersex (a temporary condition) or a functional female, depending on its level of food reserves.

Male decapods can copulate only when their exoskeleton is fully hardened, but in the crabs, at least, females need not be fully hardened. In the Cancridae, Portunidae, and some Majidae, the females can mate only following a molt, when they are soft. In other majids, *Hyas coarctus* and *Maia squinado*, the females may be either hard or soft. In crabs of the families Xanthidae, Grapsidae, and Ocypodidae most mating females are hardened. R.G. Hartnoll suggests that soft-female mating, correlated with a simple vagina structure, is the primitive condition, and from this, on several occasions, hard-female mating evolved, correlated with a more complex vagina structure.

Ecology. Although most decapods, in terms of major groups and numbers of species as well, have adapted to a benthic existence, members of some shrimp groups are well adapted to life in the open ocean. Among these are shrimps of the penaeidean families Penaeidae and Sergestidae, as well as some caridean shrimps, principally of the families Oplophoridae, Pasiphaeidae, Pandalidae, and Bresiliidae. These shrimps are generally streamlined and relatively soft bodied, usually with a very thin, membranous integument, although in some oplophorids the carapace may be quite hard. The swimming appendages are well developed. In most groups the thoracic legs are provided with functional exopods and the pleopods are also strongly developed.

Decapods of the open ocean

Shrimps living near the surface are usually unpigmented, clear in colour, or with some pigment in scattered chromatophores; those living just below the surface may show more pink coloration. Mesopelagic shrimps (free-swimming in midwater) are usually red, sometimes quite dark. In some species of *Notostomus* the abdomen is scarlet and the carapace darker, almost black. Some of these shrimps live to a depth of 5,000 m (about three miles), but below 500–1,000 m the red pigment appears black, for the red rays of sunlight are filtered out by the surface waters.

Some deep-sea shrimps possess light organs (photophores); particularly species of *Sergestes* and representatives of two genera of Oplophoridae, *Systellaspis* and *Oplophorus*. *Sergestes* may have 125 to 150 photophores, which in *S. prehensilis* are thought to light up in succession from head to tail. The photophores may occur on the eyestalks, the margin of the carapace, the roof of the branchial chamber, on the abdomen, or on the legs of the thorax and abdomen. There appears to be a correlation between photophore development and eye size, suggesting that the photophores aid in feeding or species recognition. Some oplophorids are able to discharge clouds of luminescent material from glands near the mouth.

Pelagic shrimps undergo daily vertical migrations, possibly in response to changes in light intensity and food concentrations. Shrimps living between 200 and 1,200 metres in the sea have been found to undergo vertical migrations of 200 to 600 in a 24-hour period. *Acanthephyra purpurea* has been found at 800 m at noon and at 200 m at midnight, a journey of 600 m in 12 hours or a speed of vertical movement of some 50 m per hour.

Decapods have also taken advantage of one specialized open-ocean habitat, that afforded by the Sargassum weed, *Sargassum natans*. A portunid crab, *Portunus sayi*, the grapsid crabs *Planes minutus* and *Pachygrapsus marinus*, and several shrimps are commonly found living on the weed at sea. All are a mottled brown colour like the weed. Two of the shrimps, *Latreutes fucorum* and *Leander tenuicornis*, also occur on seaweeds in shallow waters. Species of *Planes* are also commonly found on flotsam at sea.

Terrestrial decapods

The terrestrial environment has been invaded by relatively few decapods. Some representatives of the anomuran family Coenobitidae, including the species of the hermit crab genus *Coenobita* and the robber crab *Birgus latro*, and true crabs of the family Gecarcinidae, are terrestrial. Ocypodid crabs, including species of the fiddler crab *Uca* and the ghost crab *Ocypode*, also live on the shore but are more dependent on the sea than the gecarcinids. In colonizing terrestrial habitats, these decapods have evolved the ability to regulate the internal concentrations of their body fluids and have accompanying mechanisms to protect against dessication and the allied problem of overheating. In addition, they have developed methods of communication by visual and acoustic signals. According to D.E. Bliss, behavioral modifications involving primarily burrowing activity and physical movements, help to minimize water loss, aid exposure to favourable rather than extreme environmental conditions, and maintain a balance between evaporative cooling and dehydration. As a general rule, terrestrial decapods spawn in the sea.

In adult land crabs, water balance is maintained by the gills, gut, and pericardial sacs, which assimilate, store, and recirculate salts and water under the control of the central nervous system.

Respiration in terrestrial decapods has been made possible by extensive vascularization of the gill surfaces. Apparently invasion of the terrestrial environment has been accompanied by reduction in the number or the volume of the gills, vascularization of epithelial surfaces in the branchial chambers, enlargement of the gill chambers (in crabs such as *Gecarcinus* and *Uca*), or vascularization of the epithelial surfaces (in the hermit crab *Coenobita*, which apparently can survive amputation of the gills).

Most terrestrial or semiterrestrial crabs are found in the tropics, but some species of *Uca* extend northward into temperate climates. Bliss has described the zonation of *Ocypode*, *Cardisoma*, and *Gecarcinus* on Bimini, Bahama Islands. *Ocypode quadrata* remains on beaches where it can enter the ocean freely to keep its gills moist and help maintain normal concentrations of body fluids. *Cardisoma guanhumi* is only found in areas of relatively low elevation, where it constructs burrows extending into fresh water. *Gecarcinus lateralis* shows some overlap with *Cardisoma* but is found over the entire island. Its burrows do not extend into water and it can live almost independently of the aquatic environment. Among modifications developed by this crab are the highly vascularized epidermis of the branchial chambers, enormously developed pericardial sacs, which extend into the branchial chambers, and the ability to absorb water from dew or rain through tufts of hairs on the underside of the body.

In the West Indies the hermit crab *Coenobita clypeatus* also has evolved a high degree of independence from the sea. On the island of Dominica it is found some two miles from the shore and at an altitude of nearly 400 metres (1,300 feet). Like the marine hermit crabs, *Coenobita* must inhabit an empty gastropod shell to protect its uncalcified abdomen. Its relative, the robber or coconut crab of the Pacific, *Birgus latro*, is unusual among the hermit crabs in having a calcified abdomen. *Birgus* is an active

crab and an agile climber, apparently omnivorous, but possibly with a preference for coconuts. It is eaten by man in areas where it still exists.

Fresh water decapods

Relatively few decapod groups have successfully colonized fresh water. Those with fresh water representatives include the caridean shrimp family Palaemonidae (especially species of the genera *Macrobrachium* and *Palaemonetes*), all of the shrimps of the family Atyidae, all members of the crayfish families Astacidae (Europe, North and Central America) and Parastacidae (South America, Australia, Indo-Malaya, and Madagascar), and all species of the genus *Aegla* (the only anomurans restricted to fresh water). Among the crabs, the families Potamidae (Old World) and Pseudothelphusidae and Trichodactylidae (both New World) are exclusively freshwater. Some other crabs, including members of the family Grapsidae such as the Chinese mitten crab (*Ereiocheir sinensis*) and species of *Platycheirograpsus*, also live in fresh water. One species, *Metopaulias depressus*, lives in fresh water in West Indian plants of the bromeliad family. Although some of these species, including certain members of the genera *Macrobrachium* and *Ereiocheir*, must return to salt water to spawn, many go through their entire life cycle in fresh water. In most of those spawning in fresh water, larval development is abbreviated and the young often hatch as miniature adults.

The fresh water habitat has probably been invaded independently by several evolutionary lines. The crayfishes appear to be the oldest inhabitants of fresh water among the decapods, judged by the fact that their entire life cycle is spent there. Those species that must return to the sea to spawn are considered to be more recent invaders of the habitat.

Penetration of fresh water is dependent upon the organism's ability to maintain its blood concentration at a level different from the medium. Blood concentration of marine decapods is usually equal to the medium (isotonic), whereas in fresh water form it is much higher. In adapting for life in fresh water, decapods have developed several mechanisms, including reduction of the permeability of the body surface, development of greater tolerance for variations in blood concentrations, and the ability to transport inorganic ions in or out of the body against a concentration gradient.

Symbiosis and parasitism

Representatives of several different decapod groups have developed commensal relationships (in which the guest symbiont does not use the host tissue as food) with other animals, particularly echinoderms, mollusks, and coelenterates (cnidarians), including corals, sea anemones, and sea whips. Commensal relationships exist in more than 80 species of the palaemonid shrimp subfamily Pontoniinae, certain hippolytid shrimps, some snapping shrimps, and several stenopodidean shrimps. Most anomuran decapods are free-living, but some galatheid lobsters live on crinoids and numerous porcellanid crabs live as commensals. The West Indian *Porcellana sayana* often is found in the shell carried by the large hermit crab *Petrochirus*. Although hermit crabs are free-living, several carry anemones or bryozoan colonies on their shells, and in some species the bryozoans dissolve the shell, the colony becoming the shelter for the crab.

Among the crabs, the dromiaceans or sponge crabs carry a live sponge or other object on their backs. Most of the brachyrhynch crabs are free-living, but a few groups are commensal with sponges, coelenterates, mollusks, polychaete worms, or echinoderms. A pea crab, *Pinnotheres ostreum*, which lives in the American oyster and sometimes in other invertebrates, may be considered as a parasite, for its feeding activities on the mucous masses of the bivalve gills leaves evidence of gill damage. Another pinnotherid, *Dissodactylus*, can be found living on echinoderms. The family Hapalocarcinidae contains the gall crabs, of which the female forms a pocket within a living coral and spends her adult life there. Relatively few of the oxyrhynch crabs are commensals, but all members of the subfamily Eumedoninae, family Parthenopidae, are commensals of echinoderms.

Decapods may also serve as hosts of commensal animals; many of the species that form burrows act as hosts for a variety of animals. The land crab *Gecarcinus* may share its burrow with the minute fly, *Drosophila*, which lays its eggs on the crab; larval flies may be found on the mouthparts or in the gill chambers. One of the common thalassinidean shrimp from the west coast of the United States, the ghost shrimp or mud shrimp *Callianassa californiensis*, shares its burrow with no less than nine other species: a gobiid fish, three species of pea crabs, a snapping shrimp, two species of copepods, a polychaete worm, and a small clam. As many as five of these, the fish, the clam, the worm, a pea crab, and a copepod, can be found in the burrow at one time. The blue mud shrimp of the California coast (*Upogebia pugettensis*) also serves as host to a number of commensals and, in addition, may be parasitized by a pair of isopods.

Shrimps of several families in the Atlantic and Pacific oceans have developed a very specialized symbiotic relationship with fishes. An individual shrimp apparently occupies a "cleaning station" to which fish come to have parasities removed. Some fish, ordinarily predatory on crustaceans, will allow the shrimp to enter the mouth and gill chambers to search for parasites. Brightly coloured shrimps of the genera *Periclimenes* (family Palaemonidae), a genus containing many species commensal on other invertebrates, *Hippolysmata* (family Hippolytidae), and *Stenopus* (family Stenopodidae) are known to exhibit cleaning behaviour.

FORM AND FUNCTION

Basic body plan

All decapods share a basic body plan, in which the body is a modified tube made up of 19 distinct rings or segments, to each of which is attached one pair of appendages. Each body segment or somite comprises a dorsal portion, the tergum; a ventral portion, the sternum; and, on each side, a flap of the tergum that extends below the sternum, the pleuron. The appendages are attached to the sternum by a membrane. In addition to the 19 somites bearing appendages, there is an anterior portion of the body bearing the stalked, compound (multifaceted) eyes, and a terminal portion, the telson, through which the alimentary canal opens to the exterior via the anus; the telson lacks appendages. Decapods, like other eumalacostracan crustaceans, have three distinct body regions: head, thorax, and abdomen. In decapods the head and thorax are fused and may be referred to as the cephalothorax. One of the characteristics of decapods is the head shield, or carapace, that covers the cephalothorax dorsally and extends laterally over the gills at the bases of the thoracic appendages. There are five pairs of head appendages, eight pairs of thoracic appendages, and six pairs of abdominal appendages.

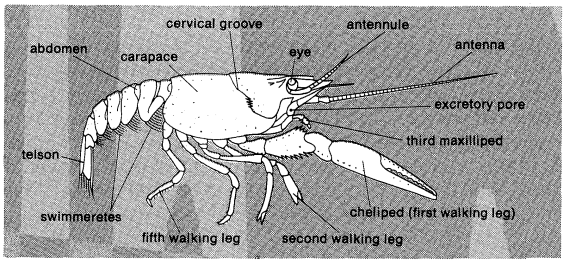


Figure 3: Typical external structure of the decapod body.

Appendages. Decapod appendages all appear to have been modified from one basic plan, in which two branches, an inner endopod and an outer exopod, arise from a basal portion, the protopod, which is attached to the body. The protopod comprises two indistinguishably fused segments, the praecoxa and coxa, and a distinct segment, the basis. The endopod and exopod both attach to the basis. The coxa may bear a separate branch, the epipod. The endopod typically consists of five segments, the ischium, merus, carpus, propodus, and dactylus; the walking legs and chelipeds of decapods are formed by the endopod.

The anterior two pairs of head appendages, the antennules (first antennae) and the antennae (second antennae), serve sensory functions, including olfaction, touch, and balance. The antennules typically consist of a three-segmented stalk with two many-jointed whiplike flagella. In some shrimps one of the flagella may be secondarily subdivided and there may be stylocerite or antennular scale. The proximal segment of the stalk of the antennule usually contains an organ of balance, the statocyst, absent in some shrimps. The second antennae have a five-segmented stalk and a single flagellum. A movable, articulated exopod, the antennal scale or squama, may be present. In adult decapods, the proximal segment of the antennae each contain the opening of the excretory organ, the green gland.

Mouthparts

The posterior three head appendages serve as mouthparts. The heavily calcified mandibles may have distinct molar and incisor processes and often have a three-segmented mandibular palp. The mandibles are the crushing and chewing mouthparts. The maxillules or first maxillae and the maxillae (second maxillae) are the flattened, multilobed food manipulators. The exopod of the maxilla, called the scaphognathite or gill bailer, extends over the gills and acts as a pump to help maintain a constant stream of water over the gills.

The anterior three thoracic appendages or thoracopods, called the first, second, and third maxillipeds, also serve as mouthparts. The first and second maxillipeds are leaflike, resembling the maxillae. The third maxillipeds are elongate in the shrimps, but in the crabs each is highly modified as an operculum, or cover, for the mouth and mouthparts.

The posterior five pairs of thoracic appendages are the walking legs or pereiopods. They may be modified by loss of exopods and epipods, fusion of segments, development of extra segments, formation of pinching claws (chelae or subchelae) or swimming paddles. They may be reduced or even lost. In the brachyurous or crablike decapods, the first pereiopods usually are modified as chelae, and the fifth pereiopods may be modified into paddles, reduced in size, or absent. In many anomurans, the fifth pereiopods are chelate. In the macrurous or shrimplike decapods all five pereiopods may be chelate, but usually only the anterior two or three are so modified; in certain shrimps some segments of the chelae are multiarticulate.

Basically, the anterior five pairs of abdominal appendages are bilobed, leaflike swimmerets or pleopods, and, in macrurous forms, they may be used as swimming appendages. The anterior two pairs may be modified as sexual appendages, or gonopods, in males, serving in sperm transfer. They may be vestigial or even absent in females. In brachyurous forms the pleopods have lost their swimming function. In females they serve as attachments for the egg mass and in males the anterior two are modified into gonopods. The sixth or last pair of abdominal appendages are the uropods. In macrurous forms these form a tail fan with the telson, the terminal segment of the body. In brachyurous forms the uropods are usually absent, but in hermit crabs the uropods are modified into structures that help to hold the abdomen of the crab in the gastropod shell in which it lives.

Nervous system. The central nervous system in decapods is made up of a supraesophageal ganglion with lateral connections to a subesophageal ganglion. The ventral nerve cord may be connected to additional thoracic and abdominal ganglia. In most brachyurans the ganglia are fused into a central mass from which nerves radiate.

Sense organs

A variety of structures serve as sense organs. The eyes are usually well developed with a pigmented, multifaceted cornea; but they may be reduced in size, or lacking pigment, or even absent in some deep-sea and cave species. Balance organs (statocysts), which may contain a sand grain or other inclusion in a cavity lined with fine hairs (setae), are on the basal segment of the antennule. Tactile setae may be present on various parts of the body and olfactory setae are present on the antennular flagella.

Circulatory system. As in other crustaceans, the circulatory system is open. The heart is located to the rear of the carapace (cardiac region) above the gut. Blood flows from the heart through several arteries to various parts of the body, returns through the tissues into a ventral sinus, and passes through the gills to the pericardial sinus around the heart. The blood contains a respiratory pigment, hemocyanin.

Respiratory system. A series of gills, attached to the body wall of the thorax or to the bases of the thoracic legs, serves as respiratory organs. A variety of gills is found: podobranchs, attached to the coxae of the thoracic legs; arthrobranchs, attached to the articular membranes of the appendages; and pleurobranchs, attached to the body wall. In some cases the epipods may have a respiratory function. The number of gills present is always less than the theoretical maximum of 32 pairs. The number, position, and kind of gill are important characters for the higher taxonomic categories. Three basic gill types can be recognized: dendrobranchiate, with the main axes branched; phyllobranchiate, with the branches flattened and arranged in two series; and trichobranchiate, with the filaments arranged in series around the stem.

Water currents are maintained over the gills by the motion of the scaphognathite, the exopod of each maxilla. The current usually runs anteriorly from the bases of the legs, but it is reversed in some burrowing species.

Digestion and excretion. The gut or alimentary tract of decapods is basically a straight tube in which three regions can be recognized: the stomodeum, or foregut, the mesenteron, or midgut, and the proctodeum, or hindgut. The foregut, expanded anteriorly to form a stomach, and the hindgut have a chitinized or calcified lining; the cardiac or anterior portion of the stomach contains a movable, calcified masticatory apparatus known as the gastric mill. The midgut is ramified with numerous blind pockets or ceca and to it are attached ducts from the hepatopancreas or digestive gland. The highly ramified hepatopancreas extends through much of the cephalothorax; in the Paguroidea it extends into the abdomen.

The green gland or antennal gland, which opens on the basal segment of the antenna, is the primary excretory organ of the decapods. A maxillary gland with the same function, believed to be more primitive, is found in some decapod larvae. The opening of the antennal gland may be operculate in brachyurous forms.

EVOLUTION AND PALEONTOLOGY

The decapods first appeared in the Early Triassic or Late Permian periods with forms identified with peneideans (shrimps and prawns) and the fossil family Erymidae in the nephropideans. There was an expansion in the Early Jurassic Period, with representatives of fossil Nephropidea, Palinuroidea, Thalassinoidea, and Paguroidea, as well as one from the Brachyura. In the Middle and Late Jurassic, the decapods apparently adapted to the varied environments afforded by coral reefs, and a diversified crab fauna was developed, mainly with representatives of the fossil family Prosoponidae. The oldest crayfishes have been found in deposits of the Late Jurassic and Early Cretaceous periods, and in the latter period nephropids and palinurans began to outnumber the Erymidae and Glypheoidea, respectively. The crabs began to increase in numbers in the Early and Late Cretaceous, and by the Tertiary Period most modern groups were established. The Tertiary was a period of rapid diversification of the crabs, accompanied by a reduction in numbers of the reptant macrurans. Dromiaceans and raninids decreased in the Tertiary, whereas oxyrhynchs and xanthids flourished. The Tertiary also marked the appearance of freshwater shrimps and crabs.

Relatively little is known of the origin and evolution of the Decapoda, for many of the early shrimps left no fossil record, presumably because they were soft bodied. Both reptant and natant forms, the latter represented by forms similar to modern penaeideans, were established by the Triassic Period. M.F. Glaessner has pointed out that there is no evidence showing the relation between

the early penaeideans and the other decapods, but some zoologists consider the penaeideans to be the ancestral decapods. It has been suggested that the lobster-like fossil order Pygocephalomorpha or representatives of the unspecialized, shrimp-like fossil order Eocaridacea, or both, may be related to ancestral decapods.

CLASSIFICATION

Distinguishing taxonomic features. In classifying the Decapoda taxonomists primarily use characters based on external morphology. In the Natantia, the form of the appendages, the shape and decoration of the carapace and abdomen, and the number and kinds of gills are primary taxonomic characters. In certain Natantia and in the section Brachyura of the Reptantia the characters afforded by the male gonopod are often important at the specific level. In the Brachyura, features of the outer maxillipeds, the five pereiopods, and the structure and ornamentation of the carapace are used; the position and nature of the genital openings are important distinguishing features of the higher categories.

Annotated classification. The classification given below is based primarily on the higher categories as proposed by L.A. Borradaile (1907), with modifications within the Caridea proposed by L.B. Holthuis (1955); with a few other modifications, it is used by modern taxonomists. An alternative classification, proposed by M.F. Glaessner, follows the discussion of classification below.

Order Decapoda

Malacostracan crustaceans in which the anterior three thoracic appendages are modified as mouthparts (maxillipeds). No more than the posterior five thoracic appendages are locomotory. Gills are arranged in several series on the thoracic appendages. Larvae hatch as a nauplius or in a more advanced stage. Marine and freshwater, a few terrestrial. More than 8,500 Recent species.

Suborder Natantia (shrimps)

Body usually laterally compressed; rostrum usually well-developed. First abdominal somite not markedly smaller than remainder. Male genital apertures in articular membrane. Pleopods well-developed, used for swimming. Contains about 2,000 species.

Section Penaeidea. Pleura of second abdominal somite not overlapping those of first somite. Third pereiopods chelate, not stronger than first pair. First pleopods of male with petasma (a hooked plate). Female with thoracic thelycum (seminal receptacle). Eggs shed into water, larvae hatch as nauplius. Gills dendrobranchiate. About 325 living species.

Superfamily Penaeoidea. Fourth and fifth pereiopods and gills not reduced. Adults benthic in shallow water or pelagic in midwater or deep water.

Superfamily Sergestoidea. Fourth and fifth pereiopods are reduced. Gills reduced or absent. Adults pelagic.

Section Caridea. Pleura of second abdominal somite overlap those of first and third somites. Third legs not chelate. Males lacking petasma; second pleopod with two styliform appendices. Females lacking thelycum; second pleopod with one styliform appendix. Eggs attached to pleopods, hatching at stage later than nauplius. Gills phyllobranchiate. Over 1,600 Recent species.

Superfamily Oplophoroidea. Pereiopods usually with exopods. First pair of pereiopods chelate, usually slenderer than second, fingers of chela not unusually long. Carpus of second pereiopods not secondarily subdivided. Over 200 species, pelagic in open ocean (Oplophoridae) or benthic in fresh water (Atyidae).

Superfamily Stylodactyloidea. Pereiopods without exopods. First pair of pereiopods chelate, fingers long and slender. Distal two joints of second maxilliped set side by side, at end of antepenultimate segment. One family; less than 10 species.

Superfamily Pasiphaeoidea. First and second pereiopods chelate; cutting edges of fingers pectinate. One family containing about 60 species of pelagic shrimps.

Superfamily Bresilioidea. First pair of pereiopods chelate, usually stronger but shorter than second. About 15 species in 4 families.

Superfamily Palaemonoidea. Pereiopods without exopods. First pair of pereiopods chelate, usually slenderer than second. Carpus of second pereiopods not secondarily subdivided. Over 400 species in four families; the river shrimp, *Macrobrachium*, family Palaemonidae, contains the largest living shrimps, up to 19 cm in length.

Superfamily Psalidopodoidea. First pair of pereiopods with both fingers movable. One family with three species.

Superfamily Alphaeoidea. Carpus of second pereiopods is usually subdivided into two or more segments. First pair of pereiopods with distinct chelae. Over 600 species in four families, including the snapping shrimp, family Alpheidae.

Superfamily Pandaloidea. Carpus of second pereiopods usually subdivided into two or more segments. Chelae of first pair of pereiopods small or absent. About 120 species in three families, some pelagic.

Superfamily Crangonoidea. First pair of periopods subchelate. Fewer than 200 species in two families; benthic.

Section Stenopodoidea. Pleura of second abdominal somite not overlapping those of first. Third pereiopods chelate, usually stouter than first. Males lacking petasma; females lacking thelycum. Eggs carried on pleopods by female. Gills trichobranchiate. About 20 species in one family.

Suborder Reptantia

Body usually depressed. Rostrum often lacking. First abdominal somite usually smaller than remainder. Male genital apertures coxal or sternal. Pleopods usually reduced or absent, not locomotory.

Section Macrura. First and third pereiopods similar, subcylindrical or with chelae. Flagella of anterior three maxillipeds directed anteriorly. Abdomen natant-like, with well-developed pleura and broad tail fan. About 700 species in four superfamilies.

Superfamily Eryonidea. Carapace depressed, fused with epistome. Integument soft. Anterior four, or all five, pereiopods chelate. First pleopods absent. About 40 species, blind inhabitants of deep sea.

Superfamily Nephropidea. Carapace usually depressed; integument hard. Rostrum present. Anterior three pereiopods chelate, first usually stoutest. Includes the marine lobsters (Nephropidae) and the freshwater crayfishes. More than 300 species.

Superfamily Scyllaridea. Carapace depressed, integument hard. Rostrum absent. No legs chelate, except sometimes the fifth in female. Includes the spiny lobsters, Palinuridae, and Spanish or shovel lobsters, Scyllaridae; together, about 100 species.

Superfamily Thalassinidea. Carapace compressed, integument usually soft. First pereiopods chelate or subchelate, third pereiopods not chelate. Contains several anomuran-like families of uncertain position. Most of the 250 species are secretive, burrowing forms.

Section Anomura. Fifth pereiopods always differing from third in length, size, and shape. Abdomen not natant-like; usually reduced. Uropods usually present.

Superfamily Galatheidea. Abdomen recurved anteriorly, bent under carapace. First pereiopods chelate. Tail fan well-developed. Includes the squat lobsters (Galatheidae, Uroptychidae) and the porcelain crabs or rock sliders (Porcellanidae); the latter resemble true crabs but have uropods and chelate fifth pereiopods. About 600 species, mostly marine; one family, Aeglidae, lives entirely in fresh water.

Superfamily Paguridea. Abdomen asymmetrical, usually soft and unprotected; sometimes hardened and bent under thorax. First and occasionally fifth pereiopods chelate. Tail fan, if present, adapted for holding abdomen in shelter. Includes several families of marine hermit crabs, the terrestrial hermit crabs, and the king crabs; about 700 species are known.

Superfamily Hippidea. Abdomen symmetrical, bent under thorax. First pereiopods styliform or subchelate. Not chelate. Tail fan adapted for burrowing. Two families. Includes about 60 species of marine burrowing crabs, many of which live in sand in the intertidal zone.

Section Brachyura (true crabs). Carapace broad, lateral margins almost always well-developed. Abdomen symmetrical, bent under thorax, tail fan absent. First pereiopods chelate or subchelate. About 4,500 Recent species.

Subsection Gymnopleura. Carapace elongate, not covering anterior abdominal somites. First pereiopods subchelate. Oviducts open on coxa. Fifth pereiopods subdorsal in position. About 30 species of marine burrowing crabs in one family Raninidae; they may be related to subsection Oxystomata, below, but their position is uncertain.

Subsection Dromiacea. Carapace usually not elongate. First pereiopods chelate. Fifth pereiopods dorsal in position, modified to hold shells, sponges, and other objects over the body. Oviducts open on coxae.

Superfamily Dromiidea. Sternum of female with longitudinal grooves. Remnants of uropods usually present. About 175 species in three families; marine; includes sponge crabs.

Superfamily Homolidea. Sternum of female lacking longitudinal grooves. Uropods absent. Contains about 25 species living in deep water.

Subsection Oxystomata. Mouth frame triangular, produced anteriorly. Female openings usually sternal. Fifth pereiopods normal or modified as in Dromiacea. Includes purse crabs (Leucosiidae), box crabs (Calappidae), and masking crabs (Dorippidae); about 500 species known, all marine.

Subsection Brachygnatha. Mouth frame quadrate. Female openings usually sternal. Fifth pereiopods normal, not reduced or dorsal in position. Includes about 3,800 species of true crabs.

Superfamily Brachyrhyncha. Body not narrowed anteriorly. Rostrum reduced or absent. Orbits well developed. Includes the swimming crabs (Portunidae), mud crabs (Xanthidae), land crabs (Gecarcinidae), fiddler crabs (Ocypodidae), square back crabs (Grapsidae), pea crabs (Pinnotheridae), and several families of freshwater crabs; about 2,900 species.

Superfamily Oxyrhyncha. Body narrowed anteriorly. Rostrum usually present. Orbits usually poorly developed. Includes the spider crabs (Majidae) and several smaller families of crabs; about 900 species, all marine.

Critical appraisal. Present classifications and phylogenetic schemes suggested for the Decapoda are not accepted by all authorities, for several reasons. Evidence derived by neontologists from Recent species, based on mouthfield and mouthpart structure, gill number and position, and internal morphology, conflicts with evidence on fossil forms, based mostly on carapace morphology, presented by paleontologists. Further, evidence from studies on larval forms of Recent species also seems to be contradictory. For example, zoologists have suggested that the Brachyura arose from astacidean stock (morphology) or from thalassinidean stock (larval development), whereas the fossil evidence strongly supports the derivation of the Brachyura from the Glypheoida, a superfamily of fossil palinurans. In 1969 M.F. Glaessner reviewed earlier classifications and the phylogeny of decapods and proposed the following classification (groups indicated by a dagger are known only from fossils):

Order Decapoda
- Suborder Dendrobranchiata
  - Infraorder Penaeidea
    - Superfamily Penaeoidea
    - Superfamily Sergestoidea (no fossil record)
- Suborder Pleocyemata
  - Infraorder Stenopodidea (no fossil record)
  - †Infraorder Uncinidea
  - Infraorder Caridea
  - Infraorder Astacidea
  - Infraorder Palinura
    - †Superfamily Glypheoidea
    - Superfamily Eryonoidea
    - Superfamily Palinuroidea
  - Infraorder Anomura
    - Superfamily Thalassinoidea
    - Superfamily Paguroidea
    - Superfamily Galatheoidea
    - Superfamily Hippoidea
  - Infraorder Brachyura
    - Section Dromiacea
      - Superfamily Dromiodea
      - Superfamily Homoloidea
      - †Superfamily Dakoticancroidea
    - Section Oxystomata
      - Superfamily Dorippoidea
      - Superfamily Calappoidea
      - Superfamily Raninoidea
    - Section Oxyrhyncha
    - Section Cancridea
    - Section Brachyrhyncha
      - Superfamily Portunoidea
      - Superfamily Xanthoidea
      - Superfamily Ocypodoidea

In this classification, the first eight categories (the suborder Dendrobranchiata and the suborder Pleocyemata, up to and including Caridea) correspond to the suborder Natantia of the above annotated classification. The Thalassinoidea are considered by Glaessner to be anomurans. This classification emphasizes the distinctness of the Penaeidea and provides other modifications that make the conventional classification more acceptable to the paleontologist.

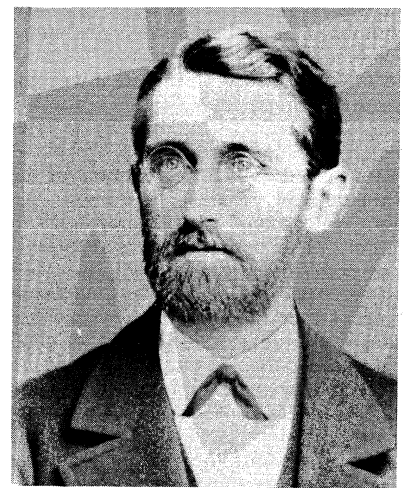
BIBLIOGRAPHY. H. BALSS and H.E. GRUNER, "Decapoda," in H.G. BRONN, *Klassen und Ordnungen des Tierreichs*, vol. 5, pt. 1, bk. 7 (1940–61), the most comprehensive survey available on the decapods, with an excellent bibliography; W.T. CALMAN, "Crustacean," in R. LANKESTER (ed.), *A Treatise on Zoology*, pt. 7, fasc. 3 (1909), a good basic introduction to the Crustacea, although the classification is out of date; D.B. CARLISLE and F. KNOWLES, "Endocrine Control in Crustaceans," *Camb. Monogr. Exp. Biol.*, no. 10 (1959), a review article on the roles played by hormones; M.F. GLAESSNER, "Decapoda," in R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology*, pt. R, *Arthropoda 4* (1969), the most recent comprehensive survey of decapods, based primarily on fossils, with a good bibliography; J. GREEN, *A Biology of Crustacea* (1961), a good, basic, general account; R. GURNEY, *Larvae of Decapod Crustacea* (1942), the classic account of decapod larvae, with general information on kinds of larvae and modes of development, bibliography excellent to time of publication; T.H. WATERMAN (ed.), *The Physiology of Crustacea*, 2 vol. (1960–61), a series of review articles by specialists on various aspects, each with a comprehensive bibliography; A.P.M. LOCKWOOD, *Aspects of the Physiology of Crustacea* (1967), a general survey of crustacean physiology that is a good supplement to Waterman; W.L. SCHMITT, *Crustaceans* (1965), an excellent popular introduction to the biology of crustaceans.

(R.B.M.)

Dedekind, Richard

Highly original, the German mathematician Julius Wilhelm Richard Dedekind developed a major redefinition of irrational numbers (see below) in terms of arithmetic concepts. Although not fully recognized in his lifetime, his treatment of the ideas of the infinite and what constitutes a real number continue to influence modern mathematics.

By courtesy of the Library of the Swiss Federal Institute of Technology, Zurich



Dedekind.

Dedekind was born the son of a lawyer on October 6, 1831, in Brunswick (modern Braunschweig), Germany. While attending the Gymnasium Martino-Catharineum in 1838–47 in Brunswick, he was at first interested primarily in chemistry and physics. At the Caroline College in 1848–50, however, he turned to calculus, algebra, and analytic geometry, which helped qualify him to study advanced mathematics at the University of Göttingen under the mathematician Carl Friedrich Gauss.

Early training

After two years of independent study of algebra, geometry, and elliptic functions, Dedekind served as *Privatdozent* ("unsalaried lecturer") in 1854–58 at the University of Göttingen, where, in his lectures, he introduced, probably for the first time, the Galois (*q.v.*) theory of equations and attended the lectures of the mathematician Peter Gustav Lejeune Dirichlet. These experiences led Dedekind to see the need for a redefinition of irrational numbers in terms of arithmetic properties. By contrast, the geometric approach had led Eudoxus in the 4th century BC to define them as approximations by rational numbers (*e.g.*, a series of non-repeating decimals, as $\sqrt{2} = 1.414213...$).

In 1858 Dedekind joined the faculty of the Zürich Polyteknikum, where he remained for five years. In 1862 he accepted a position in the Technische Hochschule in Brunswick, where he remained in comparative isolation for the rest of his life.

While teaching there, Dedekind developed the idea that both rational and irrational numbers could form a continuum (with no gaps) of real numbers, provided that the real numbers have a one-to-one relationship with points on a line. He said that an irrational number would then be that boundary value that separates two specially constructed collections of rational numbers.

The Dedekind cut

Dedekind perceived that the character of the continuum need not depend on the quantity of points on a line segment (or continuum) but rather on how the line submits to being divided. His method, now called the Dedekind cut, consisted in separating all the real numbers in a series into two parts such that each real number in one part is less than every real number in the other. Such a cut, which corresponds to a given value, defines an irrational number if no largest or no smallest is present in either part; whereas a rational is defined as a cut in which one part contains a smallest or a largest. Dedekind would therefore define the square root of 2 as the unique number dividing the continuum into two collections of numbers, such that all the members of one collection are greater than those of the other; or, that cut, or division, separating a series of numbers into two parts such that one collection contains all the numbers whose squares are larger than 2, and the other contains all the numbers whose squares are less than 2.

Dedekind developed his arithmetical rendering of irrational numbers in 1872 in his *Stetigkeit und Irrationale Zahlen* (Eng. trans., "Continuity and Irrational Numbers," published in *Essays on the Theory of Numbers*, 1901). He also proposed, as did the German mathematician Georg Cantor (*q.v.*), two years later, that a set—a collection of objects or components—is infinite if its components may be arranged in a one-to-one relationship with the components of one of its subsets. By supplementing the geometrical method in analysis, Dedekind contributed substantially to the modern treatment of the infinitely large and the infinitely small.

While vacationing in Interlaken, Switzerland, in 1874, Dedekind met Cantor. Dedekind gave a sympathetic hearing to an exposition of the revolutionary idea of sets that Cantor had just published, which later became prominent in the teaching of modern mathematics. Because both mathematicians were developing highly original concepts, such as in number theory and analysis, which were not readily accepted by their contemporaries, and because both lacked adequate professional recognition, a lasting friendship developed.

Dedekind's sets and "ideals"

Continuing his investigations into the properties and relationships of integers—that is, the idea of number—Dedekind published *Über die Theorie der ganzen algebraischen Zahlen* (1879; "On the Theory of Algebraic Numbers"). There he proposed the "ideal" as a collection of numbers that may be separated out of a larger collection, composed of algebraic integers that satisfy polynomial equations with ordinary integers as coefficients (see ALGEBRAIC STRUCTURES). The ideal is a collection of all algebraic integer multiples of a given algebraic integer. For example, the notation (2) represents such a particular collection, as $\ldots -8, -6, -4, -2, 0, 2, 4, 6, 8 \ldots$. The sum of two ideals is an ideal that is composed of all the sums of all their individual members. The product of two ideals is similarly defined. Ideals, considered as integers, can then be added, multiplied, and hence factored. By means of this theory of ideals, he allowed the process of unique factorization—that is, expressing a number as the product of only one set of primes, or 1 and itself—to be applied to many algebraic structures that hitherto had eluded analysis.

Dedekind died at Brunswick on February 12, 1916.

BIBLIOGRAPHY. There is no full-length biographical treatment of Dedekind and his work. See ERIC T. BELL, *Men of Mathematics*, ch. 27 (1937, reprinted 1961); and CARL B. BOYER, *A History of Mathematics* (1968). Recent editions of Dedekind's publications include his *Gesammelte Mathematische Werke*, 3 vol. 1930–32, reprinted 1969), his collected mathematical writings; and *Essays on the Theory of Numbers* (1901, reprinted 1963).

(Ed.)

Defoe, Daniel

Daniel Defoe, English novelist, journalist, and pamphleteer, is best known as the author of *Robinson Crusoe*, one of the most famous books ever written. A man of many talents and author of an extraordinary range and number of works, he has been called the father of the English novel and of modern journalism, while his political poem *The True-Born Englishman*, which appeared in 1701, is said to have been the most widely sold poem that had ever been published in English up to that time. Nevertheless, Defoe remains in many ways an enigmatic figure. A man who made many enemies, he has been accused of double-dealing, of dishonest or equivocal conduct, of venality. Certainly in politics he served in turn both Tory and Whig; he acted as a secret agent for the Tories and later served the Whigs by "infiltrating" extremist Tory journals and toning them down. But Defoe always claimed that the end justified the means, and a more sympathetic view may see him as what he always professed to be, an unswerving champion of moderation, skillfully helping those whom he advised and supported to steer a middle course. In an age of violent partisanship, he sought by his writings and his actions to modulate the bitterness of party and religious strife. At the age of 59 Defoe embarked on what was virtually a new career, producing in *Robinson Crusoe* the first of a remarkable series of novels and other fictional writings that took up much of his abundant energy over the next five years. In so doing he was able to bring together his brilliant gifts as a journalist and the accumulated experience of a lifetime of busy activity to create works of an extraordinary and vivid power.



Defoe, engraving by M. Van der Gucht (1660–1725) after a portrait by J. Taverner, first half of the 18th century.
By courtesy of the National Portrait Gallery, London

Nonconformist background

Two important circumstances helped to shape Defoe's life both as man and writer: his Nonconformist background and the fact that, from his early 30s, he was never quite clear of the shadow of debt. That he did not conform to the Church of England meant that he was always something of an "outsider," not part of the polite world of 18th-century literature; the satirist Jonathan Swift referred to him contemptuously as "that fellow who was pilloried" and affected to forget his name, and Alexander Pope, though privately able to praise Defoe's achievement as a writer, nevertheless found a place for him in his satiric poem *The Dunciad*. The second circumstance—the existence of unsatisfied creditors—meant that his enemies always had a weapon they could use against him, while his political masters could use his need for protection to keep him toeing the line. Defoe's last years were clouded by legal controversies over al-

legedly unpaid bonds dating back a generation, and it is thought that he died in hiding from his creditors. His character Moll Flanders, born in Newgate Prison, speaks of poverty as "a frightful spectre," and it is a theme of many of his books.

Early life. Defoe was born in London, probably in the latter half of 1660. His father, James Foe, was a hard-working and fairly prosperous tallow chandler (perhaps also, later, a butcher), of Flemish descent. By his middle 30s, Daniel was calling himself "Defoe," probably reviving a variant of what may have been the original family name. As a Nonconformist, or Dissenter, Foe could not send his son to Oxford or Cambridge; he sent him instead to the excellent academy at Newington Green kept by the Rev. Charles Morton. Here he received an education in many ways better, and certainly broader, than any he would have had at an English university. Morton was an admirable teacher, later becoming first vice president of Harvard College; and the clarity, simplicity, and ease of his style of writing—together with the Bible, the works of John Bunyan, and the pulpit oratory of the day—may have helped to form Defoe's own literary style.

Ventures in trade and bankruptcy

Although intended for the Presbyterian ministry, Defoe decided against this and by 1683 had set up as a merchant. He called trade his "beloved subject," and it was one of the abiding interests of his life. He dealt in many commodities, travelled widely at home and abroad, and became an acute and intelligent economic theorist, in many respects ahead of his time; but misfortune, in one form or another, dogged him continually. He wrote of himself:

> "No man has tasted differing fortunes more,
> And thirteen times I have been rich and poor."

It was true enough. In 1692, after prospering for a while, Defoe went bankrupt for £17,000. Opinions differ as to the cause of his collapse: on his own admission, Defoe was apt to indulge in rash speculations and projects; he may not always have been completely scrupulous, and he later owned himself as one of those tradesmen who had "done things which their own principles condemned, which they are not ashamed to blush for." But undoubtedly the main reason for his bankruptcy was the loss he sustained in insuring ships during the war with France—he was one of 19 "merchants insurers" ruined in 1692. In this matter Defoe may have been incautious, but he was not dishonorable, and he dealt fairly with his creditors (some of whom pursued him savagely) paying off all but £5,000 within ten years. He suffered further severe losses in 1703, when his prosperous brick-and-tile works near Tilbury failed during his imprisonment for political offenses, and he did not actively engage in trade after this time.

Soon after setting up in business, in 1684, Defoe married Mary Tuffley, the daughter of a well-to-do Dissenting merchant. Not much is known about her, and he mentions her little in his writings, but she seems to have been a loyal, capable, and devoted wife. She bore him eight children, of whom six lived to maturity, and when Defoe died the couple had been married for 47 years.

Mature life and works. With Defoe's interest in trade went an interest in politics. The first of many political pamphlets by him appeared in 1683. When the Roman Catholic James II ascended the throne in 1685, Defoe—as a staunch Dissenter and with characteristic impetuosity—joined the ill-fated rebellion of the Duke of Monmouth, managing to escape after the disastrous Battle of Sedgemoor. Three years later James had fled to France, and Defoe rode to welcome the army of William of Orange—"*William*, the Glorious, Great, and Good, and Kind," as Defoe was to call him. Throughout William's reign, Defoe supported him loyally, becoming his leading pamphleteer. In 1701, in reply to attacks on the "foreign" king, Defoe published his vigorous and witty poem *The True-Born Englishman*, an enormously popular work that is still very readable and relevant in its exposure of the fallacies of racial prejudice.

Political pamphleteer

Foreign politics also engaged Defoe's attention. Since the Treaty of Rijswijk (1697), it had become increasingly probable that what would, in effect, be a world war would break out as soon as the childless king of Spain died. In 1701, five gentlemen of Kent presented a petition, demanding greater defense preparations, to the House of Commons (then Tory-controlled) and were illegally imprisoned. Next morning Defoe, "guarded with about 16 gentlemen of quality," presented the speaker, Robert Harley, with his famous document "Legion's Memorial," which reminded the Commons in outspoken terms that "Englishmen are no more to be slaves to Parliaments than to a King." It was effective: the Kentishmen were released, and Defoe was feted by the citizens of London. It had been a courageous gesture, and one of which Defoe was ever afterward proud; but it undoubtedly branded him in Tory eyes as a dangerous man who must be brought down.

What did bring him down, only a year or so later, and consequently led to a new phase in his career, was a religious question—though it is difficult to separate religion from politics in this period. Both Dissenters and "Low Churchmen" were mainly Whigs, and the "high-fliers"—the High-Church Tories—were determined to undermine this working alliance by stopping the practice of "occasional conformity" (by which Dissenters of flexible conscience could qualify for public office by occasionally taking the sacraments according to the established church). Pressure on the Dissenters increased when the Tories came to power, and violent attacks were made on them by such rabble-rousing extremists as Dr. Henry Sacheverell. In reply, Defoe wrote perhaps the most famous and skillful of all his pamphlets, "The Shortest-Way With The Dissenters" (1702). His method was ironic: to discredit the high-fliers by writing as if from their viewpoint but reducing their arguments to absurdity. The pamphlet had a huge sale, but the irony blew up in Defoe's face: Dissenters and High Churchmen alike took it seriously, and—though for different reasons—were furious when the hoax was exposed. Defoe was prosecuted for seditious libel and was arrested in May 1703. The advertisement offering a reward for his capture gives the only extant personal description of Defoe—an unflattering one, which annoyed him considerably: "a middle-size spare man, about 40 years old, of a brown complexion, and dark-brown coloured hair, but wears a wig, a hooked nose, a sharp chin, grey eyes, and a large mole near his mouth." Defoe was advised to plead guilty and rely on the court's mercy, but he received harsh treatment, and, in addition to being fined, was sentenced to stand three times in the pillory. It is likely that the prosecution was primarily political, an attempt to force him into betraying certain Whig leaders; but the attempt was evidently unsuccessful. Although miserably apprehensive of his punishment, Defoe had spirit enough, while awaiting his ordeal, to write the audacious "Hymn To The Pillory" (1703); and this helped to turn the occasion into something of a triumph, with the pillory garlanded, the mob drinking his health, and the poem on sale in the streets.

Prosecution and arrest

Triumph or not, Defoe was led back to Newgate, and there he remained while his Tilbury business collapsed and he became ever more desperately concerned for the welfare of his already numerous family. He appealed to Robert Harley, who, after many delays, finally secured his release—Harley's part of the bargain being to obtain Defoe's services as a pamphleteer and intelligence agent.

Defoe certainly served his masters with zeal and energy, travelling extensively, writing reports, minutes of advice, pamphlets. He paid several visits to Scotland, especially at the time of the Act of Union in 1707, keeping Harley closely in touch with public opinion. These trips bore fruit in a different way two decades later: in 1724–26 the three volumes of Defoe's admirable and informative *Tour Thro' the whole Island of Great Britain* were published, in preparing which he drew on many of his earlier observations.

Publication of *The Review*

Perhaps Defoe's most remarkable achievement during Queen Anne's reign, however, was his periodical, the *Review*. He wrote this serious, forceful, and long-lived paper practically single-handed from 1704 to 1713. At first a weekly, it became a thrice-weekly publication in

1705, and Defoe continued to produce it even when, for short periods in 1713, his political enemies managed to have him imprisoned again on various pretexts. It was, effectively, the main government organ, its political line corresponding with that of the moderate Tories (though Defoe sometimes took an independent stand); but in addition to politics as such, Defoe discussed current affairs in general, religion, trade, manners, morals, and so on, and his work undoubtedly had a considerable influence on the development of later essay periodicals (such as Richard Steele and Joseph Addison's *The Tatler* and *The Spectator*) and of the newspaper press.

Later life and works. With George I's accession (1714) the Tories fell; but the Whigs in their turn recognized Defoe's value, and he continued to write for the government of the day and to carry out intelligence work. At about this time, too (perhaps prompted by a severe illness), he wrote the best known and most popular of his many didactic works, *The Family Instructor* (1715). Not all the writings so far mentioned, however, would have procured literary immortality for Defoe; this he achieved when in 1719 he turned his talents to an extended work of prose fiction and (drawing partly on the memoirs of voyagers and castaways such as Alexander Selkirk) produced *Robinson Crusoe*. A German critic has called it a "world-book," a label justified not only by the enormous number of translations, imitations, and adaptations that have appeared but by the almost mythic power with which Defoe creates a hero and a situation with which every reader can in some sense identify himself.

Literary technique and style

Here (as in his works of the remarkable year 1722, which saw the publication of *Moll Flanders*, *A Journal of the Plague Year*, and *Colonel Jack*) Defoe displays his finest gift as a novelist—his insight into human nature. The men and women he writes about are all, it is true, placed in unusual circumstances; they are all, in one sense or another, solitaries; they all struggle, in their different ways, through a life that is a constant scene of jungle warfare; they all become, to some extent, obsessive. They are also ordinary human beings, however, and Defoe, writing always in the first person, enters into their minds and analyzes their motives. His novels are given verisimilitude by their matter-of-fact style and their vivid concreteness of detail; the latter may seem unselective, but it effectively helps to evoke a particular, circumscribed world. Their main defects are shapelessness, an overinsistent moralizing, occasional gaucheness, and naïveté. Defoe's range is narrow, but within that range he is a novelist of considerable power, and his plain, direct style, as in almost all of his writing, holds the reader's interest.

In 1724 he published his last major work of fiction, *Roxana*, though in the closing years of his life, despite failing health, he remained active and enterprising as a writer. He died on April 24, 1731.

MAJOR WORKS

An Essay upon Projects (1697; reissued as *Essays Upon Several Projects*, 1702); "An Argument Shewing, That a Standing Army, With Consent of Parliament, Is not Inconsistent with a Free Government" (1698); "The Two Great Questions Consider'd" (1700); *The True-Born Englishman; A Satyr* (1700, for 1701); "The Shortest-Way With The Dissenters: Or Proposals For The Establishment Of The Church" (1702); "A Hymn To The Pillory" (1703); "Giving Alms no Charity, And Employing the Poor A Grievance to the Nation" (1704); "A True Relation of the Apparition of one Mrs. Veal" (1706); *The History of the Union of Great Britain* (1709); "Memoirs of the Life and Eminent Conduct Of that Learned and Reverend Divine, Daniel Williams, D.D." (1718); *The Life and Strange Surprizing Adventures of Robinson Crusoe, of York, Mariner. Written by Himself* (1719); *The Farther Adventures of Robinson Crusoe; Being the Second and Last Part of his Life* (1719); *The History Of The Life and Adventures Of Mr. Duncan Campbell, A Gentleman, who, tho' Deaf and Dumb, writes down any Stranger's Name at first Sight; with their future Contingencies of Fortune* (1720); *Memoirs Of A Cavalier* (1720); *The Life, Adventures, and Pyracies, Of the Famous Captain Singleton* (1720); *Serious Reflections During The Life and Surprising Adventures of Robinson Crusoe: With His Vision Of The Angelick World* (1720); *The Fortunes And Misfortunes Of the Famous Moll Flanders* (1721, for 1722); *A Journal of the Plague Year;* (1722); *An Impartial History Of The Life and Actions of Peter Alexowitz, The Present Czar of Muscovy* (1723, for 1722); *The History and Remarkable Life of . . . Col. Jacque, commonly Call'd Col. Jack* (1723, for 1722); *The Fortunate Mistress: Or, A History Of The Life And Vast Variety Of Fortunes Of Mademoiselle de Beleau . . . Being the Person known by the Name of the Lady Roxana, in the Time of King Charles II* (1724); *A Tour Thro' the whole Island of Great Britain*, 3 vol. (1724, 1725, 1726, dated 1727); *A Narrative of all the Robberies, Escapes, etc. of John Sheppard* (1724); *A New Voyage Round The World, By A Course never sailed before* (1725, for 1724); "The True and Genuine Account of the Life and Actions of the Late Jonathan Wild" (1725); *The Complete English Tradesman* (1726, for 1725); *The Political History Of the Devil, As Well Ancient as Modern* (1726); *The Four Years Voyages of Capt. George Roberts* (1726); *Mere Nature Delineated: Or, A Body without a soul . . . also a brief dissertation upon the usefulness and necessity of fools, whether political or natural* (1726); "Some Considerations upon Street-Walkers. With A Proposal for Lessening the Present Number of them" (1726); "Augusta Triumphans: Or, The Way To Make London The most flourishing City in the Universe" (1728); *A Plan Of The English Commerce* (1728); *The Memoirs of an English Officer . . . Capt. George Carleton* (1728); *The Compleat English Gentleman* (published posthumously, 1890).

BIBLIOGRAPHY. The fullest bibliography is J.R. MOORE, *A Checklist of the Writings of Daniel Defoe* (1960, rev. 1962). The greatest single collection of Defoe material is in the British Museum. Other important collections include the Indiana University Library, the Huntington Library (San Marino, California), the Clark Memorial Library (Los Angeles), the Bodleian Library (Oxford), the Cambridge University Library, and the Boston Public Library. No collected edition contains more than a very small portion of Defoe's works; the best general editions are *Novels and Selected Writings*, 14 vol. (Shakespeare Head edition, 1927); and *Romances and Narratives*, ed. by G.A. AITKEN, 16 vol. (1895). The *Tour* was edited by G.D.H. COLE, 2 vol. (1927; rev. and enlarged by D.C. BROWNING, 1962); and the *Review* by A.W. SECORD, 22 vol. (1938). *The Best of Defoe's Review*, ed. by W.L. PAYNE (1951), is a good selection. The *Letters* have been edited by G.H. HEALEY (1955).

The most up-to-date and fully-documented biography is J.R. MOORE, *Daniel Defoe, Citizen of the Modern World* (1958); also recommended are JAMES SUTHERLAND, *Defoe*, rev. ed. (1950); and the pioneering study by WILLIAM LEE, *Daniel Defoe: His Life, and Recently Discovered Writings*, 3 vol. (1869). Earlier lives include those of GEORGE CHALMERS (1786) and WALTER WILSON (1830).

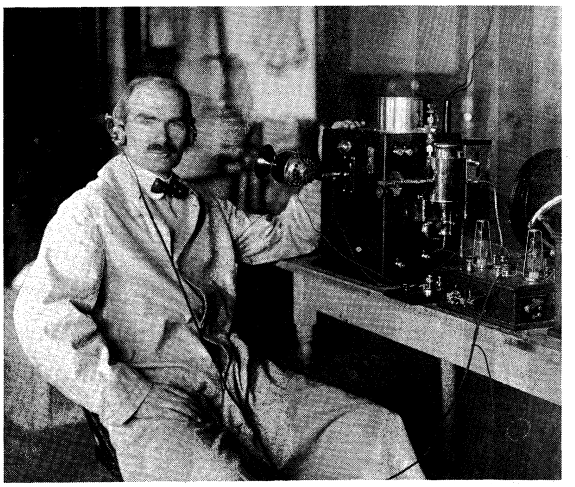
Critical studies include: PAUL DOTTIN, *Daniel De Foe et ses romans* (1924; Eng. trans. *Daniel Defoe and His Novels*, 1929); A.W. SECORD, *Studies in the Narrative Method of Defoe* (1924); IAN WATT in *The Rise of the Novel* (1957); M.E. NOVAK, *Economics and the Fiction of Daniel Defoe* (1962) and *Defoe and the Nature of Man* (1963); G.A. STARR, *Defoe and Spiritual Biography* (1965); and M. SHINAGEL, *Daniel Defoe and Middle-Class Gentility* (1968). There are numerous studies of individual novels, particularly *Robinson Crusoe* and *Moll Flanders*.

(R.P.C.M.)

De Forest, Lee

By his invention of the Audion vacuum tube and his promotion of wireless communication, Lee De Forest, prominent American inventor, substantially altered the character of 20th-century life. Among his 200 patents, secured during a professional career of more than 50 years, the Audion tube—the elementary form of the modern radio tube—exerted the most profound technological, sociological, and cultural effects. Until the invention of the transistor in 1947, followed by the development of solid-state electronics in the 1950s, the Audion remained the key component of all sophisticated radio, telephone, radar, television, and computer systems.

Lee De Forest was born on August 26, 1873, in Council Bluffs, Iowa. When he was six years old, his father, a Congregational minister, moved the family to Alabama, where he assumed the presidency of the nearly bankrupt Talladega College for Negroes. Ostracized by citizens of the white community who resented his father's efforts to educate the Negro, Lee made his friends from among the black children of the town and, together with his brother and sister, spent a happy although sternly disciplined childhood in this rural community.



De Forest, 1907.
Culver Pictures

As a child he was fascinated with machinery and was often excited when hearing of the many technological advances during the late 19th century. By the age of 13 he was an enthusiastic inventor of mechanical gadgets such as a miniature blast furnace and locomotive, and a working silverplating apparatus.

Early interest in radio

His father had planned for him a career in the clergy, but Lee insisted on science and, in 1893, enrolled at the Sheffield Scientific School of Yale University, one of the few institutions in the United States then offering a first-class scientific education. Frugal and hardworking, he supplemented his scholarship and the slim allowance provided by his parents by working at menial jobs during his college years, and, despite a not too distinguished undergraduate career, he went on to earn the Ph.D. in physics in 1899. By this time he had become interested in electricity, particularly the study of electromagnetic wave propagation, then being pioneered chiefly by the German Heinrich Rudolf Hertz and the Italian Guglielmo Marconi. De Forest's doctoral dissertation on the "Reflection of Hertzian Waves from the Ends of Parallel Wires" was possibly the first doctoral thesis in the United States on the subject that was later to become known as radio.

His first job was with the Western Electric Company in Chicago, where, beginning in the dynamo department, he worked his way up to the telephone section and then to the experimental laboratory. While working after hours on his own, he developed an electrolytic detector of Hertzian waves similar to one that had already been patented by the Englishman Sir John Ambrose Fleming. The device was modestly successful, as was an alternating-current transmitter that he designed. In 1902 he and his financial backers founded the De Forest Wireless Telegraph Company. To dramatize the potential of this new medium of communication, he began, as early as 1902, public demonstrations of wireless telegraphy for businessmen, the press, and the military.

The "Audion"

A poor businessman and a poorer judge of men, De Forest was defrauded twice by his own business partners. By 1906 his first company was insolvent, and he had been squeezed out of its operation. But in 1907 he patented a much more promising detector, which he called the "Audion"; it was capable of more sensitive reception of wireless signals than were the electrolytic and Carborundum types then in use. It was a thermionic grid-triode vacuum tube—a three-element electronic "valve" similar to a two-element device patented by John Ambrose Fleming in 1904. The same year De Forest was able to broadcast experimentally both speech and music to the general public in the New York City area.

A second company, the De Forest Radio Telephone Company, began to collapse in 1909, again because of some of his partners. In the succeeding legal confusion, De Forest was indicted in 1912 but later acquitted of federal charges of using the mails to defraud by seeking to promote a "worthless device"—the Audion tube.

In 1910 he broadcast a live performance by Enrico Caruso at the Metropolitan Opera in order to further popularize the new medium. In 1912 De Forest conceived the idea of "cascading" a series of Audion tubes so as to amplify high-frequency radio signals far beyond what could be accomplished by merely increasing the voltage on a single tube. He fed the output from the plate of one tube, through a transformer to the grid of a second, and the output of the second tube's plate to the grid of a third, and so forth, thereby allowing for an enormous amplification of a signal that was originally very weak. This was an essential development for both radio and telephonic long-distance communication. He also discovered in 1912 that by feeding part of the output of his triode vacuum tube back into its grid, he could cause a self-regenerating oscillation in the circuit. The signal from this circuit, when fed to an antenna system, was far more powerful and effective than that of the crude transmitters then generally employed and, when properly modulated, was capable of transmitting speech and music. When appropriately modified, this single invention was capable of either transmitting, receiving, or amplifying radio signals.

Throughout De Forest's lifetime the originality of his more important inventions was hotly contested, by both scientists and patent attorneys. In time, realizing that he could not succeed in business or manufacturing, he reluctantly sold his patents to major communications firms for commercial development. Some of the most important of these sales were made at very low prices to the American Telephone & Telegraph Company, which used the Audion as an essential amplification component for long-distance repeater circuits.

Further inventions

In 1920 he began to work on a practical system for recording and reproducing sound motion pictures. He developed a sound-on-film optical-recording system called phonofilm and demonstrated it in theatres between 1923 and 1927. Although basically correct in principle, its operating quality was poor, and he found himself unable to interest film producers in its possibilities. Ironically, within a few years' time the motion-picture industry converted to talking pictures by using a sound-on-film process similar to that of De Forest.

During the 1930s De Forest developed Audion-diathermy machines for medical applications and, during World War II, conducted military research for Bell Telephone Laboratories. Although bitter over the financial exploitation of his inventions by others, he was widely honoured as the "father of radio" and the "grandfather of television." He was supported strongly but unsuccessfully for the Nobel Prize for Physics.

Although he worked for many organizations during his lifetime, De Forest was fundamentally an individualist and produced most of his inventions as a free-lance worker. A complex and private kind of man, De Forest was assailed by self-doubts, indecision, and egocentricity. Divorced twice, his third marriage, to Marie Mosquini, was a happy one and endured from 1930 until his death, in Hollywood, California, on June 30, 1961.

BIBLIOGRAPHY. GEORGETTE CARNEAL, *A Conqueror of Space* (1930), an authorized biography that is uncritically laudatory in its appraisal of the inventor's work; LEE DE FOREST, *Father of Radio* (1950), a romantically written autobiography of his life and achievements, psychologically revealing but historically biassed; ISRAEL E. LEVINE, *Electronics Pioneer: Lee de Forest* (1964), a simple, popularly-written biography; W. RUPERT MACLAURIN, *Invention and Innovation in the Radio Industry* (1949), an authoritative, economically-oriented survey of major technological contributions to the radio industry.

(R.E.Fi.)

Deformation and Flow

Rheology, from the Greek word *rhein* meaning "to flow," is the science of deformation and flow. Deformation is the motion of one part of the body relative to another part so that the body changes size or shape; *i.e.*, a change of distances occurs between points in the body; and flow is a continuous change of deformation with time. The section of mechanics that deals with such motions is called continuum

mechanics or the mechanics of deformable media. Many examples of deformations are observed in everyday experience: the motion of a liquid being poured, of water in a flowing river, of a flag fluttering in the wind, of water boiling, of a rubber band being stretched, and of a violin string when it is plucked.

Applications in technology

Continuum mechanics is widely applied in technology. Structures must be designed and construction materials chosen so that neither fracture nor too much deformation will occur in use. In airplane design, the force of the air on the wings must be calculated for maximum lifting power and efficiency. Paints must be formulated so that they will flow from the brush but will not drip afterward. Pipelines must be designed with appropriate diameters so that fluids can be pumped at optimum rates.

The function of many body fluids depends on their flow characteristics. The viscosity of blood plasma increases when the organism is diseased. The synovial fluid from healthy joints has more complicated flow properties than those from rheumatoid joints.

When forces are applied to some materials they assume a deformed shape in equilibrium and return to their original shapes after the forces are removed. Such materials are called solids. Other materials can maintain an equilibrium shape only when subjected to hydrostatic pressures (the force is perpendicular to the surface). These materials, both liquid and gaseous, are known as fluids. Under other types of forces, they deform indefinitely as long as the forces are applied and do not return to their original form when these forces are removed. Such an irrecoverable deformation is called flow.

Rheology and constitutive equations

An important part of continuum mechanics is the formulation of constitutive equations—that is, relations between force and deformation—that describe the mechanical behaviour of a given material. Rheology is the part of continuum mechanics that deals with constitutive equations. Such equations provide an approximate description of behaviour over a certain range of circumstances. For example, a metal under a light load might be considered rigid; under a heavier load or with more accurate measurement of length, a linearly elastic solid; under a very large load, a plastic solid; and under a small oscillatory motion, a linearly viscoelastic solid. These three types of behaviour will be discussed later.

Two of the basic concepts of continuum mechanics are those of strain—the measure of the amount of deformation—and of stress, in units of force per unit area, as a measure of the contact force; that is, the force exerted by one part of the body on a neighbouring part.

HISTORY

Fracture, or breaking, was the chief subject in early work on the mechanics of solids. The notebooks of Leonardo da Vinci (1452–1519) contain a discussion of the strength of columns and beams and a sketch of a device for measuring the tensile strength of a wire to which a basket was attached, using sand as a variable load. Galileo Galilei noted in 1638 that the strength of a bar in simple tension is independent of its length and is proportional to its cross-sectional area; he was, thus, apparently the first to describe the property of a material in terms of a force per unit area.

An English scientist, Robert Hooke, in 1678, on the basis of a number of different experiments involving a wide variety of materials, concluded that the elongation of elastic materials is proportional to the force applied.

In a mathematical treatment of the shape of a loaded elastic bar, a Swiss mathematician, James Bernoulli, in 1705 was the first to state that the elastic law should be formulated as a relation between the force per unit area and the ratio of elongation to the original length. Although another Swiss mathematician, Leonhard Euler, was the first to assume in 1727 that this relation was linear, the proportionality constant is called Young's modulus, after Thomas Young, an English physicist who, in 1804, introduced it in connection with the elongation of an elastic bar. Euler also, in 1752, explicitly set forth the concept of the internal pressure in a fluid. Internal tension in solids (force acting perpendicular to the cross section) was recognized and used by Gottfried Wilhelm Leibniz, the German mathematician and physicist, in 1684, and by James Bernoulli in 1691; and shear stresses (forces acting parallel to the cross sections) by the French mathematician Antoine Parent in 1713 and explicitly by the French physicist Charles-Augustin Coulomb in 1773. The general concept now current, that contact forces may be represented by the stress, the introduction of the stress tensor (see below), the differential equations representing the balance of force, and the application of the balance of moment of momentum were introduced by the French mathematician Augustin-Louis Cauchy in 1822. Balance of force had been recognized by another member of the famous Bernoulli family, Daniel Bernoulli, in 1727, and both balance laws by Euler in 1775.

Introduction of stress tensor

Euler in 1775 proposed the first constitutive equation, that for a perfect fluid, based on the assumption that the force exerted by one part of the fluid on the neighbouring part is always perpendicular to the surface separating the two parts. The linearly viscous fluid is sometimes called the Newtonian fluid because Isaac Newton in his *Principia* (1687) proposed a simple case of such a material. The equations governing the motion of this fluid are called the Navier–Stokes equations after the French engineer Claude-Louis-Marie Navier (1821) and the British physicist George Gabriel Stokes (1845). Two Frenchmen, Siméon-Denis Poisson (1830) and Barré de Saint-Venant (1843), also made important contributions to the theory. The rate of efflux through narrow tubes, calculated on the basis of the Navier–Stokes equations, was found to be in good agreement with the experimental results on water obtained during the next decade or so.

Navier, on the basis of a molecular hypothesis, arrived in 1827 at a theory of elasticity of isotropic solids (solids in which elasticity is uniform in all directions) that contained only one elastic constant. Using some phenomenological concepts, Cauchy (1823), Stokes (1845), and their contemporaries produced the modern theory that involves two elastic constants, the shear modulus and bulk modulus, to be discussed later. Careful experiments confirming the modern theory were performed shortly thereafter.

In 1835 Wilhelm Weber, a German physicist, found that a load suddenly applied to a silk thread produced an immediate extension that was followed by a further lengthening with time, a behaviour called creep. Further experimental investigations of this time-dependent behaviour were performed; and in 1874 Ludwig Boltzmann, an Austrian physicist, formulated the theory of linear viscoelasticity to explain the phenomenon. That some fluids show a similar viscoelastic behaviour was exhibited on fluids of high consistency (hot glass, pitch) in 1905 and on more mobile fluids (gelatin solutions) in 1915.

Experiments around the beginning of the 20th century led to the conclusion that colloidal solutions (solutions in which one component consists of particles invisible under ordinary microscopes but thousands of times larger than ordinary molecules) do not act like Newtonian fluids in steady flow, but that their viscosity depends on the rate of deformation. The method of calculating this viscosity function from efflux experiments in the '20s greatly advanced this field of study.

The transition from the empirical approach to the development of appropriate constitutive equations has gone through several stages. Nonlinear theories proposed in the 1940s and 1950s have been superseded by theories that assume that stress depends upon the history of the strain that has been applied but not in a simple linear relationship. These latter theories explain not only nonconstant viscosity in steady flow but also time-dependent behaviour and normal stress effects, unexpected distributions of normal tractions on bounding surfaces in many experiments. Experimental work on this latter phenomenon began during World War II and is still continuing.

CONTINUUM MECHANICS

The study of mechanics falls in general into two parts: kinematics, which deals with the motion of particles and bodies, and dynamics, which is concerned with forces and

how they effect the kinematics. In continuum mechanics, which is used to give theoretical insight into deformation and flow, the kinematics deals mainly with motions that are deformations; the dynamics, with the stress; that is, with the forces exerted on one part of the body by a neighbouring part. The detailed discrete nature of the atomic and molecular structure is ignored; it is assumed that the net effect of the atomic and molecular forces and motions can be adequately represented by the stress, the strain or rate of strain, and an appropriate constitutive relation. For a thorough understanding of continuum mechanics the use of mathematics, particularly vector and tensor analysis, is required. In the following sections the basic concepts are first presented without the use of mathematics and then a more quantitative discussion is given, using vectors and tensors only on an elementary level.

Kinematics. Before presenting the general notion of strain, some simple deformations that are commonly discussed in many branches of continuum mechanics will be defined.

1. In uniform expansion, all volume elements of the body are transformed to geometrically similar elements of greater dimensions. Such a deformation occurs when a body is uniformly deformed in all directions so that any arbitrary particle, originally having the coordinates (x_1, x_2, x_3) moves to a point whose coordinates are (ax_1, ax_2, ax_3), in which a is a multiplicative factor (Figure 1). The displacement of the particle may be represented by the vector $\boldsymbol{u}$ whose components (u_1, u_2, u_3) are $[(a-1)x_i]$, $i = 1, 2, 3$. Any physical quantity that has direction, such as displacement, velocity, force, electric field, can be represented in magnitude and direction by a vector. The vector $\boldsymbol{u}$, for instance, can be represented by an arrow that points in the direction of the motion of the particle with the length representing the displacement or distance moved. Each component is the length of the axis intersected by lines drawn perpendicular to it from the head and tail of the vector. For example, the tail intersects the X_1-axis in the Cartesian coordinate system (X_1, X_2 and X_3 coordinates mutually perpendicular) at x_1 and the head intersects at ax_1, so that the displacement along the X_1-axis is a component u_1 of length $ax_1 - x_1$, or $(a-1)x_1$.

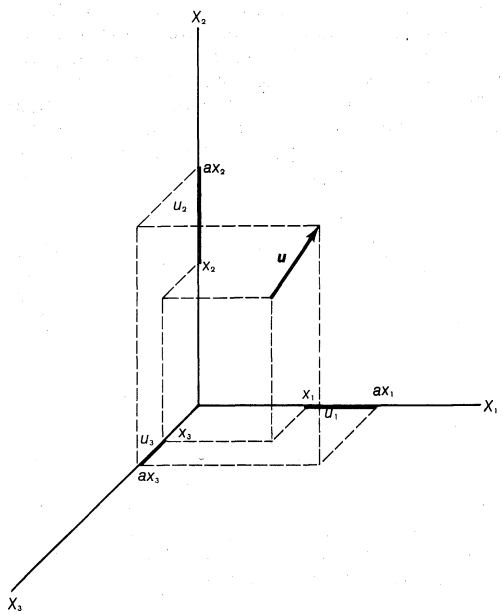


Figure 1: The vector $\boldsymbol{u}$ and its components in a Cartesian coordinate system X_1, X_2, X_3 (see text).

Vector notation

2. Simple shear is usually envisaged as taking place in a material between large parallel plates and may be thought of as sheets of material sliding parallel to one another. Simple shear can be described more precisely by taking a Cartesian coordinate system with the origin O on one of the plates, say the fixed plate, the X_2-axis perpendicular to the plates and the X_1-axis parallel to the direction of motion (Figure 2). In simple shear an arbitrary particle in the plane ABCD moves in the X_1-direction to a point in the plane ABC'D' by an amount proportional to its distance x_2 from the fixed plate, sx_2, in which s is a proportionality factor; *i.e.*, it undergoes a vector displacement with components $(sx_2, 0,0)$. In simple shearing flow, each particle of the material moves with a constant velocity $\boldsymbol{v}$ in the X_1-direction, the direction of motion of the sliding plate. The components of the velocity are $(\kappa x_2, 0,0)$ in which κ is a constant called the rate of shear.

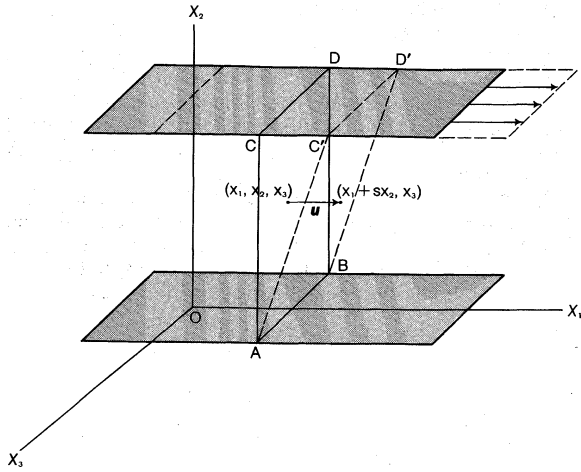


Figure 2: Simple shear in a Cartesian coordinate system, showing the vector displacement $\boldsymbol{u}$ of a particle from x_1 to $x_1 + sx_2$.

3. General infinitesimal strain. In more complicated deformations, the strain is more difficult to specify. The components (u_1, u_2, u_3) represent the displacement $\boldsymbol{u}$ of a particle in the body in the deformed state from its position (x_1, x_2, x_3) in the undeformed state. If the spatial derivatives (a spatial derivative is the change of a component with distance along any one of the three axes) of the displacement are small, strictly speaking, infinitesimally small, the complete description of the strain at a given point in the body may be given in terms of the six new components made up of these spatial derivatives:

Components of strain tensor

$$e_{11} = \partial u_1/\partial x_1;\ e_{22} = \partial u_2/\partial x_2;\ e_{33} = \partial u_3/\partial x_3;$$
$$e_{21} = e_{12} = \tfrac{1}{2}(\partial u_1/\partial x_2 + \partial u_2/\partial x_1);$$
$$e_{13} = e_{31} = \tfrac{1}{2}(\partial u_1/\partial x_3 + \partial u_3/\partial x_1);$$
$$e_{23} = e_{32} = \tfrac{1}{2}\partial u_2/\partial x_3 + \partial u_3/\partial x_2);$$

or in general,

$$e_{ij} = e_{ji} = \tfrac{1}{2}(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$$

for $i = 1, 2, 3$ and $j = 1, 2, 3$. The components e_{11}, e_{22}, e_{33} in the above equations represent relative, but not actual, elongations parallel to the X_1, X_2, and X_3 axes respectively; the other components, called shear components, are measures of changes of angle. It is of great importance to the theory that these components are the Cartesian components of a tensor called the infinitesimal strain tensor, or sometimes simply the strain. As a result, the powerful theorems of tensors, a well-developed area of mathematics dealing with such objects, may be exploited in continuum mechanics.

Applying the foregoing definitions to the components (u_1, u_2, u_3) of the vector $\boldsymbol{u}$ for uniform expansion, one finds, if $(a-1)$ is small, that only three components are not zero (namely, $e_{11} = e_{22} = e_{33} = a - 1$, or $e_{ii} = a - 1$ for $i = 1, 2, 3$) and that the shear components are all zero. When the spatial derivative equations are applied to components $(sx_2, 0,0)$ of the vector $\boldsymbol{u}$ for simple shear with $s << 1$ (s vanishingly small), all components of strain vanish except $e_{12} = e_{21}$, which is equal to $s/2$.

4. For finite strains, the infinitesimal strain is no longer a valid measure of deformation (for example, a zero value does not necessarily mean that the material is undeformed). Many kinds of tensor relations are in current use, all of

which include terms that are quadratic (squared) in the spatial derivatives of the displacement.

The flow of fluids

5. The flow of fluids is discussed in terms of the symmetric rate of deformation tensor **D**, with components defined as $D_{ij} = \frac{1}{2}(\partial v_i/\partial x_j + \partial v_j/\partial x_i) = D_{ji}$, for $i = 1, 2, 3$ and $j = 1, 2, 3$—that is $D_{11} = \partial v_1/\partial x_1$; $D_{22} = \partial v_2/\partial x_2$; $D_{33} = \partial v_3/\partial x_3$; $D_{12} = D_{21} = \frac{1}{2}(\partial v_1/\partial x_2 + \partial v_2/\partial x_1)$; $D_{13} = D_{31} = \frac{1}{2}(\partial v_3/\partial x_1 + \partial v_1/\partial x_3)$; $D_{23} = \frac{1}{2}(\partial v_2/\partial x_3 + \partial v_3/\partial x_2)$ in which v_1, v_2, and v_3 are the components of the velocity of an arbitrary particle. Applying these equations for the components of **D**, to the simple shearing flow mentioned earlier, in which $v_1 = \kappa x_2$, the only ones that do not vanish are the components $D_{21} = D_{12} = \kappa/2$.

Dynamics. In continuum mechanics it is assumed that the net effect of all forces between atoms may be considered the sum of two parts: body forces, due to gravity, magnetism, and electricity; and contact forces exerted on one part of the body by a neighbouring portion. In Figure 3,

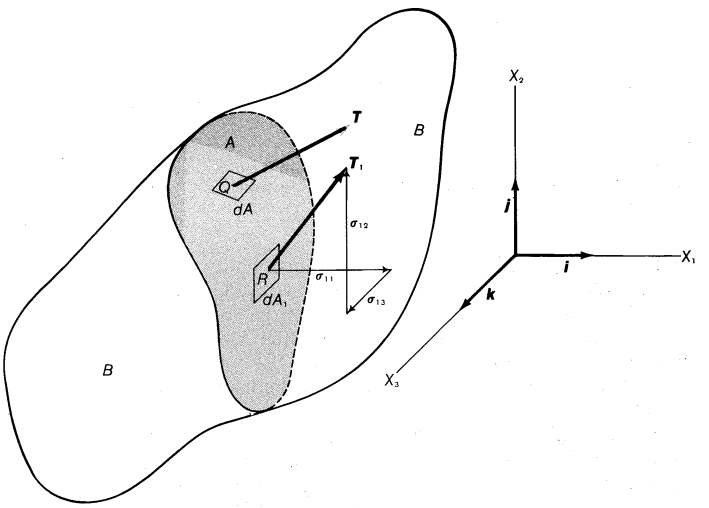


Figure 3: Stress vector $\boldsymbol{T}$ acting on an arbitrary surface A that divides the deformed body B into two parts. A vector component $\boldsymbol{T}_1$ can be decomposed along three coordinate directions, X_1, X_2, and X_3.

B is a body that is deformed and A is an arbitrary surface dividing B into two parts, with dA representing an infinitesimal portion of this area at the point Q. It is assumed that the force $d\boldsymbol{F}$ exerted by the matter on one side of the surface on that on the other is given by the equation $d\boldsymbol{F} = \boldsymbol{T}dA$, in which $\boldsymbol{T}$ is known as the stress vector. The existence of such a stress vector is one of the basic assumptions of continuum mechanics. Generally, $\boldsymbol{T}$ depends on both the position in the body and orientation of the surface dA at that point. The stress vector may be resolved into the vector sum of a normal (in this sense, perpendicular) stress that is at right angles to that surface and a shear stress parallel to that surface.

Stress vector

It is common to consider the stress vector across surfaces parallel to the orthogonal planes of a chosen Cartesian coordinate system (Figure 3). A plane normal or perpendicular to the X_1-direction is called an X_1-plane. At a given point R across such a plane, the stress vector $\boldsymbol{T}_1$ (not necessarily acting in the X_1-direction) may be decomposed into components (σ_{11}, σ_{12}, σ_{13}) in the coordinate directions; that is, the stress vector $\boldsymbol{T}_1 = \boldsymbol{i}\sigma_{11} + \boldsymbol{j}\sigma_{12} + \boldsymbol{k}\sigma_{13}$, in which $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are unit vectors (vectors of given directions and unit length) in the three coordinate directions. The first number in the subscript indicates the direction of the normal or perpendicular to the plane across which the contact force acts, and the second number indicates the direction of the component of the force. A similar set of three components (σ_{21}, σ_{22}, σ_{23}) may be obtained from the stress vector acting across the X_2-plane through R and (σ_{31}, σ_{32}, σ_{33}) acting through the X_3-plane. The stress components with repeated index (*i.e.*, σ_{11}, σ_{22}, σ_{33}) are the normal stress components; those with different indices (*e.g.*, σ_{12}) are the shear stresses, or σ_{ij} for $i = j$ represent normal stress components and for $i \neq j$ represent shear stresses.

It can be shown that the set of these nine components

$$\begin{matrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{matrix}$$

constitutes the components of a tensor, called the stress tensor, or simply the stress σ_{ij} for $i = 1, 2, 3, j = 1, 2, 3$, at the point with reference to the chosen Cartesian coordinate system. The stress tensor gives a complete characterization of the stress at a point. It can also be shown, as a consequence of the laws of mechanics, that the stress tensor is symmetric (*i.e.*, $\sigma_{12} = \sigma_{21}$, $\sigma_{13} = \sigma_{31}$, and $\sigma_{23} = \sigma_{32}$ or $\sigma_{ij} = \sigma_{ji}$ for $i = 1, 2, 3, j = 1, 2, 3$. Thus, only six components are required to specify completely the stress tensor.

Stress tensor

CONSTITUTIVE EQUATIONS

Perfect fluid. In many cases, the motion of fluids such as air and water can be described to a good approximation by the equations of the perfect fluid. The concept of perfect fluid, therefore, forms the basis of much of aerodynamics, hydrodynamics, and fluid mechanics.

By definition, no shear stresses can exist in the model of perfect fluid. As a result, the stress components are given by

$$\sigma_{ii} = \sigma_{11} = \sigma_{22} = \sigma_{33} = -p;$$

and for $i \neq j$,

$$\sigma_{ij} = \sigma_{12} = \sigma_{13} = \sigma_{23} = 0,$$

which are independent of the orientation of the Cartesian coordinate system. Such a stress is called an isotropic stress, and p, in the equation, is the hydrostatic pressure.

Linearly viscous fluid. The model of the linearly viscous (also called a Newtonian or a viscous) fluid describes the flow behaviour of fluids consisting of small molecules under all conditions in which the concept of a perfect fluid is not adequate (*e.g.*, near solid walls) except under extreme circumstances (*e.g.*, ultrasonic waves). It also describes the behaviour of more complex fluids, such as molten plastics at very low rates of deformation.

If a linearly viscous fluid is undergoing simple shearing flow (see definition in the section under *Kinematics*), the components of the stress with respect to the coordinate system defined above are

$$\begin{aligned} \sigma_{ii} = \sigma_{11} = \sigma_{22} = \sigma_{33} &= -p \\ \sigma_{12} &= \eta\kappa \\ \sigma_{13} = \sigma_{23} &= 0, \end{aligned}$$

in which η is a property known as the viscosity.

The general equation is based on the assumption that the stress is isotropic when the fluid is at rest, and is a linear function of the velocity gradients, or the partial derivatives of the velocity components with respect to the coordinates, $\partial v_1/\partial x_1$, $\partial v_1/\partial x_2$, etc., when it is in motion. Expressing these concepts mathematically in such a way that the constitutive equation is independent of the observer, it is found that the stress components may be expressed, for incompressible fluids, in terms of the rate of deformation tensor, D_{ij}, as:

$$\begin{aligned} \sigma_{11} &= -p + 2\eta D_{11}, \\ \sigma_{12} &= 2\eta D_{12}, \end{aligned}$$

for an arbitrary motion and arbitrary choice of Cartesian coordinate systems. The additional equations for σ_{22} and σ_{33} are analogous to that for σ_{11} and those for σ_{13} and σ_{23} to that for σ_{12}. The flow of a viscous fluid is an energy-dissipative process.

The viscosity of a fluid is commonly determined by measuring the rate of flow through a tube of small diameter when a pressure difference is applied. In the metric system the unit of η is called poise in honour of a French physiologist, J.-L.-M. Poiseuille. In the neighbourhood of 20° C, the viscosity of water is 0.01 poise, of air 1.8×10^{-4} poise, of mercury 0.016 poise, of pure glycerin 15 poise.

Linearly elastic solid. The linearly elastic (also called Hookean) solid provides a good model for the behaviour of many solids; *e.g.*, metals, salts, organic crystals, glasses, and rubber at very small strains under static conditions. The model becomes inadequate for some of these materials at elevated temperatures or for some dynamic experiments where linear viscoelasticity may prove a better description.

If an isotropic linearly elastic solid undergoes simple shear, the only nonzero components of the stress are $\sigma_{12} = \sigma_{21} = \mu$, where μ is a property of the material known as the shear modulus, or the rigidity.

If a solid obeying Hooke's law is subjected to a small uniform expansion [*i.e.*, $(a - 1) << 1$] the stress is isotropic:

$$\sigma_{11} = \sigma_{22} = \sigma_{33} = -p = 3K(a - 1),$$

in which $3(a - 1)$ is the fractional change of volume and K is known as the bulk modulus.

The general constitutive relation between stress and strain for the isotropic linearly elastic solid is the set of equations of the form

$$\sigma_{11} = \tfrac{1}{3}\mu e_{11} + (K - \tfrac{2}{3}\mu)(e_{11} + e_{22} + e_{33})$$

for the normal stresses, and of the form $\sigma_{12} = \tfrac{1}{2}\mu e_{12}$ for the shear stresses. Thus, two constants, the shear modulus μ and the bulk modulus K, specify the properties of an isotropic linearly elastic solid. If the material is anisotropic (not uniform in all directions), as for example a crystal, up to 21 constants may be required to specify its properties.

When a stress is applied to an elastic solid, it deforms; when the stress is removed, the solid returns to its original shape. The deformation of an elastic solid is an energy-storing process.

The shear modulus μ may be determined by measuring the torque required to twist a cylinder of the material. Sometimes an elastic modulus may be found from the speed of propagation of sound or from the natural frequency of oscillation of a mechanical system containing the material. The constants, μ and K, for steel have values of 8×10^{11} and 2×10^{12}, respectively; for lead 6×10^{10} and 4×10^{11}; for rubber 7×10^{6} and 2×10^{10}; all in units of dynes/cm^2.

Linear viscoelastic materials. This model provides a good description for the time-dependent behaviour of many materials under low stresses: for example, attenuation of ultrasonic waves in gases and liquids; oscillations, creep, and stress relaxation in metals, glass, rubber, plastics, and fluids containing long molecules (*e.g.* molten plastics and protein solutions).

If a shear stress T_0 is suddenly applied to a viscoelastic material in simple shear and kept constant thereafter, the deformation does not increase and attain its steady-state (unchanging with time) condition at once. This type of behaviour is known as creep. The variation of the shear strain $s(t)$ with time t is determined by the property creep compliance $J(t)$ and is represented by the equation: $s(t) = T_0 J(t)$. For a viscoelastic solid, the steady state is a constant strain; for the fluid, it is a constant rate of strain.

Creep behaviour

If a shear strain S_0 is suddenly imposed on a material in simple shear and kept constant thereafter, the shear stress $T(t)$ initially is very large and then decreases to a constant value, called a steady state value, which is zero for a fluid. Such an experiment is known as a stress–relaxation experiment. This decay of stress is determined by a property of the material, the relaxation modulus $G(t)$, according to the equation $T(t) = S_0 G(t)$.

The special feature of viscoelasticity is that the stress at a given time depends on all previous values of the strain, with more recent strains having greater influence.

Finite elasticity. There are materials, notably various types of rubber, that can undergo very large strains and return to their original shape after the force is removed. In the theory of isotropic nonlinear elasticity, which provides a good description of the equilibrium behaviour of these materials, the stress is a nonlinear function of the deformation.

Effects arise in nonlinear theories that do not occur in linear ones. According to linear elasticity theory, torsion of a circular cylinder requires only the application of a torque. The theory of finite elasticity demands, in addition, the imposition of normal forces on the plane ends of the cylinder; otherwise the cylinder will change in length.

Nonlinear viscoelastic fluids. High polymers, when in solution or a melted state, show many types of deviations from the behaviour predicted by the theory of Newtonian fluids. They exhibit time-dependent behaviour—for example, creep. In steady flow, the rate of efflux from a tube of small diameter is not proportional to the pressure drop; it appears that the viscosity decreases at the higher rates of flow. If such a fluid undergoes steady shear in the annulus between a stationary plate and a cone rotating at constant angular speed, a greater normal force (per unit area) is exerted by the fluid on the plate near the axis of rotation than at the outside. Based on the assumption that the stress is a nonlinear function of the previous history of the deformation, the theory of the nonlinear viscoelastic fluid is able to explain all of these phenomena.

Fluids exhibiting these phenomena show other technologically important deviations from Newtonian behaviour. The diameter of a stream of fluid may be considerably larger than the orifice from which it was extruded. A rotor in such a fluid may cause a completely different flow pattern from that seen in a Newtonian fluid.

Plasticity. When a rod composed of some solids, such as aluminum and some steels at ordinary temperature, is subjected to increasing tension, it behaves as a linear elastic solid. As extension increases, the force–elongation curve, as shown in Figure 4, deviates from proportionality at a

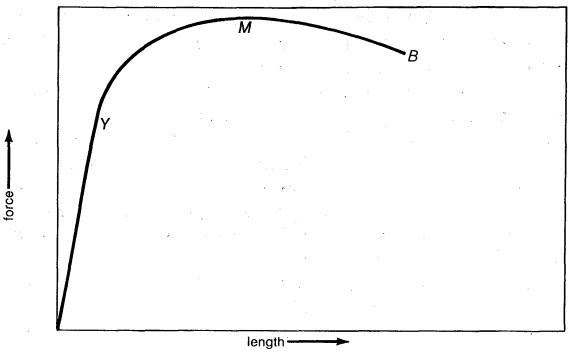


Figure 4: Force-elongation curve.

point Y when the tensile force exceeds a value, f_Y. The ratio f_Y/A_0, in which A_0 is the original cross section, is called the yield stress. The force then goes through a maximum at M before the sample finally breaks at point B. When the rod is extended beyond Y, some irrecoverable deformation, called plastic flow, takes place; the rod does not return to its original length when the force is removed.

Plastic flow

One of the most important considerations of the engineer is that structures must be designed and materials of construction chosen so that plastic deformation does not occur to any appreciable extent. On the other hand, many fabrication processes (for example, machining, forging, and drawing) depend on the behaviour in the plastic regime.

FRACTURE

It is customary to classify fracture of solids into two types: brittle and ductile. Generally speaking, brittle fracture occurs when the sample behaves as a Hookean elastic solid until a critical load is reached, at which time it breaks. In ductile fracture, plastic flow precedes fracture. A given material may change its mode of fracture with varying temperature or conditions of the deformation. Thus, many metals and polymers have a brittle–ductile transition temperature that is a function of the rate of straining.

If an increasing tensile force is applied to a rod of material until it breaks in a brittle manner, the measure of the strength of the material is the tensile strength, defined as ratio of the force at break to the original cross section of the rod. Glass, cast iron, stone, and plastics under ordinary conditions exhibit this type of fracture. If the fracture took place by the simultaneous rupture of all molecular bonds in the surface of separation, the tensile strength would be ten or 100 times greater than the generally observed values for the usual brittle materials. The discrepancy is due to microscopic sharp long cracks that exist in the material.

Tensile strength

Fracture can occur in a tensile experiment even if the stress always remains below the tensile strength. If the metal is subjected to alternating cycles of imposition and removal of stress, it may break if the number of cycles exceeds a critical number. Fatigue is the term used for this type of failure.

When a bar of some transparent glassy plastic (*e.g.*, polystyrene or polymethyl methacrylate) is subjected to a tensile force, microscopic, highly reflective, planar defects form in planes normal (at right angles) to the applied tension. These defects are called crazes, which consist of microscopic voids and of polymeric material oriented in the direction of the applied force.

MOLECULAR THEORIES

Gases. Newton's law of viscous flow has been derived from the kinetic theory of gases. If the gas is dilute and the molecules are considered to be rigid spheres that do not attract one another, the viscosity is found to be independent of pressure. A more detailed kinetic theory calculation gives the viscosity in terms of forces between molecules. Viscosity data are now used to estimate the forces between molecules. A rigorous kinetic theory developed in 1956 shows that a rarefied gas is not strictly a Newtonian fluid; normal stress effects are to be expected.

At ultrasonic frequencies, of the order of 10^6 cycles per second, viscoelastic effects are indicated by changes in the velocity of propagation of the sound waves and the attenuation of their amplitude. These effects have been attributed to distortions in the distribution of velocities of the gas molecules, and in the distribution of molecules among vibrational and rotational states.

Crystalline solids. Viscoelastic effects in crystalline materials usually have been attributed to imperfections in the crystal, such as the motion of interstitial solute atoms or the slip of one crystal grain past a neighbour along a grain boundary. Plastic behaviour is understood in terms of the motion of dislocations.

Liquids. On the molecular level, the liquid is pictured as having neighbouring molecules in contact, as does the crystalline solid, but not packed together in a systematic order; hence there are many "holes" in the structure. In the liquid at rest, a molecule can move from its location into a neighbouring hole over an energy barrier if its thermal energy is high enough and the hole large enough. A shear stress induces a preferred direction for such motions and thus leads to flow.

Polymer behaviour

Polymer molecules in the liquid state (as a melt, rubber, or solute) are loose coils that can change their conformations readily; a barrel full of long, excited snakes has been suggested as a useful analogy. Because the molecules must move through tortuous paths to become displaced relative to one another, the viscosity is high and sensitive to the length of the chain. Viscoelastic effects occur because it takes time for the long entwined molecules to adjust to the applied force. If the temperature is lowered to the point at which large enough holes are rare, the motion becomes so restricted that the polymer acts like a glassy solid.

If a reaction is performed on a polymer (as in the vulcanization of rubber) such that the polymer chains are connected by covalent bonds and therefore cannot move past one another, flow is impossible. Because the polymer chains on the average have a coiled conformation, however, a great extensibility of a rubber sample is possible since the molecule can, with little change of internal energy, assume a greatly extended conformation. The deformation of rubber chiefly involves a change of entropy—a change to a more ordered state of about the same energy—and deformation of crystals and glassy solids, a change of internal energy.

See also ELASTICITY; FLUID MECHANICS; RELAXATION PHENOMENA; TRANSPORT PHENOMENA; VIBRATIONS; SOLID STATE, THEORY OF; LIQUID STATE, THEORY OF; GLASSY STATE, THEORY OF; METALS, THEORY OF; POLYMERS; RUBBER; PLASTICS AND RESINS.

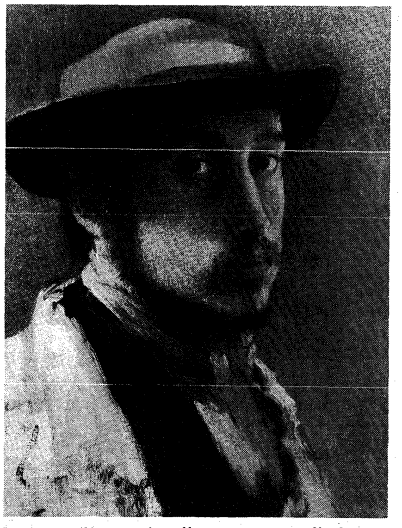
BIBLIOGRAPHY. F.R. EIRICH (ed.), *Rheology: Theory and Applications*, 5 vol. (1956–70), various areas covered by experts, articles differ greatly in style and level of presentation; A.H. COTTRELL, *The Mechanical Properties of Matter* (1964), an advanced undergraduate engineering text, with emphasis on solids; J.C. JAEGER, *Elasticity, Fracture, and Flow*, 2nd ed. (1962), many basic aspects, with emphasis on geological applications; C. TRUESDELL, *Elements of Continuum Mechanics* (1966), modern mathematical theory, *Essays in the History of Mechanics* (1968), a readable account of the development of some basic concepts, with copious references to original sources; H. ROUSE and S. INCE, *History of Hydraulics* (1957); S.P. TIMOSHENKO, *History of Strength of Materials* (1953); J.D. FERRY, *Viscoelastic Properties of Polymers*, 2nd ed. (1970), a discussion of properties, experimental methods, and molecular theory; K.F. HERZFELD and T.A. LITOVITZ, *Absorption and Dispersion of Ultrasonic Waves* (1959), an authoritative work, especially on gases and ordinary liquids; B.D. COLEMAN, H. MARKOVITZ, and W. NOLL, *Viscometric Flows of Non-Newtonian Fluids* (1966), a small monograph covering theory and experiments on nonlinear steady flows; A.C. ERINGEN, *Mechanics of Continua* (1967), modern developments; A.S. LODGE, *Elastic Liquids: An Introductory Vector Treatment of Finite Strain Polymer Rheology* (1964); B. ROSEN (ed.), *Fracture Processes in Polymeric Solids* (1964), many authoritative articles; L.R.G. TRELOAR, *The Physics of Rubber Elasticity*, 2nd ed. (1958), a discussion of molecular theory and phenomenological finite elasticity theory and experiment.

(He.M.)

Degas, Edgar

Edgar Degas was one of the greatest of 19th-century French artists and a master of the human figure in movement. Though Degas was a consistent exhibitor with the Impressionists, his position mainly as a painter of closely calculated indoor figure groups places him slightly apart from his colleagues. He was, however, very much concerned with depicting the fleeting and momentary, while his subjects were those of the school that followed Gustave Courbet, and his colours, moreover, became progressively more brilliant and divided, like those of the Impressionists.

Sterling and Francine Clark Art Institute, Williamstown, Massachusetts



Degas, self-portrait, oil on paper applied to canvas, *c.* 1857–58. In the Sterling and Francine Clark Art Institute, Williamstown, Massachusetts. 26 cm × 19 cm.

Early life and works. Hilaire-Germain-Edgar Degas was born in Paris on July 19, 1834. He came of the powerful upper bourgeoisie, his family having banking and business connections both in Italy and the United States, and he was intended for the law, which he studied for a time after leaving the Lycée Louis-le-Grand. In 1855, however, he enrolled at the École des Beaux-Arts and entered the studio of Louis Lamothe, a pupil of the painter Jean-Auguste-Dominique Ingres, whose long-established position as defender of academic orthodoxy in draftsmanship and subject matter was being challenged by the realism of Courbet as well as by the romanticism of Eugène Delacroix.

Early orthodoxy

It seems likely that as a young man Degas wished to succeed along orthodox lines as a painter of historical subjects in the grand French tradition. To further his aim he augmented his studies by visiting Florence, Assisi, Rome, and Naples and by closely observing and copying the

works of Andrea Mantegna, Sandro Botticelli, Hans Holbein the Younger, and Nicolas Poussin, all notable for their scrupulousness in figure draftsmanship. Before 1860, Degas had produced some splendid family portraits in which the effect of this discipline, though clear, is heightened by a taut, alert urbanity that belongs unmistakably to the mid-19th century. The "Portrait of the Duchess of Morbilli" is typical of this group. It is broadly designed with large, simple surfaces rather flatly modelled in the manner of Ingres; the paint is solid, yet delicate, and the colours cool and restrained, with many black and neutral passages. In 1860, Degas made his debut as a painter of classical subjects with his "Young Spartans Exercising"; but here the nude figures, though arranged in balanced groups, are those of real adolescents in a natural landscape instead of idealized nudes in an Arcadian setting.

After 1861, when Degas painted "Semiramis Founding Babylon," again with academic intentions, he seems to have abandoned historical painting and begun to seek his subject matter in the fast-moving city life of Paris. In this he was probably inspired by contemporaries like Courbet and Édouard Manet (whom he met in 1862), by contemporary novelists, and by the discovery, late in the 1850s, of the astonishing formal yet documentary quality of Japanese graphic art. Nor did he overlook the brilliant work of contemporary French graphic artists such as Paul Gavarni and Honoré Daumier. It is not surprising that by 1862 he was painting the riders, their mounts, and the smart spectators at Longchamp racecourse, soon afterward beginning the portrait groups of musicians and stage subjects, which, like all subjects in which the sitters were absorbed in practiced movement, fascinated him throughout his life. Among the first of the latter is the "Mlle Fiocre in the Ballet 'La Source' ". His portraits of the 1870s show greater ease and naturalism than the very first group but are still based on a firm discipline traceable right back to Holbein and the great north Italian portraitists.

Mature life and works. Degas served in the artillery during the Franco-German War of 1870–71. On his return, he began to undertake ambitious figure groups, seen informally and in movement, and continued his studies of stage and orchestral groups. From these he passed to instantaneous renderings of both outdoor and indoor scenes, using displaced figure grouping and unorthodox cutting and perspective rather in the manner of a cameraman. Yet his magnificent formal sense and skill is always present to provide an equilibrium, however momentary, to these exacting subjects. The "Place de la Concorde (Vicomte Lepic and His Daughters)" is a fine example and an outdoor counterpart to the ballet subjects that gave Degas endless scope for multifigure groups seen in fast intercepted movement. Degas visited the United States in October 1872, staying for five months and painting one of his best known scenes of figures in absorbed "occupational" movement. This was the "New Orleans Cotton Office" of 1873, now in the museum at Pau, France; it shows that although Degas had completely abandoned his early ambition of historical figure painting, he had nevertheless put to full use the structural principles of the formal tradition.

Development of figure groupings

During the 1870s, most of Degas's figure groups were arranged against fairly extended background space, in which the figures themselves were given plenty of room. By the end of the decade, however, he was becoming interested in the pictorial possibilities of more closely juxtaposed and superimposed groups and giving more attention to the formal qualities of the voids between them. The famous "Repasseuses" ("Two Laundresses") of 1884 shows this tendency at an advanced stage, with an artificially shallow picture space and a reconciliation of solid form and surface reminiscent of Venetian mid-16th-century art in method but here applied with a documentary eye to a casual workaday subject. By this time Degas had begun to work in pastel, sometimes using a mixed technique with volatile oil mediums, and his indoor series of women at their ablutions carries on the researches mentioned above. Some of the later ones reach an astonishing compromise between plasticity and surface pattern, the flesh colours being built up of strips of pure colour more closely knit than the *taches* of the Impressionists but like them, merging at a certain distance to give the illusion of solid modelling.

After 1880 Degas was practicing occasionally as a sculptor, and a group of small bronzes deriving from his models of dancers, bathing women, and horses again show his power of revealing the potentialities of the ordinary unobserved movements of human beings and animals. Degas, in fact, perhaps for the first time in history, viewed his animal and human models with the same dispassionate eye when making these studies. He was interested in photography, and there is an affinity between his vision and that of a high-speed camera. Degas's eyesight failed in later life; he became completely blind in one eye and nearly so in the other. He died September 27, 1917, leaving an important collection of the drawings and paintings of his contemporaries, and a notebook of poetic compositions, mostly in sonnet form.

Bronze sculptures

MAJOR WORKS

PAINTINGS: "Portrait of the Duchess of Morbilli" (1855–56; Louvre, Paris); "Achille de Gas in the Uniform of a Cadet" (1856–57; National Gallery of Art, Washington, D.C.); "Young Spartans Exercising" (1860; National Gallery, London); "Portrait of Mme Brunet" (1860; Charles S. Payson Collection, New York); "Semiramis Founding Babylon" (1861; Louvre); "The Bellelli Family" (1860–62; Louvre); "At the Races" (1860–62; Fogg Art Museum, Cambridge, Massachusetts); "Self-Portrait" (*c.* 1862; National Gallery of Art, Washington, D.C.); "Portrait of a Young Woman" (1867; Louvre); "Duke and Duchess of Morbilli" (1867; Museum of Fine Arts, Boston); "Mlle Fiocre in the Ballet 'La Source' " (*c.* 1867–68; Brooklyn Museum, New York); "Jacques-James Tissot" (1868; Metropolitan Museum of Art, New York); "Les Musiciens à L'orchestre" (1868–69; Louvre); "Madame Camus au piano" (1869; Alfonse Kann Collection, Saint-Germain-en-Laye); "Mlle Hortense Valpinçon" (1869; Minneapolis Institute of Arts); "The False Start" (1869–72; John Hay Whitney Collection, New York); "Chevaux de courses devant les tribunes" (1869–72; Louvre); "Madame Camus" (1869–70; National Gallery of Art, Washington, D.C.); "Carriage at the Races" (1870–73; Museum of Fine Arts, Boston); "Le Foyer de la danse à l'opéra de la rue le Pelletier (1872; Louvre); "The Dancing Class" (1872; Metropolitan Museum of Art, New York); "Mme René de Gas" (1872–73; National Gallery of Art, Washington, D.C.); "Foyer de Danse" (*c.* 1873; Corcoran Gallery of Art, Washington, D.C.); "New Orleans Cotton Office" (1873; Musée des Beaux-Arts, Pau, France); "Sulking (Bouderie)" (*c.* 1873–75; Metropolitan Museum of Art, New York); "Place de la Concorde (Vicomte Lepic and His Daughters)" (*c.* 1875; Gerstenberg Collection, Berlin); "Women Combing Their Hair" (1875–76; Phillips Collection, Washington, D.C.); "L'Absinthe" (1876; Louvre); "The Rehearsal" (1877; Glasgow Art Gallery); "Miss Lala at the Cirque Fernando" (1879; Tate Gallery, London); "Éventail" (1879; Louvre); "The Dancing Class" (1880; Denver Art Museum, Colorado); "Jockeys" (*c.* 1881–85; Yale University Art Gallery, New Haven, Connecticut); "Les Repasseuses" ("Two Laundresses," 1884; Louvre); "The Millinery Shop" (*c.* 1885; Art Institute of Chicago); "Ballet Scene" (1907; National Gallery of Art, Washington, D.C.).

PASTELS: "Woman Drying Her Foot (*c.* 1866; Metropolitan Museum of Art, New York); "L'Etoile" (*c.* 1878; William Coxe Wright Collection, Philadelphia); "The Rehearsal on the Stage" (1878–79; Metropolitan Museum of Art, New York); "Ballet at the Opera" (1878–80; Leigh B. Block Collection, Chicago); "At the Milliner's" (*c.* 1882; Museum of Modern Art, New York); "Blue Dancer" (*c.* 1883; Charles S. Payson Collection, New York); "The Dancers" (1899; Toledo Museum of Art, Ohio).

SCULPTURE: "A Dancer at the Age of Fourteen" (1880; Metropolitan Museum of Art, New York); "Dancer Looking at the Sole of Her Foot" (*c.* 1900; Metropolitan Museum of Art, New York).

BIBLIOGRAPHY. Books on Degas are extremely abundant. One hundred items are included in JOHN REWALD, *The History of Impressionim*, rev. ed. (1961). The essential work is PAUL A. LEMOISNE, *Degas et son oeuvre*, 4 vol. (1946–49). Other important studies are PIERRE CABANNE, *Degas: danseuses* (1960; Eng. trans., *Degas: Dancers*, 1963), a discerning text with excellent reproductions; DOUGLAS COOPER, *Pastels by Edgar Degas* (1954), an important contribution to the study of a medium developed to a high point by Degas; FRANCOIS

FOSCA, *Degas* (1953), a very good treatment of his stylistic development; EDOUARD HUTTINGER, *Degas* (1959); J.B. MANSON, *The Life and Work of Edgar Degas* (1927); CAMILLE MAUCLAIR, *Degas*, rev. and enl. ed. (1941), which includes quotations from Degas' letters; MARGUERITE REBATET, *Degas* (1944), a very well-written short study with good reproductions; and JOHN REWALD, *Degas, das plastische Werk*, 2nd ed. (1957; Eng. ed., *Degas: Works in Sculpture*, 1944), a valuable study, with 141 illustrations.

Testimony of his contemporaries: JACQUES EMILE BLANCHE, *Propos de peintre de David à Degas*, 3 vol., with a preface by MARCEL PROUST (1919–28), informative material by a fellow artist; MARCEL GUERIN (ed.), *Lettres de Degas*, new ed. (1945; Eng. trans., *Letters*, 1948), letters to his friends that reveal a man affable, tender, sometimes gay, and sympathetic to the sufferings of others; GEORGE MOORE, "Memories of Degas," *Burlington Magazine* (July–August, 1928); DENIS ROUART (ed.), *Correspondance de Berthe Morisot* (1950); PAUL VALERY, *Degas danse dessin* (1938; Eng. trans., *Degas Dance Drawing*, 1948), rich in considerations of art in general—Valery's ideas about literature were very similar to Degas's ideas on his art.

Catalogs: Catalogues des tableaux, pastels et dessins, par Edgar Degas, et provenant de son atelier (1918), catalogs of paintings, pastels, and drawings by Degas, originating in his studio (2,290 pieces); *Degas*, Museé de l'Orangerie, Paris (1937), an excellent analysis of Degas' style, with an introduction by PAUL JAMOT; *La Peintre au Museé du Louvre* by PAUL JAMOT (1929), an explanation of Degas' techniques of painting; *Le Peintre graveur illustré* (1919), discusses 45 major etchings and 20 lithographs; *Degas*, Wildenstein Galleries, New York (1960), a distinguished general catalog; *Drawings by Degas*, City Art Museum of St. Louis (1966), an important, recent study.

(D.C.Th.)

Dehydration

Dehydration is the loss of water from the body; it almost invariably is associated with varying kinds and degrees of disturbance of salt (sodium chloride) metabolism. Because of this, dehydration can be classified simply as follows: (1) loss of more water than sodium; (2) loss of more sodium than water; (3) isotonic (*i.e.*, proportionate) losses of sodium and water. In addition to the losses of sodium, there may be losses of other electrolytes such as potassium, as well as alterations of acid-base balance. The treatment of any form of dehydration, therefore, requires not only the replacement of the water lost from the body but also the restoration of the normal concentration and distribution of electrolytes within the body fluid.

Causes of dehydration. Dehydration may be caused by restriction of water intake or by excessive water loss.

Restriction of water intake. The commonest cause of dehydration is failure to drink. The deprivation of water is far more serious than the deprivation of food. Man loses approximately 2.5 percent of his total body water per day (about 1,200 millilitres, or 1.25 quarts) in urine, in expired air, by insensible perspiration, and from the gastrointestinal tract. If, in addition to this loss, his loss through perspiration is greatly increased—as in the case of the shipwrecked sailor in tropical seas or the traveller lost in the desert—in only a few hours he may lose so much body water that he goes into shock and dies.

When swallowing is difficult in extremely ill persons, or when people cannot respond to a sense of thirst because of age or illness or dulling of consciousness, the failure to compensate for the daily loss of body water will rapidly result in dehydration and its consequences.

Excessive water loss. Loss of water from the body invariably is associated with more or less loss of salt. In the following examples more water is lost than salt.

Excessive urination (diuresis) occurs when there is sugar in the urine as a result of diabetes mellitus, particularly when the blood and body fluids are abnormally low in alkalinity. It also occurs in seriously ill persons fed a high-protein diet, who cannot retain nitrogen and must excrete the urea formed by protein breakdown.

Water loss from diabetes

Persons with diabetes insipidus (which results from insufficiency of antidiuretic hormone) are unable to concentrate urine, and in consequence maintain the most phenomenal intake of water and excretion of urine. It has been said of these unfortunates that they "spend their lives running from the kitchen to the bathroom." The water that they lose in this fashion has an extremely low concentration of salt. A similar clinical picture can be seen in persons whose diseased kidneys are unable to reabsorb water.

In persons with extensive burns, a large volume of water may be lost through the damaged skin. Similarly, excessive sweating without adequate water intake can produce severe dehydration. A rare cause of dehydration is loss of water from the lungs in the presence of hyperventilation —excessive rate and depth of breathing.

The second major group of diseases in which dehydration occurs are those in which there is a loss of salt which is greater than the accompanying loss of water. Such a disease is adrenal cortical insufficiency, also called Addison's disease. The hormone aldosterone, secreted by the outer substance, or cortex, of the adrenal gland, controls the reabsorption of sodium in the kidney. In Addison's disease there is lessened secretion of the hormone, with a resultant increase in excretion of sodium in the urine and a lowered level of sodium in the blood (hyponatremia). There is then increased excretion of water to restore the composition of the blood to normal. Severe dehydration is the ultimate result.

Persons whose kidneys are unable to function adequately, a condition called chronic renal insufficiency, may develop an inability to conserve salt. Such a defect may be aggravated by injudicious salt restriction to the point where the clinical state resembles adrenal cortical insufficiency.

Dehydration and hyponatremia may result from bleeding under the arachnoid, the middle layer of the brain covering, or from encephalitis, inflammation of the brain.

Water loss from vomiting and diarrhea

The normal daily turnover of fluids in the gastro-intestinal tract amounts to about eight litres (slightly more than eight quarts) in 24 hours, nearly 20 percent of the total body water. In vomiting or diarrhea large volumes of water may be lost, always with an associated loss of electrolytes (*e.g.*, sodium, potassium). The exact nature of the resultant electrolyte deficit will depend in part on the site of the lesion causing symptoms, and in part on whether the ill person is receiving other fluids. In uncomplicated, untreated vomiting or diarrhea, the dehydration that results is ordinarily a balanced loss of electrolytes and water.

Short- and long-term effects of dehydration. The symptoms of dehydration depend in part on the cause, and in part on whether there is associated salt deprivation as well. Since the body is divided into a number of compartments so far as its body fluids are concerned, alterations in the water or salt content of one compartment results in rapid shifts of water or salt in an attempt to re-establish the overall balance. To further complicate matters, the electrolytes and protein are in different proportions in the blood plasma and in the cell fluid. This means that the human body must regulate not only the amount of water and the concentration of electrolytes in each compartment, but also that it must assure that each compartment contains the right kind of salt (electrolyte) for its own normal function.

When loss of water is disproportionately greater than loss of electrolytes, the osmotic pressure of the extracellular fluids becomes higher than in the cells. Since water passes from a region of lower to a region of higher osmotic pressure, water flows out of the cells into the extracellular fluid, tending to lower its osmotic pressure and increase its volume toward normal. As a result of the flow of water out of the cells, they become dehydrated. This results in the thirst that always accompanies "pure" water depletion.

In those diseases in which there is loss of salt in excess of water, the decreased concentration of sodium in the extracellular fluid and in the blood serum results in decreased osmotic pressure, and water therefore enters the cells to equalize the osmotic pressure. Thus there is extracellular dehydration and intracellular hydration—and no thirst.

A comparison of the results of water and salt (electrolyte) deprivation is given in the Table. The degree of deprivation that is present affects these results. As is already

apparent, it is rare for either salt or water to be depleted by itself, since compensatory mechanisms result in significant alteration in the other when the concentration of one is changed.

Comparison of Effects of Water and Salt Depletion*

manifestation	pure water depletion	pure salt depletion
Thirst	+ + +	Absent
Lassitude	+	+ + +
Fainting upon standing up	Absent until late	+ + +
Urine volume	Scanty	Normal until late
NaCl in urine	Often +	Always absent except in Addison's disease
Vomiting	Absent	May be + + +
Cramps	Absent	May be + + +
Plasma NaCl	Slight increase or normal	Diminished + + +
Plasma volume	Normal until late	Decreased + + +
Blood urea	+	+ + +
Decrease of fluid content of blood	Not until late and slight	+ + +
Blood viscosity	Normal until late	Increased + + +
Blood pressure	Normal until late	Fall + + +
Water absorption	Rapid	Slow

* Plus signs indicate degree of severity, from slight (+) to extreme (+ + +).

Early and late effects of water deprivation

Gradual weight loss occurs in both types of dehydration, amounting to two to three pounds per day. Thirst is the inevitable accompaniment only of water deprivation, without concomitant salt loss. Dryness of the mouth, a craving for fluid, decreased production of saliva, and impaired swallowing all follow prolonged restriction of water intake. It is probable that thirst is the result of the subsequent intracellular dehydration and increased intracellular osmotic pressure. Experimentally, thirst can be produced when the cells have lost about one percent of their intracellular water.

As dehydration progresses, the tissues tend to shrink, the skin becomes dry and wrinkled, and the eyes become shrunken and the eyeballs soft. Fever develops, which may be mild, but may become marked as dehydration progresses. Dehydration itself probably affects the temperature regulatory centres in the brain. As dehydration and salt loss progress, however, the plasma volume and the output of the heart decrease, with consequent decrease of the blood supply to the skin. Sweating decreases and may stop completely, and the main avenue for heat loss is closed. The body temperature may then rise precipitously.

There are marked changes in the volume of the extracellular and intracellular fluids, but the blood plasma volume changes the last and the least.

Plasma volume maintained at expense of tissue fluids

The plasma volume is maintained more or less constant at the expense of the tissue fluids. When the plasma volume decreases, the concentration and the osmotic pressure of the plasma protein rise, and the venous and capillary pressures fall; these are processes that tend to maintain a normal plasma volume. If these mechanisms fail, and the plasma volume falls, the output of the heart also fails, and the pulse rate climbs, indicative of a dangerous physical state.

The renal (kidney) changes that occur in man during prolonged water depletion similarly tend to maintain a normal balance. Initially the renal blood flow is maintained and the glomerular filtration rate (the rate at which protein-free fluid containing nitrogenous waste products and other substances in solution are filtered out of the blood) may only be decreased 10–20 percent. If water deprivation continues and the plasma volume falls, however, with subsequent further decrease in the glomerular filtration rate, the output of urine will be drastically reduced. As long as urine output of over 30 millilitres (about one fluid ounce) per hour is maintained, the kidney can excrete nitrogenous and nonnitrogenous solids with maximum efficiency. Once the urine flow is decreased below this level, the kidney is unable to function efficiently, the substances are retained in the body, and their concentration in the blood rises.

The final result of prolonged dehydration is now apparent. The normal distribution of salt and water in the body is destroyed, the plasma volume decreases, and the blood viscosity increases. As a result of these changes renal function is impaired, the glomerular filtration rate falls, the urinary output falls, and waste products accumulate. Far more life-threatening, however, is decreased loss of moisture from the skin, with the subsequent rise in temperature, and the fall in cardiac output with the attendant irreversible shock.

Prognosis and nature of the treatment. Once renal failure occurs, about 8 percent of the total body water has been lost (four litres). When five to ten litres of body water have been lost a person is acutely and severely ill, with a contracted plasma volume, increased concentration and viscosity of the blood, renal failure and excessive urea in the blood, and a falling blood pressure. In a previously healthy adult death follows the loss of 12–15 litres of body water. In the very young, the very old, or the debilitated, death occurs at a lower level of dehydration.

Effects of degrees of dehydration

The treatment of any form of dehydration depends not only on restoring the depleted water, but also on the re-establishment of normal levels of body electrolytes (salt) and limitation of the production of nitrogenous waste products.

Before any of these therapeutic measures can be applied, however, the initiating cause must be removed. The sailor or the desert traveller must be rescued, the vomiting infant must be cured, or the underlying disease (*e.g.*, Addison's disease) must be treated. Then, after accurate biochemical determinations of the levels of various electrolytes and other blood components have been made and the plasma volume has been measured, the physician can commence giving measured quantities of the appropriate mixtures of salt and water. Given the right amounts of salt and water, the human body will gradually restore the normal relationships between the cells, the extracellular fluid, and the plasma volume. That done, the complicated functions of the kidney will clear the circulating blood of the retained waste products, and the body will have restored its own normal balance.

BIBLIOGRAPHY. T.R. HARRISON, *Principles of Internal Medicine*, 6th ed. by M.M. WINTROBE *et al.* (1970), an up-to-date reference text in internal medicine; W.A. SODEMAN and W.A. SODEMAN, JR., *Pathological Physiology: Mechanisms of Disease*, 4th ed. (1967), a basic work on the pathophysiology underlying the disease state; J.C. TODD and A.H. SANFORD, *Clinical Diagnosis by Laboratory Methods*, 14th ed. by I. DAVIDSOHN and J.B. HENRY (1969), a standard reference text in clinical pathology, regularly updated; S. WRIGHT, *Applied Physiology*, 11th ed. by C.A. KEELE and E. NEIL (1965), one of the basic teaching texts in medical physiology; J.H. BLAND (ed.), *Clinical Metabolism of Body Water and Electrolytes* (1963), one of the most comprehensive books available on clinical water and electrolyte metabolism and their alteration in health and disease.

(D.C.H.L.)

Deism

Deism, as the word is customarily employed, describes an unorthodox religious attitude that found expression among a group of English writers beginning with Edward Herbert (later 1st Baron Herbert of Cherbury) in the first half of the 17th century and ending with Henry St. John, 1st Viscount Bolingbroke, in the middle of the 18th century. In general, however, it refers to what can be called natural religion, the acceptance of a certain body of religious knowledge that is inborn in every person or that can be acquired by the use of reason, as opposed to knowledge acquired through either revelation or the teaching of any church.

The concept of natural religion

Nature and scope. Though an initial use of the term occurred in 16th-century France, the later appearance of the doctrine on the Continent was stimulated by the translation and adaptation of the English models. The high point of Deist thought occurred in England from about 1689 through 1742, during a period when, despite widespread counterattacks from the established Church of England, there was relative freedom of religious expression following upon the Glorious Revolution that

ended the rule of James II and brought William and Mary to the throne. Deism took deep root in 18th-century Germany after it had ceased to be a vital subject of controversy in England.

At times in the 19th and early 20th centuries, the word Deism was used theologically in contradistinction to theism, the belief in an immanent God who actively intervenes in the affairs of men. In this sense Deism was represented as the view of those who reduced the role of God to a mere act of creation in accordance with rational laws discoverable by man and held that, after the original act, God virtually withdrew and refrained from interfering in the processes of nature and the ways of man. So stark an interpretation of the relations of God and man, however, was accepted by very few Deists during the flowering of the doctrine, though their religious antagonists often attempted to force them into this difficult position. Historically, a distinction between theism and Deism has never had wide currency in European thought. As an example, when encyclopaedist Denis Diderot, in France, translated into French the works of Anthony Ashley Cooper, 3rd earl of Shaftesbury, one of the important English Deists, he often rendered "Deism" as *théisme*. The term is not in current usage as a metaphysical concept, and its significance is really limited to the 17th and 18th centuries.

The historical Deists. *The English Deists.* In 1754–56, when the Deist controversy had passed its peak, John Leland, an opponent, wrote a historical and critical compendium of Deist thought, *A View of the Principal Deistical Writers that Have Appeared in England in the Last and Present Century; with Observations upon Them, and Some Account of the Answers that Have Been Published Against Them.* This work, which began with Lord Herbert of Cherbury and moved through the political philosopher Thomas Hobbes, Charles Blount, the Earl of Shaftesbury, Anthony Collins, Thomas Woolston, Matthew Tindal, Thomas Morgan, Thomas Chubb, and Viscount Bolingbroke, fixed the canon of who should be included among the Deist writers. In subsequent works Hobbes usually has been dropped from the list and John Toland included, though he was closer to pantheism than most of the other Deists were. Herbert was not known as a Deist in his day, but Blount and the rest who figured in Leland's book would have accepted the term Deist as an appropriate designation for their religious position. Simultaneously, it became an adjective of opprobrium in the vocabulary of their opponents. Bishop Edward Stillingfleet's *Letter to a Deist* (1677) is an early example of the orthodox use of the epithet.

Herbert's five fundamental beliefs

In Lord Herbert's treatises five religious ideas were recognized as God-given and innate in the mind of man from the beginning of time: the belief in a supreme being, in the need for his worship, in the pursuit of a pious and virtuous life as the most desirable form of worship, in the need of repentance for sins, and in rewards and punishments in the next world. These fundamental religious beliefs, Herbert held, had been the possession of the first man, and they were basic to all the worthy positive institutionalized religions of later times. Thus, differences among sects and cults all over the world were usually benign, mere modifications of universally accepted truths; they were corruptions only when they led to barbarous practices such as the immolation of human victims and the slaughter of religious rivals.

In England at the turn of the 17th century this general religious attitude assumed a more militant form, particularly in the works of Toland, Shaftesbury, Tindal, Woolston, and Collins. Though the Deists differed among themselves and there is no single work that can be designated as the quintessential expression of Deism, they joined in attacking both the existing orthodox church establishment and the wild manifestations of the dissenters. The tone of these writers was often earthy and pungent, but their Deist ideal was sober natural religion without the trappings of Catholicism and the High Church in England and free from the passionate excesses of Protestant fanatics. In Toland there is great emphasis on the rational element in natural religion; in Shaftesbury more worth is ascribed to the emotive quality of religious experience when it is directed into salutary channels. All are agreed in denouncing every kind of religious intolerance because the core of the various religions is identical. In general, there is a negative evaluation of religious institutions and the priestly corps who direct them. Simple primitive monotheism was practiced by early men without temples, churches, and synagogues, and modern men could readily dispense with religious pomp and ceremony. The more elaborate and exclusive the religious establishment, the more it came under attack. A substantial portion of Deist literature was devoted to the description of the noxious practices of all religions in all times, and the similarities of pagan and Roman Catholic rites were emphasized.

The Deists who presented purely rationalist proofs for the existence of God, usually variations on the argument from the design or order of the universe, were able to derive support from the vision of the lawful physical world that Sir Isaac Newton had delineated. Indeed, in the 18th century, there was a tendency to convert Newton into a matter-of-fact Deist—a transmutation that was contrary to the spirit of both his philosophical and his theological writings.

When Deists were faced with the problem of how man had lapsed from the pure principles of his first forebears into the multiplicity of religious superstitions and crimes committed in the name of God, they ventured a number of conjectures. They surmised that men had fallen into error because of the inherent weakness of human nature; or they subscribed to the idea that a conspiracy of priests had intentionally deceived men with a "rout of ceremonials" in order to maintain power over them.

The role of Christianity in the universal history of religion became problematic. For many religious Deists the teachings of Christ were not essentially novel but were, in reality, as old as creation, a republication of primitive monotheism. Religious leaders had arisen among many peoples—Socrates, Buddha, Muḥammad—and their mission had been to effect a restoration of the simple religious faith of early men. Some writers, while admitting the similarity of Christ's message to that of other religious teachers, tended to preserve the unique position of Christianity as a divine revelation. It was possible to believe even in prophetic revelation and still remain a Deist, for revelation could be considered as a natural historical occurrence consonant with the definition of the goodness of God. The more extreme Deists, of course, could not countenance this degree of divine intervention in the affairs of men.

Natural religion was sufficient and certain; the tenets of all positive religions contained extraneous, even impure elements. Deists accepted the moral teachings of the Bible without any commitment to the historical reality of the reports of miracles. Most Deist argumentation attacking the literal interpretation of Scripture as divine revelation leaned upon the findings of 17th-century biblical criticism. Woolston, who resorted to an allegorical interpretation of the whole of the New Testament, was an extremist even among the more audacious Deists. Tindal was perhaps the most moderate of the group. Toland was violent; his denial of all mystery in religion was supported by analogies among Christian, Judaic, and pagan esoteric religious practices, equally condemned as the machinations of priests.

The Deists were particularly vehement against any manifestation of religious fanaticism and enthusiasm. In this respect Shaftesbury's *Letter Concerning Enthusiasm* (1708) was probably the crucial document in propagating their ideas. Revolted by the Puritan fanatics of the previous century and by the wild hysteria of a group of French exiles prophesying in London in 1707, Shaftesbury denounced all forms of religious extravagance as perversions of true religion. These false prophets were directing religious emotions, benign in themselves, into the wrong channels. Any description of God that depicted his impending vengeance, vindictiveness, jealousy, and destructive cruelty was blasphemous. Because sound religion could find expression only among healthy men, the argu-

ment was common in Deist literature that the preaching of extreme asceticism, the practice of self-torture, and the violence of religious persecutions were all evidence of psychological illness and had nothing to do with authentic religious sentiment and conduct. The Deist God, ever gentle, loving, and benevolent, intended men to behave toward one another in the same kindly and tolerant fashion.

French Deists

Deists in other countries. Ideas of this general character were voiced on the Continent at about the same period by such men as Pierre Bayle, a French philosopher famous for his encyclopaedic dictionary, even though he would have rejected the Deist identification. During the heyday of the French Philosophes in the 18th century, the more daring thinkers—Voltaire among them—gloried in the name Deist and declared the kinship of their ideas with those of Rationalist English ecclesiastics, such as Samuel Clarke, who would have repudiated the relationship. The dividing line between Deism and atheism among the Philosophes was often rather blurred, as is evidenced by *Le Rêve de d'Alembert* (written 1769; "The Dream of d'Alembert"), which describes a discussion between the two "fathers" of the *Encyclopédie:* the Deist Jean Le Rond d'Alembert and the atheist Diderot. Diderot had drawn his inspiration from Shaftesbury, and thus in his early career he was committed to a more emotional Deism. Later in life, however, he shifted to the atheist materialist circle of the Baron d'Holbach. When Holbach paraphrased or translated the English Deists, his purpose was frankly atheist; he emphasized those portions of their works that attacked existing religious practices and institutions, neglecting their devotion to natural religion and their adoration of Christ. The Catholic Church in 18th-century France did not recognize fine distinctions among heretics, and Deist and atheist works were burned in the same bonfires.

German Deists

English Deism was transmitted to Germany primarily through translations of Shaftesbury, whose influence upon thought was paramount. In a commentary on Shaftesbury published in 1720, Gottfried Wilhelm Leibniz, a Rationalist philosopher and mathematician, accepted the Deist conception of God as an intelligent Creator but refused the contention that a god who metes out punishments is evil. A sampling of other Deist writers was available particularly through the German rendering of Leland's work in 1755 and 1756. H.S. Reimarus, author of many philosophical works, maintained in his *Apologie oder Schutzschrift für die vernünftigen Verehrer Gottes* ("Defense for the Rational Adorers of God") that the human mind by itself without revelation was capable of reaching a perfect religion. Reimarus did not dare to publish the book during his lifetime, but it was published in 1774–78 by Gotthold Ephraim Lessing, one of the great seminal minds in German literature. According to Lessing, common man, uninstructed and unreflecting, will not reach a perfect knowledge of natural religion; he will forget or ignore it. Thus, the several positive religions can help men achieve more complete awareness of the perfect religion than could ever be attained by any individual mind. Lessing's *Nathan der Weise* (1779; "Nathan the Sage") was noteworthy for the introduction of the Deist spirit of religion into the drama; in the famous parable of the three rings, the major monotheistic religions were presented as equally true in the eyes of God. Although Lessing's rational Deism was the object of violent attack on the part of Pietist writers and the more mystical thinkers, it influenced such men as Moses Mendelssohn, a German Jewish philosopher who applied Deism to the Jewish faith. Immanuel Kant, the most important figure in 18th-century German philosophy, stressed the moral element in natural religion; moral principles are not the result of any revelation but originate from the very structure of man's reason. English Deists, however, continued to influence German Deism. Witnesses attest that virtually the whole officer corps of Frederick the Great was infected with Deism and that Collins and Tindal were favourite reading in the army.

American Deists

By the end of the 18th century, Deism had become a dominant religious attitude among intellectual and upper class Americans. Benjamin Franklin, the great sage of the Colonies and then of the new republic, summarized in a letter to Ezra Stiles, president of Yale College, a personal creed that almost literally reproduced Herbert's five fundamental beliefs. The first three presidents of the United States also held Deistic convictions, as is amply evidenced in their correspondence. "The ten commandments and the sermon on the mount contain my religion," John Adams wrote to Thomas Jefferson in 1816.

BIBLIOGRAPHY. JOHN LELAND, *A View of the Principal Deistical Writers . . .*, 3rd ed., 3 vol. (1754; also 1837 ed.), the first historical account; FRITZ MAUTHNER, *Der Atheismus und seine Geschichte im Abendlande*, 4 vol. (1921–23), a complete history; ERNST CASSIRER, *Die Philosophie der Aufklärung* (1932; Eng. trans., *The Philosophy of the Enlightenment*, 1951), a description of Deism and its philosophical background; HAROLD G. NICOLSON, *The Age of Reason* (1960), on the nature of 18th-century Rationalism and its connection with Deism; JAMES COLLINS, *God in Modern Philosophy* (1959), a full history of Deism, here called "theism," from Nicolas of Cusa to contemporary theological theories; JOHN ORR, *English Deism: Its Roots and Its Fruits* (1934); GOTTHARD V. LECHLER, *Geschichte des englischen Deismus* (1841), the first full history after the end of Deism; HERBERT OF CHERBURY, *De Veritate* (1624; Eng. trans. by MEYRICK H. CARRE, *On Truth*, 1937), the first English translation of the reputedly "first" classic expression of Deism; MARIO M. ROSSI, *La vita, le opere, i tempi di Edoardo Herbert di Chirbury*, 3 vol. (1947), and *Alle fonti del deismo e del materialismo moderno* (1942), two works that describe Herbert's life and Deistic thought against the background of the history of Deism and the attitude of the church; DAVID HUME, *Dialogues Concerning Natural Religion*, 2nd ed. with suppl. (1947), the beginning of the Deist's self-criticism; THOMAS PAINE, *The Age of Reason*, 3 pt. (1794–1811), the work most influential on the Deism of common people; JOHN S. SPINK, *French Free-Thought from Gassendi to Voltaire* (1960), on French Deism; HENRY E. ALLISON, *Lessing and the Enlightenment* (1966); IMMANUEL KANT, *Die Religion innerhalb der Grenzen der blossen Vernunft* (1793; Eng. trans., *Religion Within the Limits of Reason Alone*, 1947), the classic work of the last stage of German Deism; G.W.F. HEGEL, *Early Theological Writings*, trans. by THOMAS M. KNOX and RICHARD KRONER (1948), early writings to show Hegel's indebtedness to Deistic polemics.

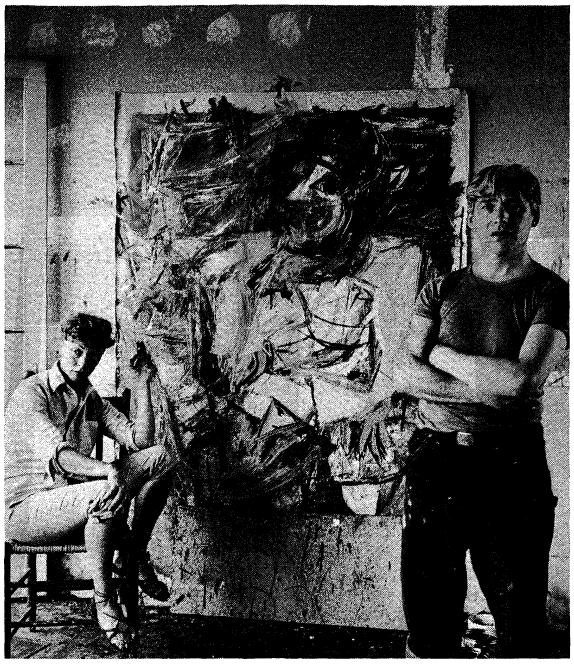
(F.E.M.)

De Kooning, Willem

Willem de Kooning is one of the most distinguished surviving painters of Abstract Expressionism, an American art movement characterized by free-associative gestures in paint and sometimes referred to as action painting. Most famous for his harsh depictions of women in the early 1950s, his work has fluctuated between figuration and abstraction since the early 1930s, when his reputation was established among his peers. His work has had substantial influence on his younger American contemporaries, especially during the late 1950s, and he enjoys a wide international reputation.

De Kooning was born in Rotterdam, Holland, April 24, 1904, the son of Leendert de Kooning and Cornelia Nobel, who were divorced about 1909. He was raised by his mother and a stepfather. In 1916 he was apprenticed to a firm of commercial artists and decorators, and about the same time he enrolled in night classes at the Academie voor Beeldende Kunsten en Technische Wetenschappen (Rotterdam Academy of Fine Arts and Techniques), where he studied creative and practical art for eight years. In 1920 he went to work for the art director of a large department store. During this period he became aware of modern art movements, especially the de Stijl group led by the painters Piet Mondrian and Theo van Doesburg, who were dedicated to geometric structure, abstraction, and simplification in art. With the exception of a few academic sketches, none of de Kooning's work from this time still exists.

In 1926 de Kooning entered the United States as a stowaway and eventually settled in Hoboken, New Jersey, where he supported himself as a house painter. In 1927 he moved to a studio in Manhattan and came under the influence of the artist, connoisseur, and art critic John Graham and the painter Arshile Gorky. Gorky became one of de Kooning's closest friends.



De Kooning and his wife, Elaine, photograph by Hans Namuth, 1952.
Hans Namuth

Early work in America

From about 1928 de Kooning began to paint still-life and figure compositions reflecting school of Paris and Mexican influences. By the early 1930s he was exploring abstraction, using biomorphic shapes and simple geometric compositions—an opposition of disparate formal elements that prevails in his work throughout his career. These early works had strong affinities with those of his friends Graham and Gorky and reflected the impact on these young artists of Pablo Picasso and the Surrealist Joan Miró, both of whom achieved powerfully expressive compositions through biomorphic forms.

In October 1935, de Kooning began to work on the WPA (Works Progress Administration) Federal Art Project. He was employed by this work-relief program until July 1937, when he was forced to resign because of his alien status. This period of about two years provided the artist, who had been supporting himself during the early Depression by commercial jobs, with his first opportunity to devote full time to creative work. He worked on both the easel-painting and mural divisions of the project (the several murals he designed were never executed). After he left the WPA he was commissioned to design a mural for the Hall of Pharmacy at the New York World's Fair of 1939, which was painted by professional muralists and destroyed when the fair ended. In 1940 he received a commission from the federal government's Section of Fine Arts to create a mural for the library of the S.S. President Jackson, a U.S. Navy ship. De Kooning's mural designs for the Federal Art Project were similar in style to his abstract works of the early 1930s; his murals for the World's Fair and the Section of Fine Arts were more representational.

Coloured abstractions

In 1938, probably under the influence of Gorky, de Kooning embarked on a series of sad, staring male figures, such as "Two Men Standing," "Man," and "Seated Figure (Classic Male)." Parallel with these works he also created lyrically coloured abstractions, such as "Pink Landscape" and "Elegy." This coincidence of figures and abstractions continued well into the 1940s with his representational but somewhat geometricized "Woman" and "Standing Man," along with numerous untitled abstractions whose biomorphic forms increasingly suggest the presence of figures. By about 1945 the two tendencies seem to fuse perfectly in "Pink Angels." In 1946, too poor to buy artists' pigments, he turned to black and white household enamels to paint a series of large abstractions; of these works, "Light in August" and "Black Friday" are essentially black with white elements, whereas "Zurich" and "Mailbox" are white with black. Developing out of these works in the period after his first show are complex, agitated abstractions like "Asheville," "Attic," and "Excavation," which reintroduce colour and seem to sum up with taut decisiveness the problems of free-associative composition he had struggled with for many years.

In 1938 de Kooning had met Elaine Fried, whom he married in 1943. During the 1940s and thereafter he became increasingly identified with the Abstract Expressionist movement and was recognized as one of its leaders in the mid-1950s. He had his first one-man show, which consisted of his black-and-white enamel compositions, at the Charles Egan Gallery in New York in 1948, and taught at Black Mountain College in North Carolina in 1948 and at the Yale School of Art 1950–51.

Variations on the theme of "Woman"

While de Kooning had painted women regularly in the early 1940s and again from 1947 to 1949, and while the biomorphic shapes of his early abstractions can be interpreted as female symbols, it was not until 1950 that he began to explore the subject of women exclusively. In the summer of that year he began "Woman I," which went through innumerable metamorphoses before it was finished in 1952. During this period he also created other paintings of women. These works were shown at the Sidney Janis Gallery in 1953 and caused a sensation, chiefly because they were figurative when most of his fellow Abstract Expressionists were painting abstractly, and because of their blatant technique and imagery. The savagely applied pigment and the use of colours that seem vomited on his canvas combine to reveal a woman all too congruent with some of modern man's most widely held sexual fears. The toothy snarls, overripe, pendulous breasts, vacuous eyes, and blasted extremities imaged the darkest Freudian insights. The "Woman" paintings II through VI are all variants on this theme, as are "Woman and Bicycle" and "Two Women in the Country." The deliberate vulgarity of these paintings contrasts with the French painter Jean Dubuffet's no less harsh "Corps de dame" series of 1950, in which the female, formed with a rich topography of earth colours, relates more directly to universal symbols.

By 1955, however, de Kooning seems to have turned to this symbolic aspect of woman; this is suggested by the title of his "Woman as Landscape," in which the vertical figure seems almost absorbed into the abstract background. There followed a series of landscapes such as "Police Gazette," "Gotham News," "Backyard on Tenth Street," "Parc Rosenberg," "Suburb in Havana," "Door to the River," and "Rosy-Fingered Dawn at Louse Point," which display an evolution from compositional and coloristic complexity to a broadly painted simplicity.

Around 1963, the year he moved permanently to the springs at East Hampton, Long Island, de Kooning returned to depicting women in "Pastorale" and "Clam Diggers." The theme is re-explored over the next years in "Woman Acabonic," "Woman and Child," and "The Visit." These paintings, as controversial as his earlier women, are satiric attacks on the female anatomy, painted with a flamboyant lubricity in keeping with the uninhibited subject matter.

Critics have charged that, compared with his paintings of the late 1940s and 1950s, his work since the 1960s seems to exploit old themes and lacks that archaeological stratification of paint surfaces so indicative of his former struggle with the processes of creation. Yet these works are not his last, and his critics' taste for more recent trends in American art may well prevent them from accepting the logical development toward freedom of a very subjective artist.

MAJOR WORKS

"Two Men Standing" (*c.* 1938; private collection, New York); "Pink Landscape" (*c.* 1938; Mr. and Mrs. Reuben Tam Collection, New York); "Man" (*c.* 1939; private collection, New York); "Elegy" (*c.* 1939; private collection); "Glazier" (*c.* 1940; private collection, New York); "Seated Figure (Classic Male)" (*c.* 1940; private collection); "Standing Man" (*c.* 1942; Wadsworth Atheneum, Hartford, Conn.); "Woman"

(*c.* 1943; private collection, New York); "Portrait of Max Margulis" (*c.* 1944; private collection); "Queen of Hearts" (1943–46; Joseph H. Hirshhorn Collection, New York); "Pink Angels" (*c.* 1945; Mr. and Mrs. Frederick R. Weisman Collection, Beverly Hills, California); "Light in August" (*c.* 1946; Elise C. Dixon Collection, Scottsdale, Arizona); "Mailbox" (1947–48; Nelson A. Rockefeller Collection, New York); "Zurich" (1947; private collection); "Painting" (1948; Museum of Modern Art, New York); "Black Friday" (1948; Mrs. H. Gates Lloyd Collection, Haverford, Pennsylvania); "Woman" (*c.* 1949; University of North Carolina); "Asheville" (1948–49; Phillips Collection, Washington, D.C.); "Attic" (1949; Muriel Newman Collection, Chicago); "Boudoir" (*c.* 1950; William Rockhill Nelson Gallery and Atkins Museum of Fine Arts, Kansas City, Mo.); "Excavation" (1950; Art Institute, Chicago); "Two Women" (1952; Art Institute Chicago); "Woman I" (1952; Museum of Modern Art, New York); "Woman II" (1952–53; Museum of Modern Art, New York); "Woman IV" (1952–53; William Rockhill Nelson Gallery and Atkins Museum of Fine Arts, Kansas City, Mo.); "Woman and Bicycle" (1953; Whitney Museum of American Art, New York); "Woman VI" (1953; Carnegie Institute, Pittsburgh, Pennsylvania); "Two Women in the Country" (1954; Joseph H. Hirshhorn Collection, New York); "Police Gazette" (1954–55; Robert C. Scull Collection, New York); "Woman As Landscape" (1955: private collection); "Gotham News" (1955–56; Albright-Knox Art Gallery, Buffalo, New York); "Easter Monday" (1956; Metropolitan Museum of Art, New York); "Backyard on Tenth Street" (1956; Baltimore Museum of Art, Baltimore, Maryland); "Parc Rosenberg" (1957; Isabel and Donald Grossman Collection, New York); "Suburb in Havana" (1958; Mr. and Mrs. Lee V. Eastman Collection, New York); "Door to the River" (1960; Whitney Museum of American Art, New York); "Rosy-Fingered Dawn at Louse Point" (1963; Stedelijk Museum, Amsterdam); "Pastorale" (1963; private collection); "Clam Diggers" (1964, private collection); "Woman Acabonic" (1966; Whitney Museum of American Art, New York); "Woman and Child" (1967; Joseph and Mildred Gosman Collection, Toledo, Ohio); "The Visit" (1967; M. Knoedler and Co., Inc.).

BIBLIOGRAPHY. THOMAS B. HESS has written two books entitled *Willem de Kooning*, the first in 1959 as a monograph in the "Great American Artist Series" and the second as the catalog of the Museum of Modern Art's travelling retrospective in 1968. The latter contains a basic bibliography and the text of de Kooning's three major published statements. An interpretative monograph has been written by HARRIET JANIS and RUDI BLESH, *De Kooning* (1960); and an important article describing the creation of "Woman I" is THOMAS B. HESS, "De Kooning Paints a Picture," *Art News*, 51:30–33 (1953).

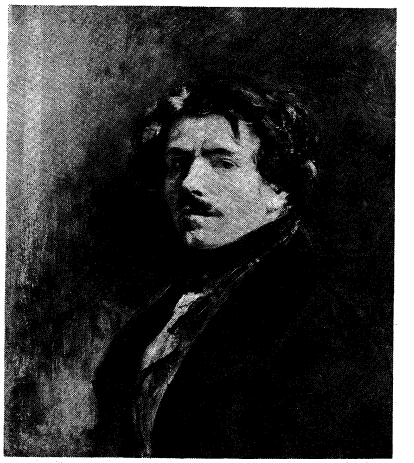
(F.V.O'C.)

Delacroix, Eugène

Eugène Delacroix, the greatest French painter of the first half of the 19th century, possessed a powerful imagination, a subtle intelligence, and a rare (and touchy) sensitivity that made him perhaps the century's least understood artist. He wrote in 1850 in his *Journal:* "The Beautiful is found but once, at a certain appointed moment in history. Too bad for the genius who comes after." He might well have added, in view of the misunderstandings that caused him so much grief, "Too bad for the genius who comes *before*." For his genius foreshadowed the subtleties and the tragic side of modern art, which was only understood in his time by the poetic intuition of Charles Baudelaire.

With the English painter J.M.W. Turner, he was the initiator of the bold technical innovations of Impressionism and the forerunner of modern Expressionism. Auguste Renoir, Claude Monet, Paul Cézanne, Paul Gauguin, Vincent van Gogh, Odilon Redon, Georges Seurat, Henri Matisse, and Pablo Picasso all recognized their debt to him—his use of colour and skill in modelling equalled that of Titian or Rubens. Too broad an artist to be confined within one aesthetic movement, he was yet characterized, about 1830, as a Romantic. Displeased, he replied, "I am a pure Classicist." He wished to transmit through art the poetry of his being. Thus he created in each work an expressive unity, a symphony of forms, of light, and especially of colour.

Early life. Ferdinand-Victor-Eugène Delacroix was born on April 26, 1798, at Charenton-Saint-Maurice, the fourth child of Victoire-Oeben, a descendant of the Oeben-Riesener family, which had created furniture for the king and court in the 17th and 18th centuries, and of Charles Delacroix, a government official who was ambassador to Holland in 1798 and who died in 1805 while prefect of Bordeaux. One tradition attributes Eugène's true paternity to the statesman Charles-Maurice de Talleyrand-Périgord. But the childhood of the future painter was never troubled by any doubt, and he would always maintain great affection and admiration for his father. Up to the age of 17 he pursued classical studies. Within his distinguished and artistic family he found a passion for music and theatre, the manners, and the relationships that would make him a perfect dandy.



Delacroix, self-portrait, oil on canvas, *c.* 1837. In the Louvre, Paris. 65 cm × 55 cm.
Giraudon

Early associations with painters

In 1815 he became the pupil of a famous academic painter, Baron Pierre-Narcisse Guérin. He knew the historical painter Antoine-Jean Gros, and as a young man visited the salon of the royalist and painter Baron François Gérard. As early as 1822 he received the backing of Louis Thiers, the statesman and historian, who, as interior minister in the 1830s, put Delacroix in charge of architectural decorations.

A child of his century, Delacroix shared the melancholy that afflicted the sons and brothers of the heroes of the Napoleonic Wars. He was affected by the Romanticism of the painter Théodore Géricault, and of his friends the English painter Richard Parkes Bonington, the Polish-born composer and pianist Frédéric Chopin, and the French writer George Sand. He did not, however, take part in the battles of the Romantic movement waged by Victor Hugo, Hector Berlioz, and others.

Development of mature style. His first masterpiece, "Dante and Virgil in Hell," was inspired by Dante's *Divine Comedy*, but its tragic feeling is reminiscent of Michelangelo and Rubens. The dramatic contemporary event of the massacre of Greeks by Turks on the island of Chios inspired a large canvas, "The Massacre at Chios," submitted to the Paris Salon of 1824. The haughty pride of the conquerors, the horror as well as despair of the innocent Greeks, and the splendour of a vast sky create an expressive unity in which the nature of his genius is evident.

Influence of English artists

Already interested in the delicate technique of his English painter friends Bonington and the Fielding brothers (Thales, Copley, Theodore, and Newton), he also admired the English landscapes of John Constable, which were exhibited in Paris in 1824. To round out his technical and cultural education, Delacroix left for London in 1825. There, his technique, developed by contact with Turner, Constable, and Sir Thomas Lawrence, acquired the freedom and suppleness that until then he had been seeking in Rubens. Between 1827 and 1832, masterpieces came in quick succession. Paintings of historical subjects included "The Battle of Nancy" and "Bataille de Poitiers." The subjects of other paintings were taken from the English poet Byron. He also made a set of 17 lithographs illustrating a French edition of Goethe's *Faust*.

Inspired by the Revolution of 1830, "Liberty Leading the People" was the last evidence of a Romanticism he finally rejected. After, he tried to avoid the characteristics of his earlier works: the dizzying colours of "The Death of Sardanapalus," the turmoil of massacres as in "The Assassination of the Bishop of Liège," and the gaudy trumpery of history. He did not intend to surrender to his literary penchant for dramatic expression any more than to his taste for exactitude in historical detail. He needed to confront a reality that would be a source of inner exaltation.

Moroccan influence

He found it on a trip to Morocco in 1832. From January to July he toured with the Comte de Mornay, King Louis-Philippe's diplomatic representative to the Sultan. The sight of exuberant nature, the beauty of the horses, of the Arabs and their dress, of primitive savagery, enhanced by the light of memory, would henceforth inspire him, even in his last works. In the Moroccan people he found the ancient stability that his cultured Frenchman's soul craved, just as he would find it in 1846 in the dignity of Ojibwa Indians brought to Europe from the United States by the ethnologist-painter George Catlin.

Building decoration. Delacroix's first commission for the decoration of a government building came in 1833. This project, a group of murals for the Salon du Roi at the Palais-Bourbon, posed problems of architectural decoration that he was to resolve through the power of colour, taking his inspiration from the Venetians, Rubens, and the work of the Renaissance Italians Il Rosso and Francesco Primaticcio at Fontainebleau. He used powerfully organized colour, unifying the elements of the composition down to the last poetic detail. From then on his decorations were on a par with those of the Renaissance geniuses. Notable were his commissions for the museum of history at Versailles: "La Bataille de Taillebourg" and "The Taking of Constantinople by the Crusaders." In these works he combined history, myth, philosophy, religion, and man's suffering into one grandiose symphony. From 1835 to 1861 he executed

Large decorative works

many large decorations. At the Palais-Bourbon he painted allegories on the ceiling and in friezes above the doorways of the Salon du Roi; on the ceiling of the library (completed 1847), he painted two hemicycles and 20 pendentives. For the library of the Palais du Luxembourg (completed 1846), he painted the hemicycle, the cupola, and the four pendentives. At the Louvre (completed 1851) he painted the central ceiling of the Galerie d'Apollon. At the Hôtel de Ville (City Hall of Paris, burned in 1871) he painted (beginning 1849) the ceiling, eight panels, and 11 tympana of the Salon de la Paix. For the Chapel of the Holy Angels at the church of Saint-Sulpice, he painted Saint Michael on the ceiling and wall frescoes of Jacob and the angel and of Heliodorus (completed 1861).

The task of creating these works strained his health. And, by 1849 he had lost all his closest relatives and friends: his nephew died in New York in 1834; his brother, Gen. Charles Delacroix, in 1845; Chopin, in 1849. After this, he lived only for the creations of his imagination. On August 13, 1863, he died in Paris in his apartment on the Place Fürstenberg, which was later made into a national museum in his honour.

MAJOR WORKS

PAINTINGS: "Dante et Virgile aux enfers" ("Dante and Virgil in Hell," 1822; Louvre, Paris); "Scènes des massacres de Scio" ("The Massacre at Chios," 1824; Louvre); "Baron Schwiter" (*c.* 1826–27; National Gallery, London); "The Execution of the Doge Marino Faliero" (*c.* 1826–27; Wallace Collection, London); "Combat Between the Giaour and the Pasha" (1827; Art Institute of Chicago); "La Mort de Sardanapale" ("The Death of Sardanapalus," 1827; Louvre); "La Grèce expirant sur les ruines de Missolonghi" ("Greece Expiring on the Ruins of Missolonghi," 1827; Musée des Beaux-Arts, Bordeaux); "L'Assassinat de l'évêque de Liège" ("The Assassination of the Bishop of Liège," 1829; Louvre); "La Liberté guidant le peuple (Le 28 juillet 1830)" (1830; Louvre); "Bataille de Poitiers" ("The Battle of Poitiers," 1830; Louvre); "Arab Fantasy" (1833; Städelsches Kunstinstitut, Frankfurt am Main); "The Battle of Nancy" (1831; Musée des Beaux-Arts, Nancy); "Boissy d'Anglas at the Convention" (1831; Musée des Beaux-Arts, Bordeaux); "Femmes d'Alger dans leur appartement" ("Women of Algiers in Their Apartment," 1834; Louvre); "Portrait de l'artiste" (*c.* 1837; Louvre); "La Bataille de Taillebourg" ("The Battle of Taillebourg," 1837; Louvre); "Furious Medea" (1838; Musée des Beaux-Arts, Lille); "Portrait de Frédéric Chopin" (1838; Louvre); "Portrait of George Sand" (1838; Øregaard-Museum, Copenhagen); "The Return of Columbus" (1839; Toledo Museum of Art); "The Justice of Trajan" (1840; Musée des Beaux-Arts et de la Céramique, Rouen); "Prise de Constantinople par les Croisés, 12 avril 1204" ("The Taking of Constantinople by the Crusaders on April 12, 1204," 1840; Louvre, Paris); "Naufrage de Don Juan" ("The Shipwreck of Don Juan," 1841; Louvre); "George Sand's Garden at Nohant" (*c.* 1842–43; Metropolitan Museum of Art, New York); "Muley-Abd-er-Rahmann, Sultan of Morocco" (1845; Musée des Augustins, Toulouse); "The Abduction of Rebecca" (1846; Metropolitan Museum of Art, New York); "The Entombment" (1848; Museum of Fine Arts, Boston); "Alfred Bruyas" (1853; Musée Fabre, Montpellier); "Christ on the Sea of Galilee" (1854; Walters Art Gallery, Baltimore); "Turkish Women Bathing" (1854; Wadsworth Atheneum, Hartford, Connecticut); "View of Tangiers from the Seashore" (1858; Minneapolis Institute of Art); "A Lion Hunt" (1858; Museum of Fine Arts, Boston); "L'Enlèvement de Rébecca" ("The Abduction of Rebecca," 1858; Louvre); "The Lion Hunt" (1861; Art Institute of Chicago).

DECORATIONS: "Orpheus Bringing Civilization"; "Attila Followed by Barbarian Hordes Tramples Italy and the Arts"; "Philosophy," "Natural Sciences," "Legislation," "Theology," and "Poetry" (commissioned in 1833, inaugurated in 1847; ceiling of library of the Chambre des Députés, Palais-Bourbon, Paris); cupola of the library (commissioned in 1840, inaugurated in December 1846; Palais du Luxembourg, Paris); "Apollo Destroying the Serpent Python" (commissioned in 1850, inaugurated in October 1851; Galerie d'Apollon, Louvre); Chapel of the Holy Angels, church of Saint-Sulpice, Paris; "The Archangel St. Michael Subduing the Demon"; "Heliodorus Driven from the Temple"; "Jacob Wrestling with the Angel" (commissioned in 1849, completed in 1861); albums of drawings and water colours from his travels in Morocco (Louvre; Musée Condé, Chantilly).

BIBLIOGRAPHY

Literary works: As an adolescent, Eugène Delacroix wished to be a poet. His early efforts in the theatre have been edited by JEAN MARCHAND, *Les Dangers de la cour*, new ed. (1960). Many important articles on aesthetics that he published during his lifetime were collected by his executor, E.A. PIRON, *Eugène Delacroix: sa vie et ses oeuvres* (1865); and later by ELIE FAURE in *E. Delacroix: oeuvres littéraires*, 2 vol. (1923). There are two important editions of the famous *Journal*, indispensable to a knowledge of the psychology of the painter. One is annotated by ANDRE JOUBIN, *Journal de Eugène Delacroix*, 3 vol. (1932), and the re-edition of 1960 with a preface by J.L. VAUDOYER. There are also several selective editions in English, the most recent by HUBERT WELLINGTON, *The Journal of Eugène Delacroix* (1951). An important edition of Delacroix's correspondence was edited by ANDRE JOUBIN, *Correspondance générale d'Eugène Delacroix*, 5 vol. (1936–38).

Critical and biographical works: Among the famous 19th-century writers and critics who have comprehended and analyzed the art of Delacroix, the poet CHARLES BAUDELAIRE was the most perspicacious. See *Curiosités esthétiques* (1868) and *L'Art romantique* (1868). The basic and indispensable work for any scientific study of the ensemble of paintings, drawings, engravings, and lithographs was compiled from primary sources by ALFRED ROBAUT and ERNEST CHESNEAU, *L'Oeuvre complet d'Eugène Delacroix* (1885, reprinted 1969), containing numerous sketches, addenda, and an analytical table. The memorial booklet of the exhibition organized at the Louvre on the occasion of the centenary of the death of the artist, MAURICE SERULLAZ, *Mémorial de l'exposition Eugène Delacroix* (1963), is an excellent tool for analytic study. Most important for the life and history of the works are ETIENNE MOREAU-NELATON, *Delacroix raconté par lui-même*, 2 vol. (1916); and RAYMOND ESCHOLIER, *Delacroix peintre, graveur, écrivain*, 3 vol. (1926–29). The pointillist painter PAUL SIGNAC, *D'Eugène Delacroix au néoimpressionnisme*, 4th ed. (1939), demonstrates the importance of Delacroix for modern art. RENE PIOT, a disciple and collaborator of Delacroix in the large decorations, has provided important details on his technique in *Les Palettes de Delacroix* (1931). RAYMOND REGAMEY, *Eugène Delacroix: l'époque de la Chapelle des Saints-Anges, 1847–1863* (1931), analyzes the thinking of the painter's later style. A work that assigns Delacroix a place in the individualism of the 19th century is LUCIEN RUDRAUF, *Eugène Delacroix et le problème du romanticisme artistique* (1942). JEAN CAS-

sou, *Delacroix* (1947), has chosen beautiful photographic details to explain the force of the "fauve" (wild beast) art. MAURICE SERULLAZ, *Les Peintures murales de Delacroix* (1963), is a systematic study of the decorations, from the classical attempts at the Abbaye dc Valmont to the Chapel of Saint Sulpice. PHILIPPE JULLIAN, *Delacroix* (1963), is essentially a study of the environment and the friends of the painter. The best British specialist on the master is LEE JOHNSON, *Delacroix* (1963). RENE HUYGHE, *Delacroix* (1963), uses analytical and psychological methods in his critical technique and his *Delacroix and Greece* (Eng. trans. 1971), seeks to determine the manner in which the symbol of the liberation of Greece, through Eugène Delacroix, was of prime importance in the thought and art of the early 19th century.

(R.Hu.)

Delaware

The first of the original 13 American states to ratify the federal Constitution of the United States, Delaware occupies a small niche in the Boston-to-Washington urban corridor along the Middle Atlantic Seaboard. Its 2,057 square miles (5,328 square kilometres) make it, after Rhode Island, the smallest state in the Union, but the more than 548,000 people counted in the 1970 census make it one of the most densely populated states. Most of its people live in the north around Wilmington, where its industry is concentrated and where the major coastal highways and railways pass into and through it from Pennsylvania and New Jersey on the north and east into Maryland on the south and west. The rest of the state comprises the northeast corner of the Delmarva Peninsula, which it shares with Maryland and Virginia.

Overview of the state

Historically, geographically, and economically, Delaware has its closest ties with Pennsylvania, particularly Philadelphia, where the Delaware River and other transportation arteries have directed its commerce. The state's three counties—New Castle, Kent, and Sussex—had been established by 1680, and, except for periods during the Revolution and the Civil War, its history has been placid. Stability and conservatism became characteristics of Delaware, especially in the southern areas, which until 1964 had maintained a grip on political life vastly out of proportion to their population. As a result, old institutions were tenaciously preserved.

The manufacturing complex in the north makes Delaware one of the most industrialized of American states and the leading state in petrochemical production. The state is often thought of as being dominated by corporations, especially by the vast du Pont industrial empire, but the industrial wealth of the Wilmington area was balanced by the political overrepresentation of the agricultural downstate region until the mid-1960s. All factions have united to perpetuate liberal incorporation laws that encourage many major American businesses to make Delaware their nominal home state. Delaware ranks among the top ten states in the per capita income of its citizens. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; and NORTH AMERICA.)

THE HISTORY OF DELAWARE

The colony. The Dutch who established the first European settlement in Delaware at Lewes in 1631 were soon massacred by Indians, and it was not until 1638 that a permanent settlement was planted—by Swedes at Ft. Christina, now Wilmington; they reputedly erected America's first log cabins in this colony of New Sweden. The Dutch from New Amsterdam (New York) conquered the Swedes in 1655, and the English seized the colony from the Dutch in 1664. Thereafter, except for a brief Dutch reconquest in 1673, Delaware was administered as part of New York until 1682, when the Duke of York, acting as proprietor, ceded it to William Penn, who wanted it so that his colony of Pennsylvania could have access to the ocean. Though he tried to unite the Delaware counties with Pennsylvania, both sides resented union. In 1704 he allowed Delaware an assembly of its own. Pennsylvania and Delaware shared a royal governor until the Revolution. Only in 1776 did the name Delaware—deriving from Sir Thomas West, Lord de la Warre, a governor of Virginia—become official, though it had been applied to the bay in 1610 and gradually thereafter to the adjoining land.

Revolution and statehood. During the Revolution, Delaware was invaded by a British army en route to Philadelphia and was constantly menaced by British ships. The event best remembered, however, is the spectacular ride (July 1–2, 1776) of Caesar Rodney from his home to Philadelphia to break a tie in the Delaware delegation and cast Delaware's vote for independence. The proudest boast of Delaware is that its speedy ratification of the Constitution, on December 7, 1787, gave Delaware its right to be called "the first state."

As national political parties arose, Delaware became a Federalist state, adhering to the party of Alexander Hamilton and John Adams well into the 1820s. In the next period, Delaware became as fervently Whig as it had been Federalist.

19th-century political sentiments

Civil War and aftermath. The advent of the Civil War did not seriously tempt Delaware to secession. Nominally Delaware had been slave territory since its days as a Dutch colony, but the number of slaves had declined drastically, mainly through voluntary manumissions (grants of freedom), from 8,900 in 1790 to 1,800 in 1860. A more important consideration was Delaware's economic bond with Pennsylvania and the North, strengthened by the river trade and the new railroad network. Though Unionist in sentiment, this border state never voted for Lincoln, and the Reconstruction that followed the Civil War drove many Delaware voters to the Democratic Party in sympathy with the occupied South. By the end of the century, however, economic realities had reasserted themselves, and Delaware became firmly Republican and remained such until well into the Great Depression of the 1930s.

Economic growth. Wilmington, meanwhile, had become a manufacturing city so populous that by 1920 it contained just under half of the state's population and at least a similar proportion of its wealth and economic energy. Diversity characterized the products of Wilmington factories, but in the 20th century the city became renowned as an administrative centre for the nation's chemical industry. Primarily, this meant E.I. du Pont de Nemours & Company and two other powder makers, Hercules Inc. and Atlas Chemical Industries, Inc., all of which transferred their major energies to basic chemicals after World War I.

THE NATURAL AND HUMAN LANDSCAPE

Except for Florida, Delaware, located mainly within the Atlantic Coastal Plain, is the most low-lying state in the nation. A long sand beach forms the ocean front, stretching from the Maryland line, at Fenwick Island, to Cape Henlopen, at the mouth of Delaware Bay. Only one major break, Indian River Inlet, occurs along the 23-mile length of that beach. Much of the beach is a low bar between the ocean and a series of lagoons or shallow bays, but at Bethany Beach, near the southern boundary, and again at Rehoboth, near the northern end of the beach, the mainland reaches directly to the ocean.

Geographical regions and drainage

The shoreline of Delaware Bay is often marshy. The mouths of tributary streams like the Murderkill, the Mispillion, and the St. Jones are so shallow that no good harbours exist except at Lewes, just inside Cape Henlopen, where an artificial harbour protects shipping from Atlantic storms. Farther north, on the banks of the Delaware River, spots of high, dry land appear, as at Port Penn, New Castle, and Edge Moor; but the state's main port, Wilmington, lies on the Christina River, a tributary of the Delaware.

Most of Delaware is drained by streams that run eastward to the Delaware River, Delaware Bay, and the Atlantic, but the Nanticoke River and its tributaries in southwestern Delaware flow into Chesapeake Bay. So does the Pocomoke River, which drains the so-called Burnt Swamp, or Cypress Swamp, in the extreme south of Delaware athwart the Maryland line.

Most of the coastal plain terrain is fertile and level, seldom over 60 feet (18 metres) above sea level, but it becomes increasingly sandy to the south. Abundant

woodlands, streams, and freshwater ponds interrupt the monotony of the landscape. Occasionally, as at Odessa, villages appear suddenly at the side of the road, with no more warning than the sight of a church steeple.

Near its northern edge the plain is intersected by a great deep ditch, the Chesapeake and Delaware Canal, which has been deepened and straightened for ocean shipping. It shortens the water route between Philadelphia and Baltimore by several hundred miles and also brings Baltimore closer to the ocean than via Chesapeake Bay. The canal is popularly considered to be the boundary between agricultural downstate Delaware and the northern industrial region. Though the land on either side of it is strikingly similar, many Delawareans are convinced that even the weather changes at the canal.

The several high bridges over the canal, as well as the giant twin bridges crossing the Delaware north of New Castle and the stacks of the refinery at Delaware City, serve as the major landmarks on the horizon below the hillier northwest corner of the state. There, north and west of a 14-mile line between Wilmington and Newark, lie the rolling hills of a section of the Piedmont extending south from Pennsylvania. The significance of this area belies its size—less than one-fifteenth of the state's total area—for within it and at its edge dwells most of Delaware's population. The highest point in the state, near Centerville, is only 438 feet (134 metres) above sea level. Probably the most peculiar features are Iron and Chestnut hills, which protrude into the plain southwest of Newark and are scarred by open pits where iron ore once was mined.

The centre of Wilmington lies on hills sloping downward toward the intersection of the Christina and its major tributary, the Brandywine. Here, navigable water brought shipping close to falls that provided power for manufacturing. The railroads and highways, which followed this fall line along the East Coast, have kept Wilmington on major transportation routes between Philadelphia and Baltimore and have promoted the tendency for the urbanization of the once open land between Wilmington and other cities.

The climate of Delaware is like that of the rest of the Middle Atlantic area. August, which has the second hottest temperatures after July, is also the rainiest month, with an average precipitation of about 5.6 inches (142 millimetres), whereas February has the least precipitation, an average of just under three inches (76 millimetres). The yearly average precipitation is nearly 45 inches (1,143 millimetres).

THE PEOPLE OF DELAWARE

Patterns of immigration. As is characteristic of the Middle Atlantic states, the colonial population of Delaware was quite varied. Swedes (and the Finns who came with them), Netherlanders, and Africans had settled in Delaware before the English, mainly in present-day New Castle County in the north. The English settlers came not only from overseas but also from Pennsylvania and Maryland. Some of the settlers from Pennsylvania were Quaker artisans and merchants; the settlers from Maryland were often planters who brought slaves with them. With the English came some Welsh settlers and, after 1715, large numbers of Irish, particularly the Presbyterians of Scottish descent known as Scotch-Irish. Downstate Delaware was mainly settled by the English and by Africans.

After the Revolution, a small group of French came to New Castle and Wilmington from the West Indies, and a few, including the progenitors of the du Pont family, came from France. The middle of the 19th century brought a large immigration of Catholic Irish and Germans, and at the end of the century Italians, Poles, and Jews came in large numbers, accompanied by smaller groups of Ukrainians, Russians, Scandinavians, and Greeks. These newcomers settled mainly in Wilmington.

Demography. Though federal laws reduced the flow of immigrants after World War I, Delaware experienced its largest population growth in the middle of the 20th century. From 1950 to 1960 its population grew by about 40 percent, making Delaware the ninth state in density of population. The 23 percent growth during the 1960s was a rate exceeded by only seven other states, most of them in the West.

Many of the newcomers were highly skilled scientists or technicians. Wilmington also received a large influx of blacks, many of them unskilled. In 1950, 15 percent of Wilmington's inhabitants were black; by 1960, 26 percent, and by 1970, 44 percent. This growth rate did not match the exodus of whites to the suburbs over the same 20 years, when the population of Wilmington proper declined from 110,000 to 80,000.

Process of suburbanization

The suburbs of Wilmington, largely unincorporated, received not only the people fleeing Wilmington but most of the white newcomers to Delaware after 1945. By 1970 more than half of the state's population lived outside the city of Wilmington but within 15 miles of it. This suburban band included the second largest city, Newark, which nearly doubled in size during the 1960s to 21,000 inhabitants, partly because of geographic extension of its boundaries. Thus, suburbia in the early 1970s was the seat of Delaware's population, political power, and wealth. It generally was populated by whites, but with a few black enclaves. The descendants of the later immigrant groups were leaving the city as hurriedly as the Anglo-Saxon Protestants, imparting diversity to suburbia. All parts of Delaware, except Wilmington itself, have shared in the state's growth to some degree. The capital, Dover, third city in size, grew from 7,000 in 1960 to 17,000 in 1970, aided by some extension of its boundaries and by employment offered by new industries and the Dover Air Force Base.

Interesting ethnic groups in rural Delaware include Polish potato growers in Kent, who have come from Long Island; Italian mushroom growers at Hockessin; a colony of Finns that originated at Iron Hill after World War I; an Amish settlement near Dover; and the historic groups of mixed-bloods, called Moors and Nanticokes, at Cheswold, in Kent County, and beside Indian River Bay, in Sussex County.

THE STATE'S ECONOMY

Delaware's prosperity depends upon its favourable location: four of the ten largest cities in the United States lie within 150 miles.

Goods produced. Though the number of farmers continued to decline, agriculture remained viable in the early 1970s. Net income per farm was double the national average. More than half of the farmers' cash income comes from poultry raising, centred in Sussex County. Soybeans are increasing in importance, and other major agricultural products are corn, milk, and vegetables, especially lima beans and asparagus. The coastal and inland waters produce fish, clams, and crabs. The only mining is of gravel and sand.

Agriculture and manufacturing

The major economic enterprise in Delaware is manufacturing, especially the chemical industry. Wilmington boasts of being the chemical capital of the world because it is the centre of administration and research of several chemical companies; du Pont, Hercules, and ICI America (formerly Atlas) are the largest. Chief chemical products are pigments, nylon, and petrochemicals. Delaware also has automobile-assembly plants, an oil refinery, a synthetic rubber plant, packaging plants, textile mills, and various food-processing plants.

Taxation. In the public sector, many things are done by the state that elsewhere would be left to local government because of Delaware's small size. Consequently, state taxes and indebtedness are relatively high, whereas local equivalents are low. The largest source of state income is the tax on personal and corporate incomes. There is a state inheritance tax, which frequently produces a windfall when a very wealthy citizen dies.

The second most important source of state revenue is the corporation franchise tax. Delaware has made a business of incorporating companies, many of which operate primarily in other states, since early in the 20th century. It offers them favourable laws that are kept up to date to reflect changing business conditions, a con-

venient location, very moderate taxation, stable institutions, and a bench and bar with experience in corporate litigation.

There is no general sales tax in Delaware and no state property tax. Real estate taxes are, however, the chief support of county and municipal governments. Though the schools are supported chiefly by the state, school districts must raise part of the money for new buildings by property and other taxes, after approval by a referendum. They may also, again with voter approval, raise money to supplement the state appropriation for school operations, including salaries.

Transportation. The chief flow of highway traffic in Delaware is between Wilmington and its suburbs and the interstate traffic crossing northern Delaware between New York or Philadelphia and Baltimore or Washington. Slightly less important is the traffic up and down the state on the Du Pont Boulevard. The state maintains all roads and bridges as well as through streets in municipalities. A joint Delaware–New Jersey agency operates both the twin bridges across the Delaware River above New Castle and a ferry between Lewes and Cape May.

There is railroad passenger service between Philadelphia and Baltimore. Freight service is also available to the southern state line and in northern Delaware. Local bus transportation in the Wilmington area is provided by a public authority. Wilmington has a marine terminal and an airport.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. The constitution of Delaware was adopted in 1897 but has been amended many times. Amendments require a two-thirds vote in two successive legislatures, with an election intervening. The governor, who has no veto on amendments, serves a four-year term and may be re-elected only once. Traditionally, the legislature has been strong and the governor relatively weak, but adoption of the cabinet form of government in 1970 centralized and strengthened executive authority. The bicameral legislature is known as the General Assembly.

An unusual feature of Delaware's judicial system is the retention of the Court of Chancery, which handles equity cases involving civil rights and litigation concerning Delaware corporations. Most other states have merged their chancery into their law courts. The highest court is the Supreme Court, which hears appeals from the Chancery Court and the Superior Court. At the lowest level in the state judiciary are the magistrate courts, presided over by justices of the peace, who seldom are lawyers. All Delaware judges are appointed by the governor.

Weak county governments have been the rule. Each was headed by an elected levy court that sets the tax rate and appropriates funds. New Castle, however, now has a strong county government with an elected and powerful county executive and council and some appointed administrators. Unique to Delaware is the county subdivision known as a hundred, an ancient English governmental unit that has survived nowhere else in the United States. It no longer has a governmental or other function and is retained purely as a geographical name.

Politics. Democrats and Republicans have been fairly evenly matched in Delaware. The Republicans have won more of the recent elections, but the Democrats have the larger number of registered voters. Many decline to list party preference, and numerous swing votes may go to either side. Primaries are not of significance; major nominations are made at a convention.

The black voter

After the Civil War, Delaware Democrats used their control of such offices as assessor and tax collector to discourage blacks from qualifying as voters, but Republicans actively sought the black vote and, with its aid, won control of the state early in the 20th century. In 1932 the Democrats at last abandoned their all-white tradition. At first they won black votes only for the national ticket, but gradually, over the next two decades, Delaware blacks, like those in other Northern states, realigned themselves with the Democratic Party. Thereafter, only an exceptionally popular Republican was able to win black support. Bipartisan black support was largely responsible for passage of a fair-housing law.

The social milieu. Delaware maintained racial segregation by both custom and law until the 1950s, when court decisions began to strike down old laws separating the races as well as acts of discrimination in housing, public accommodations, and the like. Integration was accomplished fairly smoothly in the schools, including the University of Delaware and Delaware State College, the latter integrating whites into a formerly all-black student body. In Wilmington, however, school integration and fair-housing laws had the effect of speeding the desertion of the city by whites. Federal programs to assist the poor home buyer helped blacks move into the vacated housing; some of these blacks had been dispossessed by urban renewal projects. Even with such programs and the end of legal discrimination, social pressure and economic realities have served to maintain some degree of segregation in education and housing.

Social services and special schools

The state, meanwhile, was called upon to provide an increasing number of services for its citizens. The demands for expenditures, especially in education and welfare, were brought on partly by population growth and immigration of young families with children, partly by recognition of long-ignored needs. Kindergartens, schools for the handicapped, and mental-health clinics were new services offered. Delaware ranked in the upper half of the states in per capita expenditures for welfare and health services in 1970, but lower when these outlays were related to personal income.

The University of Delaware grew from 1,500 students in 1952 to more than 12,000 in 1971. Because the small population would make the establishment of a medical school a heavy expense, the state has arrangements with Jefferson Medical College of Philadelphia to save places in each class for Delaware students. For similar reasons it sets aside funds to help Delaware students go out of state for education in such fields as law and dentistry, in which no training is offered in Delaware's public institutions. In 1970 the state ranked 2nd in per capita expenditures for education, but only 9th in relation to personal incomes.

CULTURAL LIFE AND INSTITUTIONS

Museums and libraries. Two major museums are located in the outskirts of Wilmington. The Henry Francis du Pont Winterthur Museum is noted for its collection of American decorative arts, which are displayed in authentic period rooms. The Hagley Museum portrays the development of American manufacturing through preservations of the early mills and other structures of the du Pont company, as well as by indoor exhibits. The Delaware State Museum, in Dover, has a wide range of exhibits; while local history is the theme of the Zwaanendael Museum at Lewes.

Historic buildings, houses and colonial capitol

A number of historic houses are permanently open to the public, including the John Dickinson Mansion, near Dover; the Parson Thorne Mansion, in Milford; and several houses in Odessa and New Castle. Several blocks in New Castle surrounding the colonial capitol, known as the Old Court House, remind visitors of the restorations of colonial Williamsburg in Virginia—except that in New Castle very few buildings had to be restored. Immanuel Episcopal Church, on the Green, was begun in 1703; its graveyard has many interesting stones. The Presbyterian Church across the street dates from 1707. No buildings survive from the Dutch period. Old Swedes Church in Wilmington was built in 1698 for a Swedish Lutheran congregation, but it is now Episcopalian. The Swedes brought a tradition of log construction to the New World, but none of their work remains except, perhaps, portions of a few small log structures.

The state's foremost research library is that of the University of Delaware. The major public library, the Wilmington Institute Free Library, serves all of New Castle County. The Delaware State Library Commission serves the lower counties; and most towns also support a library of their own. Among the specialized libraries, the Eleutherian Mills Historical Library, featuring business and

industrial history, and the library division of the Winterthur Museum, specializing in the decorative arts and crafts, are internationally famous.

The arts and communications. Wilmington has been the centre of a distinguished group of illustrators, many of them pupils, directly or indirectly, of Howard Pyle, whose work is displayed at the Delaware Art Museum. N.C. Wyeth, a pupil of Pyle, made his home just over the Pennsylvania line at Chadds Ford, which members of his family have made famous as the home of the Brandywine school, a group of mainly genre and narrative painters.

Wilmington has one legitimate theatre, the Playhouse, in which professional companies appear; there are also numerous amateur groups. The small village of Arden is remarkable for its theatrical traditions, both amateur and professional, which include annual productions of Gilbert and Sullivan operettas and, until recently, of Shakespeare by the townspeople.

Of three daily newspapers, the larger two are published in Wilmington and are controlled by the Christiana Securities Company, a Du Pont family holding company. There are many radio stations but only one television station, an educational outlet that has studios both in Wilmington and Philadelphia. Most Delawareans tune in Philadelphia television stations, a continuation of the historic Delaware pattern of looking toward Philadelphia.

Recreation. Delaware's ocean beaches are popular not only with Delawareans but also with people from neighbouring areas, especially Washington, D.C. Rehoboth and Indian River bays are boating, fishing, and clamming centres. State parks, such as Lum's Pond, are increasingly attractive for recreation. The week-long Delaware State Fair is held annually in Harrington. Pari-mutuel betting lures crowds to racetracks in Stanton, Dover, and Brandywine Hundred.

PROSPECTS

The 1897 constitution of Delaware overtly discriminated against Wilmington in the allocation of representatives and senators in the General Assembly. At the time, the city had almost half of the state's population, but it received only two of the 17 senators and five of the 35 representatives. Since only the legislature could change the constitution, the shackles on Wilmington—on its wealthy manufacturers and merchants as well as on its immigrant labourers—seemed to be fastened forever. By 1964, when the United States Supreme Court declared against such practices, such populous upper middle class suburbs as Brandywine Hundred, with 58,000 people, and Christiana Hundred, with 48,000, had no more representation than rural Blackbird Hundred with 1,600, or Appoquinimink Hundred with 2,500. The resultant reapportionment profited Wilmington, with its falling population, very little. The suburbs, however, came into their own, a majority of the seats in each chamber going to New Castle County outside Wilmington.

This sudden acquisition of political power by suburban Delaware, in addition to the state's continuing rapid growth in population and the changing character of Wilmington, will unquestionably be critical factors in Delaware's future. The new suburban representatives appear to speak for new population elements in the state, replacing in large measure the representatives of the long-established rural citizenry who tended to seek preservation of traditional ways, and for white-collar rather than blue-collar attitudes. Seats in the party nominating conventions were also reapportioned, and the first changes following this saw election of a research chemist as governor and a corporation lawyer as United States senator.

In state government the new spirit was shown in a reorganization for efficiency and a willingness to spend money to this end and in an interest in ecology demonstrated by restrictions on waterfront industry and efforts to protect the beaches. How far the electorate will allow greater advances in educational and other social services is a question yet to be answered in full.

BIBLIOGRAPHY. WILLIAM H.A. CARR, *The du Ponts of Delaware* (1964), the most recent popular history of the family; ALFRED D. CHANDLER and STEPHEN SALSBURY, *Pierre S. du Pont and the Making of the Modern Corporation* (1971), a serious, thorough study; PAUL DOLAN, *The Government and Administration of Delaware* (1956), though old, the best account of the state government; FEDERAL WRITERS' PROJECT, *Delaware: A Guide to the First State*, new and rev. ed. by JEANETTE ECKMAN (1955), an excellent guide that includes valuable introductory chapters on history, education, and cities; HAROLD HANCOCK, *Delaware During the Civil War: A Political History* (1961); and JOHN A. MUNROE, *Federalist Delaware, 1775–1815* (1954), two studies of Delaware at critical historical periods, and *Delaware: A Student's Guide to Localized History* (1965), short paperback history; H. CLAY REED (ed.), *Delaware: A History of the First State*, 3 vol. (1947), the most recent comprehensive history, with the first two volumes containing many excellent essays on specific topics; and compiled with MARION BJORNSON REED, *A Bibliography of Delaware Through 1960* (1966), authoritative and comprehensive; C.A. WESLAGER, *Delaware's Forgotten Folk: The Story of the Moors and Nanticokes* (1943), a study of mixed-blood groups, and *The English on the Delaware, 1610–1682* (1967).

(J.A.Mu.)

Delhi

Delhi is the third-largest city of India, surpassed in size only by Calcutta and Bombay. New Delhi, the capital of the Indian union, lies to the south of Delhi (popularly known as Old Delhi). The city is situated in north central India and stands on the west (right) bank of the Yamuna River, a tributary of the Ganges. It is located within the union territory of Delhi, which covers an area of 573 square miles (1,483 square kilometres). The union territory is bounded on the east by the state of Uttar Pradesh and on the north, west, and south by Haryana. In 1971 the union territory had a total population of over 4,040,000, of which Old Delhi's population amounted to about 3,280,000 and New Delhi's population about 290,000. It is generally presumed that the city was named after Raja Dhilu, a king of the Mauryan dynasty who reigned in the 1st century BC, and that the various names by which it has been known (Delhi, Dehli, Dillī, and Dhillī) are corruptions of this name. Delhi is a focal point in India's transportation network. It stands about 100 miles to the south of the Himalayan mountain ranges.

Delhi has been the capital city of a succession of mighty empires or powerful kingdoms; ruins mark the sites of the various cities, both ancient and medieval. According to popular tradition, the city has changed its locality seven times, although some authorities, who take smaller towns and strongholds into account, claim it has changed its site 15 times. All these locations are confined to a triangular area of about 70 square miles called the Delhi triangle. Two sides of this triangle are represented by the rocky hills of the Arāvalli Range in the west and south and the third by the shifting channel of the Yamuna. The present site of Delhi is bounded to the west by a northern extension of the Arāvalli Range known as the Delhi Ridge.

The Delhi triangle

HISTORY

The earliest reference to a settlement at Delhi is to be found in the famous epic *Mahābhārata* (a narrative about the descendants of the prince *Bhārata*), which mentions a city called Indraprastha, built about 1400 BC somewhere between the sites where the historic Purāna Qal'ah (Old Fort) and Humāyūn's tomb were later to be located. The first reference to the place-name Delhi, as already mentioned, seems to have been made in the 1st century BC, when Raja Dhilu built a city near the Quṭb Mīnār (Quṭb Tower) site and named it for himself. Thereafter Delhi faced many vicissitudes and did not reemerge into prominence until the 12th century AD, when it became the capital of the Cāhamāna ruler Pṛthvīrāja. After the defeat of Pṛthvīrāja the city passed into Muslim hands. Quṭb-ud-Dīn Aybak, founder of the Mu'izzī (Slave) dynasty and builder of the famous tower Quṭb Mīnār (completed in the early 13th century), also chose Delhi as his capital.

ʿAlāʾ-ud-Dīn Khaljī (1296–1316) built the second city of Delhi at Siri, three miles northeast of the present Delhi. The third city of Delhi was built by Ghiyās-ud-Dīn Tughluq (1320–25) at Tughlakābād but had to be abandoned in favour of the old site near the Quṭb Mīnār because of scarcity of water. Muḥammad ibn Tughluq (1325–51) extended the city farther northeast and built new fortifications around it. It then became the fourth city of Delhi, under the name Jahānpanāh. These new settlements were located between the old cities near the Quṭb Mīnār and Siri Fort. Muḥammad ibn Tughluq's successor, Fīrūz Shāh Tughluq (1351–88), however, abandoned this site altogether and in 1354 moved his capital farther north near the ancient site of Indraprastha and founded the fifth city of Delhi, Fīrūzābād, which was situated in what is now the Kotla Fīrūz Shāh area.

Invasion of Timur

After the invasion of Delhi by Timur (Tamerlane) in the latter half of the 14th century, the last of the sultan kings moved the capital to Āgra, so that Delhi experienced a temporary diminution in its importance. Bābur, the first Mughal ruler, re-established Delhi as the seat of his empire in 1526. His son Humāyūn built a new city, Dīn Panāh, on the site between Kotla Fīrūz Shāh and the Purāna Qalʿah. Shēr Shāh, who drove Humāyūn from the country in 1540, razed Dīn Panāh to the ground and built his new capital, the Shēr Shāhī (Purāna Qalʿah), as the sixth city of Delhi.

Delhi later again lost importance when the Mughal emperors Akbar (1556–1605) and Jahāngīr (1605–27) moved their headquarters to Fatehpur Sīkri and Āgra in the late 16th and early 17th centuries, but the city was restored to its former glory and prestige in 1638, when Shāh Jahān, Akbar's grandson, laid the foundations of the seventh city of Delhi, Shāhjahānābād, which is presently known as Old Delhi. The greater part of the city is still confined within the space of Shāh Jahān's walls, and several gates built by him— the Kashmir Gate, the Delhi Gate, the Turkmen Gate, and the Ajmer Gate—still stand.

With the fall of the Mughal Empire during the mid-18th century, Delhi again faced many vicissitudes—raids by the Marāthās (a people of peninsular India), the invasion by Nāder Shāh of Persia, and a brief spell of Marāthā rule—before the arrival of the British in 1803. Under British rule the city flourished, except during the Indian Mutiny in 1857, when the mutineers seized the city for several months, after which British power was restored, and Mughal rule ended. In 1912 the British moved the capital of British India from Calcutta to the new city of New Delhi, then adjacent to Delhi and now a part of it.

THE LANDSCAPE

Elevation and area. The union territory of Delhi lies at an altitude of between 700 and 1,000 feet above sea level. Out of its total area of 573 square miles (the union territory of Delhi), Old Delhi occupies 93 square miles, Delhi Cantonment 17 square miles, and New Delhi 16 square miles.

Climate and environment. The climate of Delhi is characterized by extreme dryness, with intensely hot summers and cold winters. It is associated with a general prevalence of continental air, which moves in from the west or northwest, except during the season of the monsoon (rain-bearing wind), when an easterly to southeasterly influx of oceanic air brings increased humidity. The summer season lasts from mid-March to the end of June, with average maximum and minimum temperatures of 97° F (36° C) and 77° F (25° C); it is characterized by frequent thunderstorms and squalls, which are most frequent in April and May. The monsoon season, following the hot summer, continues until the end of September, with an average rainfall of about 25 inches. The post-monsoon period of October and November constitutes a transition period from monsoon to winter conditions. The winter season extends from the last week of November to mid-February; average maximum and minimum temperatures are 70° F (21° C) and 52° F (11° C), respectively. The air in Delhi is dry for most of the year, with very low relative humidity from April to June and markedly higher humidity in July and August, when weather conditions are oppressive. The mean daily temperature is highest in May, and the monthly mean temperature is highest in June, when the night temperature is also at its maximum. The mean daily temperature may rise as high as 110° F (43° C). The coldest month is January, when both the mean maximum temperature and the mean minimum temperature are at their lowest—70°F (21° C) and 45° F (7° C), respectively.

Temperature and rainfall

Air and water pollution do not cause serious problems, although sometimes a temperature inversion (which often occurs when a warm air mass remains over a land surface that cools during the night) forms in the winter months, which traps pollutants (coal dust and soot) and increases contamination considerably. In the past, the danger of water pollution caused by drainage water mixing with the city's water supply in summer months was a serious problem. Since 1960, steps have been taken, however, to prevent this situation.

Vegetation and wildlife. The natural plant cover in the Delhi area varies according to the physical features with which it is associated. The ridges and hillsides abound in thorny trees, such as acacias. During the monsoon season, herbaceous species grow in profusion. The shisham, or sissoo, tree, which yields a dark-brown and durable timber, is commonly met with in the Bangar (Plain) area of the union territory. Riverine vegetation, consisting of weeds and grass, occurs on the banks of the Yamuna. New Delhi is famous for its avenues of flowering shade trees, such as the neem (a drought-resistant tree with a pale-yellow fruit), jaman (a tree with an edible grapelike fruit), mango, pipal (a fig tree), and shisham. It is also famous for numerous flowering plants, which provide a splash of colour during the winter. These include a large number of multicoloured seasonals: chrysanthemums, phlox, violas, and verbenas. The transition from winter to spring is very gradual, and only the flowers can testify to changing conditions, with chrysanthemums in December yielding place to roses in February.

The avenues of shade trees

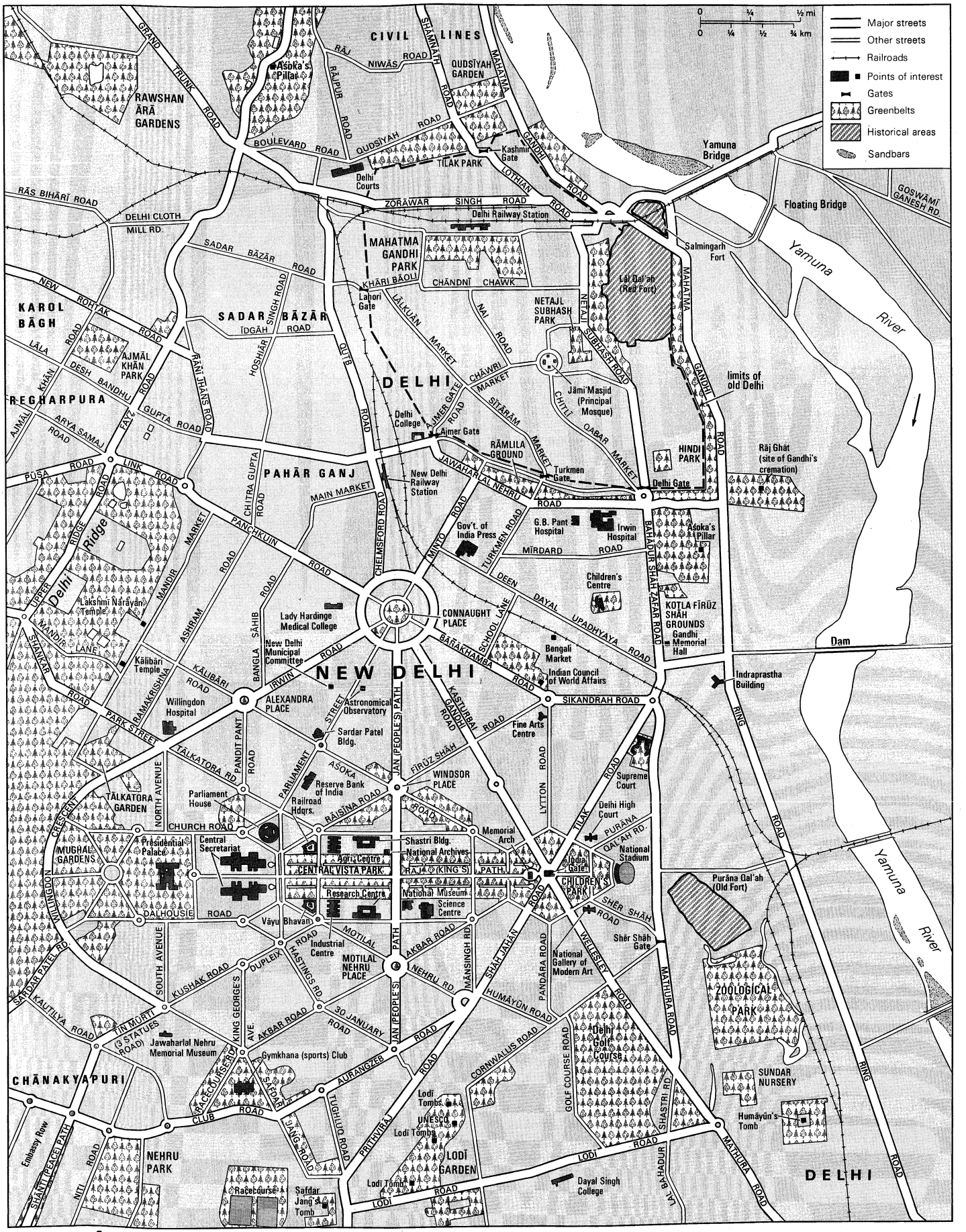
The animal life of the union territory, like its plant life, is quite diverse. Among carnivorous animals are leopards, hyenas, foxes, wolves, and jackals, which inhabit the jungles, low forests, and hilly ridges. In some places along the bank of the Yamuna, wild boars are found. Monkeys are not uncommon. Bird life includes partridge (gray and black), rock pigeons (black and blue), parrots, and bush quail. Peafowl are numerous on the hilly ridges. The Yamuna abounds in fish, and an occasional crocodile may be found.

The city plan. The city plan of Delhi is a mixture of contrasting old and new road and circulation patterns. The contrast between the convoluted form of the old city and the diagonal features of the modern traffic arteries in New Delhi is particularly striking.

The street pattern of Old Delhi reflects some of the older requirements of defense, with a few transverse streets leading from one major gate to another. Occasionally a through street from a subsidiary gate leads to the main axes. The other Old City streets tend to be irregular in direction, length, and width and are suitable only for pedestrian traffic. Thus, the pattern as a whole consists of a confusing mixture of narrow and winding streets, culs-de-sac, alleys, and byways giving access to residences and commercial areas. The population density—about 35,000 persons per square mile—is high.

In sharp contrast to the Old City, the Civil Lines (residential areas originally built by the British for senior officers) in the north and New Delhi in the south present an aspect of relative openness, characterized by green grass and trees, order, and quiet.

When the decision was made in 1911 to transfer the capital of India from Calcutta to Delhi, and a town planning committee was formed, a site was chosen three miles south of the existing city of Delhi, around Rāisīna Hill. This was a well-drained, healthy area between the ridge and the river that provided ample room for expansion. The Rāisīna Hill, commanding a view of the entire area, stood 50 feet above the plain, but the top 20 feet



Central Delhi–New Delhi.

Based upon Survey of India map with permission of the Surveyor General of India; © Government of India
Copyright 1969

were blasted off to make a level plateau for the major government buildings and to fill in depressions. With this low acropolis as the focus, the plan was laid out.

The New Delhi plan was characterized by wide avenues with trees in double rows on either side, creating vistas and connecting various points of interest. Almost every major road has a specific focal point closing the vista so that no avenue is lost in the horizon. Besides the diagonal road pattern, the most prominent feature of the entire plan is the Central Vista Park, starting from the National Stadium in the east, continuing through the memorial arch and the Central Secretariat, and culminating in the west at Rāshtrapati Bhavan (the Presidential Palace). This is the main east–west axis; it divides the New City into two parts, with the fashionable shopping centre, Connaught Place, in the north and extensive residential colonies in the south.

The Central Vista Park

Land use. The pattern of land use in Delhi has been changing with the implementation of the Delhi Town Planning organization's master plan. Broadly, the public and semipublic land use is confined to the Central Secretariat area of New Delhi and to the Old Secretariat area in the Civil Lines, with subsidiary centres developing in the Indraprastha Estate (an office complex) in the east and in Ramakrishnapuram (an office-cum-residence district) in the south. A large number of small manufacturing establishments have entrenched themselves in almost every part of the Old City, but the main industrial areas are concentrated along the Najafgarh Road, in the west, and on Kalkaji Road, in the south, where a large planned industrial estate (Okhla) has been established. Areas for commercial land use are confined to Chāndnī Chawk (Silver Street) and Khāri Bāoli (both in the north), the Sadar Bāzār of Old Delhi, the Ajmāl Khān Road of Karol Bāgh in west Delhi, and the Connaught Place area of New Delhi. A number of district and local shopping centres have also developed in other localities. The University of Delhi is located in the north, where a number of educational institutions for college education and for higher studies are located. Another educational complex is growing in south Delhi (see below *Recreation*).

Traditional city neighbourhoods and regions. In a city such as Delhi, which bears the impress of history, there is a clear distinction between areas where indigenous influences are uppermost and areas characterized by colonial and postindependence influences. Although the social structure of Delhi has changed from coherence to a heterogeneity that is in keeping with its position as the national capital, certain residential neighbourhoods in the Old City, in the Civil Lines, in government housing areas, and in recent-growth areas have acquired a specific character of their own. In the Old City, there is a strong *mohalla* ("neighbourhood") feeling, partly induced by its peculiar housing layout. Here gates or doorways open onto private residences and courtyards or onto *katras* (one-roomed tenements facing onto a courtyard or other enclosure and having access to the street by only one opening or gate). The Civil Lines area consists of upper-income-group residences. The government housing areas also exhibit segregation by income groups. In some recent growth areas, "mixed neighbourhoods" have been created. Chānakyapuri, with its concentration of foreign embassies, represents a microcosm of international architecture. Cultural "islands" are formed in such areas as the Bengali Market area or Karol Bāgh; Karol Bāgh, for example, is characterized partly by Bengali and partly by South Indian culture, although cultural distinctiveness is being eroded as other city residents move in. The Regharpura area, inhabited by Camārs (persons engaged in shoemaking and leather tanning), offers an interesting spectacle of an Indian ghetto. Similarly, the Kotla Mubārakpūr area is an example of a village enclave in the urban matrix: houses and streets retain their rural characteristics, though residents have urban employment.

THE CONTEMPORARY CITY

Transportation. The geographical position of Delhi on the great plain of India, where the Deccan tableland and the Thar Desert (*q.v.*) approach the Himalayas, producing a narrow corridor, ensures that all land routes from northwestern India to the eastern plain must pass through it, thus making it a pivotal centre in the subcontinent's network of transportation. Five national highways—connecting with Amritsar to the north, Āgra to the south, Jaipur to the southwest, Rohtak to the west, and Lucknow and Calcutta to the east—converge on Delhi. Four broad-gauge railway lines and one metre-gauge line also meet there, linking the city with all parts of the country. It is also the most important air terminal in northern India for both domestic and international air services, having two airports, at Palam and Safdar Jang. Palam airport is one of the hubs of the international airway system and is located far away from the main city. Safdar Jang was originally on the southern fringe of the city but is now in the midst of the built-up area. It is used mainly by the Delhi Flying Club.

The traffic-circulation pattern within a city that was designed for a smaller population now poses serious problems. Slow-moving bullock carts and bicycles, together with motorcars and trucks, create traffic snarls that are aggravated by peak-hour conditions. Mass-transportation facilities are inadequate. Proposals have been made for a comprehensive traffic-development program, including the construction of a ring railway.

Demography. Delhi's population of 240,000 in 1911 had increased nearly tenfold, to 2,350,000, by 1961. Between 1911 and 1941 its growth was comparatively slow, but between 1941 and 1951 the population doubled, due mainly to the immigration of a large number of refugees into the city at the time of independence. Between 1961 and 1971, a growth rate was maintained that was substantially higher than those of the other metropolitan cities in the country. The population of the union territory in 1971 was 4,044,000, including about 3,630,000 living in the urban area.

The refugee immigration

The average gross density of population of the city, which was 35,000 persons per square mile in 1971, does not suggest overcrowded conditions. But there were marked intra-urban differences, with the Old City Sadar–Pahār Ganj area having a gross density of well over 100,000 persons per square mile in 1971, while New Delhi and the Cantonment Area have densities of only 18,000 and 3,300 persons per square mile, respectively.

The population structure of Delhi in 1971 revealed a sex ratio of 798 females to 1,000 males. The younger age groups (less than 18 years old) constituted nearly half of the population in 1961 (49%); adult males and females constituted 29% and 18%, respectively in the same year; while the older age group constituted only 4%. Of the total population, 53% were unmarried. The literacy rate in 1971 was 59%. The working force constituted 30% of the population.

The composition of Delhi's population reflected its truly cosmopolitan character, with 56% of its inhabitants in 1961 coming from outside the territory. Of these, 36% were immigrants from other Indian states, 19% were immigrants from adjacent countries, and only 0.02% consisted of resident foreigners. The religious composition of the population was also varied. In 1961 it included Hindus (84%), Sikhs (about 8%), Muslims (almost 6%), Jains (1%), Christians (1%), and Buddhists (0.2%).

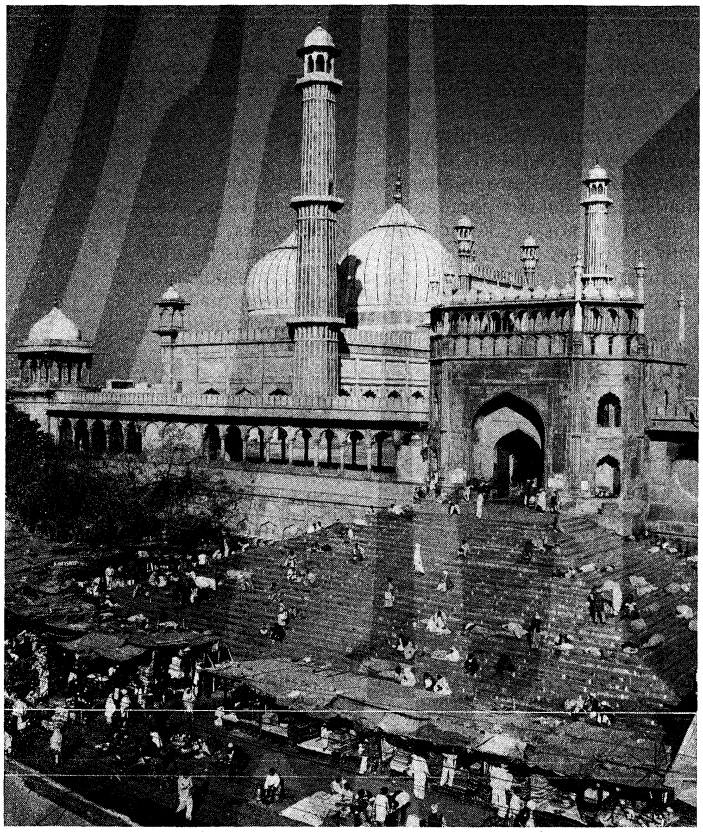
Housing. The housing situation in Delhi deteriorated after 1947 as a result of the influx of refugees and the city's emergence as the national capital of India. Building activity was insufficient to close the gap or to keep pace with the increasing population. This compelled nearly a third of the city's population to seek shelter in congested areas and in unauthorized dwellings or to settle as squatters in slums. The Delhi master plan estimated the housing shortage in 1961 at 150,000 units, a figure that later greatly increased, the shortage being most acute for low-income groups.

The houses in Old Delhi are unplanned, consisting of old structures of two, three, or more stories with a high proportion of single-room dwelling units. In the Civil Lines area, there are a number of old one-story bungalow-type houses. In New Delhi the government housing colonies have been laid out in a lavish manner and are

grouped on an income basis. The earlier constructions consisted of one-story houses, but later multistoried structures were built.

The implementation of the housing program is in the hands of various agencies, such as the union government, the Municipal Corporation of Delhi, the New Delhi Municipal Committee, and the Delhi administration, and various individuals and cooperatives.

W. Suschitzky



Market at one of the entrances to the Jāmiʿ Masjid (Principal Mosque) in Old Delhi.

Delhi's monuments

Building types and architectual features of note. There is perhaps no city in India that can compare with Delhi in the number of its monuments. These edifices illustrate the types of Indian architecture from the time of the imperial Gupta dynasty 1,600 years ago to the style of such architects as Sir Edwin Lutyens and Sir Herbert Baker, which was in evidence in New Delhi during the period of British rule. Delhi is particularly rich in material for the study of Indo-Muslim architecture. The monuments of the early Pathan style (1193–1320)—represented by the Qūwat-ul-Islām mosque, the Quṭb Mīnār, the tomb of Iltutmish, and the ʿAlāʾī Darwāzah ("gate")—reveal the adoption and adaptation of Hindu materials and style to Saracenic motifs and requirements. The later Pathan styles represented in Tughlakābād and tombs of the Sayyid Kings (1414–51) and Lodī kings (1451–1526) are characterized by finer domes and decoration and the use of finer marbles and tiles. The later Mughal architecture represented in the Red Fort and the Jāmiʿ Masjid (Principal Mosque) reveals an increasing use of marble, overelaboration with florid decoration, and the construction of bulbous domes and lofty minarets.

The Red Fort (Lāl Qalʿah) is one of the most important buildings of the city. Its massive red-sandstone walls, 75 feet in height, enclose a complex of palaces, gardens, military barracks, and other buildings. The two most famous of these are the Dīuān-e ʿĀmm and Dīuān-e Khāṣṣ. The Dīuān-e ʿĀmm (Hall of the Public Audience) has 60 red-sandstone pillars supporting a flat roof. The Dīuān-e Khāṣṣ (Hall of the Private Audience) is smaller and has a pavilion of white marble.

The architectural styles in the British period, represented by the Central Secretariat, Parliament House, and Rāshtrapati Bhavan (formerly the British viceroy's palace, now the residence of the president of India), combine the best features of the modern English school of architecture with traditional Indian forms. In the post-independence era, public buildings in Delhi show a utilitarian bias and a search for a synthesis between Indian and Western styles; the attempt, however, is not always successful, as is evident from the Supreme Court building, Vigyān Bhavan (a conference hall), and the government *bhavans* (buildings enclosing public areas). The buildings of the Bāl Bhavan (a childrens' centre) and Rabindra Bhavan (a fine-arts centre) show a trend toward a new style, using modern materials. The prevailing style, in both public and private buildings, emphasizes functionalism and humanism. Along the Yamuna River front, memorials set in flowering gardens have been built for such 20th-century national leaders as Mahatma Gandhi, Lal Bahadur Shastri, and Jawaharlal Nehru.

Economic life. Workers in the Delhi Union Territory comprise 50% of the males and about 5% of the females. In the economy of Delhi, the service sector comes first in importance (employing 22% of the total working population), the industrial sector is second (9%), and the agricultural sector is third (0.4%). The general structure of Delhi's occupational specialization can be seen from the fact that among the individual occupational categories "other services" holds the first place (46%), manufacturing comes second (21%), and trade and commerce third (19%). A substantial proportion of Delhi's working population is engaged in various services, including public administration, the professions, the liberal arts, and various miscellaneous personal, domestic, and unskilled-labour services. As a trading and commercial centre, Delhi has held a dominant position in northern India for many centuries. In modern times, it has also become a manufacturing centre with small- and medium-scale industries.

The city's role as a manufacturing centre

Traditionally, Delhi has been famous for its artistic work, such as ivory carving and painting, gold and silver embroidery, decorative ware, copperware, and brassware. More recently, it has become important for the manufacture of various sophisticated products in small-scale industry, such as electronics and engineering goods, automobile parts, precision instruments, lathes and drilling machines, and electrical appliances. Delhi has the largest number of establishments for the manufacture of electronics goods in the country. Hosiery items, ready-made garments, sports and leather goods, handloom products, and handicrafts are also produced. There are two modern industrial estates for small-scale industries—the Okhla Industrial Estate and the Badli Industrial Estate, which is a rural project.

Delhi's position as the national capital has accentuated its function as a banking, wholesale-trade, and distributive centre. It is the headquarters of the Reserve Bank of India and of the regional offices of the State Bank and other banking institutions. It is also a divisional headquarters for the insurance business and an important stock-exchange centre. In various distributive trades, such as those for cloth, dry fruits, spices and herbs, hosiery, and general merchandise, Delhi holds an important position. It also has a large share in the regional market for bicycles, fresh fruits and vegetables, furs, skins, and wool, motor parts and machinery, and iron and steel. Most of the distributive trades are carried on from within the Old Delhi area, where most of the markets are located in close proximity to each other.

Delhi's economic base is sound; this fact is reflected in its high per capita income of 872 rupees in 1960–61, as compared with the average per capita income of the country as a whole, for the same year, which was 330 rupees.

Political and governmental institutions. Delhi was a chief commissioner's province when India attained independence in 1947. It became a Part C (*i.e.*, centrally administered) state in 1952, but, as a result of the recommendations of the States Reorganisation Commission, its

The administration

status as a Part C state was abolished, and it became a union territory under the central government in 1956. A unified corporation for both urban and rural areas was established early in 1958. The administrative system was further modified by the Delhi Administration Act of 1966. Under the present arrangement, Delhi has a three-tier administration consisting of a lieutenant governor and his executive council, an elected metropolitan council, and the municipal corporation. The lieutenant governor is appointed by the president of India and is the chief administrator of the union territory of Delhi. He is assisted by an executive council of four members (headed by a chief executive councillor), which is also appointed by the president. The metropolitan council is a purely deliberative body. The municipal corporation is an elected local body, having under its control all statutory autonomous bodies, except the New Delhi Municipal Committee, the Delhi Cantonment Board, and the Delhi Development Authority. The New Delhi Municipal Committee is a body nominated by the central government. The Cantonment Board consists of partly nominated and partly ex officio members.

Public utilities. Water supply, drainage, sewerage, and conservancy and scavenging services are obligatory functions of the municipal corporation. Such functions as city transportation and the generation and distribution of electricity, though not obligatory, are performed by the corporation as activities in the public interest. Three statutory undertakings—the Delhi Water Supply and Sewerage Undertaking, the Delhi Electric Supply Undertaking, and the Delhi Transport Undertaking—perform these functions.

Water supply in Delhi is far from adequate, in spite of the fact that it has been augmented several times. The Yamuna River, the main source of supply, is practically dry during the summer months. Underground water has generally been found to be brackish in the territory; Delhi, therefore, must depend for part of its needs upon the adjoining states. The average per capita supply was 40 gallons per day in the early 1970s.

The sewerage of Delhi needs much improvement, both by way of extension to new areas and by the expansion of its capacity in older areas. Arrangements for the treatment of sewage are inadequate.

Delhi's electric-power supply depends on power generated by the local coal-burning thermal stations, augmented by a bulk supply received from the Bhākra-Nangal grid system. The supply of power has always lagged behind demand, with the result that industries have not been able to operate at full capacity.

Health and safety. *Health facilities.* In 1970 there were 50 hospitals in Delhi, with a total of 9,500 beds. Besides the hospitals, there were almost 90 allopathic dispensaries, 15 Ayurvedic and Unanī (*yūnānī*) dispensaries (*i.e.*, practicing indigenous systems of medicine that use mostly herbs and minerals), and six homeopathic dispensaries. Among the larger institutions are the All India Institute of Medical Sciences, the Willingdon Hospital, and the Safdar Jang Hospital, all administered by the central government, and the Irwin Hospital and G.B. Pant Hospital, run by the Delhi administration. The central government has its own health scheme for its employees.

Fire and police services. The Delhi Fire Service, with 13 fire stations located in various parts of the city, is under the control of the municipal corporation. Its jurisdiction extends over both the urban and rural areas of the union territory. In the rural areas, temporary stations are opened during the summer.

The Delhi Police Service is under an inspector general of police of the Delhi administration. The city is divided into four police districts, each of which is under a superintendent of police. There are about 80 police stations or posts in the city.

Education. There were almost 1,940 schools (including about 730 primary schools, 380 middle schools, and 470 higher secondary schools) in urban Delhi in 1969–70, with an enrollment of nearly 800,000 children. A Board of Higher Secondary Education regulates school educational standards. Among the institutions of higher learning is the University of Delhi (established 1922), which is an affiliating and teaching institution. In 1970 the university had an enrollment of 59,000 students, including 26,000 women. There are 39 colleges of arts, science, and commerce, including research institutions, affiliated to the university. There are also almost 30 colleges for professional and other studies. The most important among them are the Indian Agricultural Research Institute, the Indian Institute of Technology, and the All India Institute of Medical Sciences. Another university campus, the Jawaharlal Nehru University, was being developed in the early 1970s.

The University of Delhi

Cultural life. Delhi's cultural life has been considerably influenced by the cosmopolitan character of its population, which comes from different parts of India and the world and possesses variegated cultural backgrounds. The city's cultural life may be said to be still evolving. With the younger generation, cultural activities of earlier days, such as dancing, music, and poetry forums (*mushā-'ira*), are yielding place to frequenting the cinema, the cabaret, and clubs. At the same time there are also some theatre groups and institutions that foster indigenous literature and fine arts. Among the national cultural institutions are the Sangeet Natak Akademi, which promotes such cultural activities as the dance, drama, and music; the Lalit Kala Akademi, which promotes study and research in painting, sculpture, architecture, and applied arts; and the Sahitya Akademi, which fosters and coordinates literary activities in all the Indian languages. The National Gallery of Modern Art has works by nearly 100 artists.

The media. Press, television, and radio establish effective contact with Delhi's citizens. More than 1,100 newspapers and periodicals were published in the city in 1968, mostly in Hindi or English; these included 27 dailies, 180 weeklies, and about 600 monthlies. There are four radio stations. Television broadcasts, available on only one channel, were inaugurated in 1959 and can be seen by viewers within a range of 19 miles.

Recreation. Delhi is a city of gardens and fountains, with many places for recreation. Among the major recreation areas being developed in the early 1970s were the ridge (the principal green space of the city) and the Yamuna riverfront. There are also a number of parks and gardens, notably the parks at Okhla, Ḥawẓ-e Khāṣṣ, and Quṭb; the Rawshan Ārā Garden (developed in the Japanese style), and a rock garden in Patel Nagar. Delhi also has a Zoological Park situated on a 250-acre site. The National Stadium and the Kotla Fīrūz Shāh Grounds are important sports centres. An ambitious scheme for constructing swimming pools, stadia, and sports complexes is being pursued by the Delhi administration.

BIBLIOGRAPHY. For an authentic work on the history and places of interest of Delhi, see PRABHA CHOPRA (ed.), "Delhi: History and Places of Interest," *Delhi Gazetteer* (October 1970). One of the earliest historical works is G.R. HEARN, *The Seven Cities of Delhi*, 2nd ed. (1929). A. CUNNINGHAM, *Archaeological Survey of India 1862–65*, vol. 1 (1871), presents the report of the Archaeological Survey of India. V.K.R.V. RAO and P.B. DESAI, *Greater Delhi: A Study in Urbanisation, 1940–1957* (1965), is a socio-economic survey of Delhi conducted by the Delhi School of Economics. A. BOPEGAMAGE, *Delhi: A Study of Urban Sociology* (1957), deals with the sociological aspects of planning for metropolitan Delhi. The INSTITUTE OF TOWN PLANNERS (DELHI), "Papers of the Seminar on National Capital Planning and Development," *Journal of the Institute of Town Planners (India)*, no. 58 (March 1969), is a collection of papers covering various problems in the restructuring of the Delhi region.

(V.L.S.P.R./K.V.Su.)

Demography

The first person to use the word demography was the Frenchman Achille Guillard, who wrote in 1855 that the field of demography was "the natural and social history of the human species." In this broad sense, demography is closely linked with the disciplines of biology, genetics, psychology, sociology, history, and economics. In a narrower sense Guillard defined demography as "the mathe-

matical study of populations, their general movements, their physical, civil, intellectual and moral condition." Demography as a modern field of study examines the structure of human populations (their distribution by age, sex, marital status, etc.) and their dynamic aspects (births, deaths, migratory movements, etc.). This involves statistical measurement and the use of mathematical methods. Demography is first of all a quantitative science, but it obviously cannot be divorced from a knowledge of the various political, social, and biological influences that affect such statistics.

The construction of mortality tables

Historical development. The beginning of demography as a science may be found in the work of the Englishman John Graunt, who in 1662 published *Natural and Political Observations . . . Made upon the Bills of Mortality*. The "Bills of Mortality" were weekly records of deaths and baptisms going back to the end of the 16th century, which furnished Graunt with his raw material. In search of statistical regularities Graunt made an estimate of the male–female ratios at birth in London and rural communities; he also compared deaths with births, concluding that the former exceeded the latter in London, whereas the opposite situation prevailed in the countryside. But Graunt's most famous contribution was his construction of the first mortality table; by analyzing birth and death rates he was able to estimate the number of men currently of military age, the number of women of childbearing age, the total number of families, and even the population of London. Doubtless the results were imprecise; but the novelty and quality of the arguments and the originality of the enterprise sufficed to establish Graunt as the founder of demographic analysis.

Another precursor of the modern science was the German pastor Johann Süssmilch, who published the first edition of his *Die Göttliche Ordnung* in 1741. Süssmilch's researches, based on information collected in 1,056 parishes in Brandenburg and in various cities and provinces of Prussia, allowed him to describe concretely many of the dynamic aspects of populations; among other things he constructed several mortality tables, including the first one for the whole population of Prussia (1765). Working with extensive statistical series, he was probably the first to perceive the "law of large numbers."

In the 18th century an interest arose in the study of mortality statistics, resulting from the development of life insurance and from the growing attention given to public health; to this period belongs the work of Willem Kersseboom in Holland (1740), Antoine Deparcieux in France (1746), Daniel Bernoulli in Switzerland (1760), and Pehr Wargentin in Sweden (1766), all mathematicians or astronomers. In the 19th century demographic statistics developed rapidly. The establishment of a civil registry of demographic events (births, deaths, marriages) was, in effect, an extension of the old church registries. In the United States, however, the large number of religious groups and the opposition of the civil authorities in certain states greatly retarded the establishment of an adequate system of registration for the whole country; one was set up for births and deaths only in 1933.

Population censuses

Censuses of the population also developed during the 19th century—decennial censuses in the United States began in 1790 and in England in 1801, quinquennial censuses in France in 1801. Demographic research was still concentrated on mortality, notably in the mathematical works of Benjamin Gompertz (1825), William Makeham (1860), Wilhelm Lexis (1875), and Karl Pearson (1897). Interest in fertility developed much later, for it seemed to be a more fundamentally stable phenomenon and one which did not call for collective action. This lack of interest in the study of fertility is remarkable on the part of scientists living in the century of Thomas Malthus, whose thesis that reproduction tended to outrun the means of subsistence was enormously influential. When it became apparent that a considerable decline of fertility had taken place in the industrialized countries during the second half of the 19th century, however, demographers began to make studies of fertility and reproduction.

The phenomenon of differential fertility, with its implications about selection and more particularly about the evolution of intelligence, evoked widespread interest as shown in Charles Darwin's theories and in the works of Francis Galton. Wilhelm Lexis (1875), G.F.R. Böckh (1884), and Robert Kuczynski worked out measurements of population reproduction, while Alfred Lotka, starting in 1907, founded mathematical demography.

As fertility continued to decline, the period between the two world wars saw the development of studies on this subject. At the same time demography took on a broader, interdisciplinary character, as can be seen in the fundamental work of Frank Lorimer and Frederick Osborn, *Dynamics of Population*. After 1920 international demographic conferences multiplied, and in 1928 the International Union for the Scientific Study of Population was created.

In recent decades the field of demography has greatly expanded, reflecting the increased importance of population problems in both the underdeveloped and the developed countries. This has been marked by the creation of numerous specialized research institutes, the publishing of several dozen periodicals devoted exclusively to demographic research, and the great importance assigned to demographic questions by international organizations, particularly the United Nations.

The basis for most demographic research continues to lie in population censuses and the registration of vital statistics. A population census is a gigantic and costly operation, since it involves collecting a considerable amount of information from all the inhabitants of a country in a brief period of time. Various methods are used. One can try to reach the population on a de facto basis (that is, to count all persons where they happen to be at the time of the census), or one can reach the population on a de jure basis, according to their customary place of residence. In order to reach the population de facto, the census must be carried out in a very short period of time; for example, on a fixed day. Increasingly the trend appears to be toward de jure censuses that can be spread out over several weeks. No census is ever completely accurate, since omissions and double counts are inevitable; thus it is estimated that in the 1960 census of the population of the United States there was a 3% omission and a 1.3% double count, giving a net underestimate of 1.7%. Moreover, if the questions are too numerous it is not possible to obtain replies of a high quality except by making exceptional efforts, which add to the expense; to avoid this, some questions are asked of only a fraction of the population.

Birth and death statistics are based on certificates drawn up by local authorities and turned in to the regional or national government. In countries with a long tradition of birth and death registration, such statistics have considerable accuracy. Marriages are often recorded in the same way, but divorces—to the extent that they are determined in law courts—are not.

Demographic studies pursued in depth require extensive data from each individual, some of it of a qualitative nature (opinions, miscellaneous situations, intentions). The inquiry is concentrated on a limited number of persons (a few thousand), chosen in a representative manner by some sampling technique; they may be questioned not only about the situation at a given moment but about their history. Such observations are called *longitudinal*, as contrasted to *transverse* observations made when events occurring over a short period of time are collected from a rather large population. Recent developments in demography have given particular weight to longitudinal observations and the analyses that follow from them.

Longitudinal analyses

Life tables. All demographic research begins with a statistical treatment of quantitative data on the population; this is called demographic analysis. The demographer must calculate the rates of annual events in a population. The commonest are crude rates. For example, the crude death rate is the ratio of deaths in a given year in a population to the average total size of this population

during that year; in 1967 there were 1,851,000 deaths in the United States for an average population (total size of the population in the middle of the year) of 199,114,000, which yields a crude death rate of 9.4 per 1,000. The crude birth rate is defined in the same way: with 3,521,000 births in the United States in 1967, a rate of 17.8 per 1,000 is then obtained. Likewise, the marriage rate is the ratio of marriages during the year to the average population.

But demographic phenomena have variable frequencies depending on the age of the individuals; this leads to the calculation of age-specific rates. If in a given year 2,474 persons die between their 50th and 51st birthdays and the average total number of persons between 50 and 51 years of age is 282,000, the death rate at age 50 is 8.8 per 1,000. Age is not the only temporal characteristic used in analyzing demographic phenomena. Marital fertility rates, for example, become more significant if they are analyzed on the basis of the duration of the marriage.

Demographic rates are usually presented in the form of a table, the best known example of which is the mortality or life table. A mortality table describes the frequency with which deaths occur between successive birthdays in a group of persons born during the same calendar year (a birth cohort). For example, if we take a group of 10,000 women born in France in 1820 we have a mortality table that begins as in Table 1. The probability of death

Table 1: Mortality

age in years (x)	number surviving to age x (l_x)	number dying between the ages x and $x+1$ ($d_{x,\ x+1}$)	mortality rate (per 1,000 women) (q_x)
0	10,000	1,527	152.7
1	8,473	525	62.0
2	7,948	294	37.0
3	7,654	193	25.2
4	7,461		

is defined as the ratio of those dying between the ages x and $x + 1$ to the survivors of age x (or the figures in the third column divided by the figures in the second column). The probability of death is also the probability of the survivors of age x dying before they reach the age $x + 1$. The table can be used to compute life expectancies for populations having the same general characteristics.

A similar table may be constructed for marriage rates. Taking, for example, the French female birth cohort of the year 1900, we get a table that begins as in Table 2.

Table 2: Nuptiality

age in years (x)	number remaining single to age x (c_x)	number marrying between the ages x and $x+1$ ($m_{x,\ x+1}$)	marriage rate (per 1,000 women) (n_x)
15	10,000	57	5.7
16	9,943	176	17.7
17	9,767	396	40.5
18	9,371	722	77.0

The fundamental difference between a marriage table and a mortality table consists in the fact that the second column of the former does not generally end with zero, while that is necessarily the case with the mortality table.

It is possible also to construct a table relating births to the age of a particular cohort of women. For the French female cohort born in 1900, the table begins as in Table 3. The figures in the final column are cumulative, and the total of 2,094 is the sum of all the births (divided by 1,000) for that cohort of women during their reproductive span. The total of female births is known as the gross reproduction rate; when this figure is adjusted downward

Table 3: Fertility

age in years* (x)	number of births between the ages x and $x+1$† ($n_{x,\ x+1}$)	number of births by age x†
15	4	0
16	10	4
17	23	14
18	43	37
19	66	80
48	...	2,094

*Data for ages 20–47 omitted.
†Per 1,000 women.

in accordance with the survival rate of the mothers the result is called the net reproduction rate.

The preceding tables represent the *longitudinal* method of analyzing population characteristics, which is based on the life history of actual cohorts; such data are necessarily more than a century old. The application of current mortality rates to a population results in a current life table. This is compiled by having a fictitious cohort go through all the ages of life, assigning to each age a risk of death, which is the risk observed in the various real cohorts of that age during the current year. This permits an estimate of life expectancy at birth during the current year. Due to declining mortality rates, the life expectation at birth computed in this way always underestimates the average length of life of a new born generation.

Some demographers have contended that the net reproduction rate gives a better indication of the rate of population increase than other measurements, such as crude birth rates and natural increase rates, because it is based on age-specific fertility rates. In applying it to a given population, however, one must make the assumption that the age-specific fertility rates and mortality rates will not change over a period of several decades.

Applications of demographic analysis

Demography is also concerned with describing and interpreting population distributions, the simplest of which are distributions by sex and age. A graphic representation of such a distribution is known as an age pyramid. In an age pyramid, ages are plotted on the vertical axis and the total numbers of males and females in each age category on the two horizontal semi-axes; the progressive narrowing of the pyramid as ages increase is due to the effects of mortality; but the form of the pyramid is also governed by many other factors, including fertility, migration, the impact of wars, etc.

The methods of demographic analysis can be adapted to other fields such as epidemiology, sociology, human genetics, history, etc. Demography itself finds many applications in social and economic fields. Population projections are the basis of forecasts of school population, numbers of pensioners, numbers of households, the structure of the labour market, medical requirements, housing requirements, etc. In making such forecasts it must be remembered that long-range population projections cannot be regarded as predictions; rather, they are attempts to show what the consequences will be if present trends continue. This is the sense, for example, in which we must approach the current demographic projections that indicate that the earth is likely to become overpopulated in the near future.

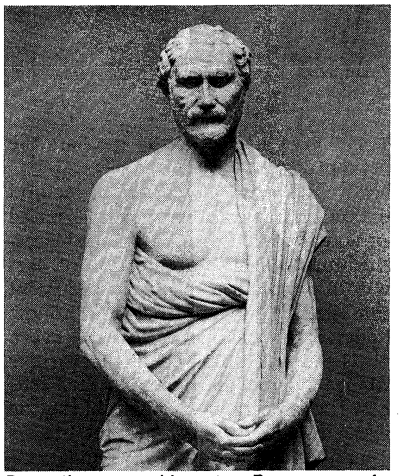
BIBLIOGRAPHY. Classic textbooks on the subject include PETER R. COX, *Demography*, 4th ed. (1970), well documented with regard to the United Kingdom; and MORTIMER SPIEGELMAN, *Introduction to Demography*, rev. ed. (1968), data specifically on the United States. GEORGE W. BARCLAY, *Techniques of Population Analysis* (1958), discusses methods usable in countries with incomplete statistics; ROLAND PRESSAT, *L'Analyse démographique* (1961; expanded Eng. trans., *Demographic Analysis*, 1972), describes techniques with the use of detailed and complete statistics. PHILIP M. HAUSER and OTIS DUDLEY DUNCAN (eds.), *The Study of Population: An Inventory and Appraisal* (1959), is a symposium by a number of authors on the present state of demography.

(R.F.P.)

Demosthenes

A Greek statesman who is often called the greatest orator of all time, Demosthenes, for almost 30 years, rallied the citizens of Athens to oppose the military power of Philip of Macedon and Philip's son Alexander the Great. Demosthenes' speech "On the Crown," a defense of his career delivered in 330 BC, has been termed "the greatest speech of the greatest orator in the world."

By courtesy of the Ny Carlsberg Glyptotek, Copenhagen



Demosthenes, marble statue, Roman copy of a Greek original, which was placed in the Athenian marketplace in 280 BC. In the Ny Carlsberg Glyptotek, Copenhagen.

Demosthenes' life falls into four periods. His first 20 years were spent preparing to regain a lost inheritance, then for another ten he wrote speeches for other speakers. The third period, when he was the virtual leader of the Athenian city-state, lasted for three decades. The last two years of his life found him twice a refugee from Athens, convicted of a grave crime, and finally sentenced to death by the people he had once led.

Heritage and youth. A contemporary of Plato and Aristotle, Demosthenes was born in a township of Athens in 384 BC, the son of a wealthy sword maker. His father died when he was seven, leaving a large inheritance, but the boy's unscrupulous guardians took advantage of their position, and when he came of age Demosthenes received very little of his estate. His strong desire to sue his guardian, Aphobus, in the courts, coupled with a delicate physique that prevented him from receiving the customary Greek gymnastic education, led him to train himself as an orator. He also studied legal rhetoric. In his *Lives of Noble Grecians and Romans* Plutarch, the Greek historian and biographer, relates that Demosthenes built an underground study where he exercised his voice, shaving one half of his head so that he could not go out in public. Plutarch adds that Demosthenes had a speech defect, "an inarticulate and stammering pronunciation" that he overcame by speaking with pebbles in his mouth and by reciting verses when running or out of breath. He also practiced speaking before a large mirror.

Early training and career

Despite this self-improvement program, his first youthful speaking efforts in the public Assembly met with disaster; he was laughed at by his audiences. His lawsuits against Aphobus and two other guardians in 363 were more successful; they produced little money, but he learned much about speaking strategy and methods of argument. Three of his speeches against Aphobus and two against the sculptor Antenor have survived.

Demosthenes as speech writer. At the age of 20, then, the young Demosthenes found himself without his fortune, without a trade or profession, and with seemingly little prospect for success in any field. But his rhetorical skill had been noticed. In 4th-century democratic Athens every citizen who wished to prosecute a lawsuit or to defend himself against accusation had to do the speaking himself. Not every citizen, of course, possessed sufficient skill to write his own speeches—a fact that gave rise to the practice of employing a speech writer (logographer) to prepare a speech for such occasions. Demosthenes' skill in his speeches against Aphobus was recognized by wealthier men in need of a logographer; he soon acquired wealthy and powerful clients willing to pay well for his services. Thus began a lifelong career that he continued even during his most intense involvement in the political struggle against Philip of Macedon, much as a modern lawyer might retain a private practice while engaged in public affairs.

Demosthenes was already 30 when, in 354, he made his first major speech before the Assembly. The speech, "On the Navy Boards," was a marked success. The Assembly or Ecclesia (Ekklēsia), a legislative body composed of all adult male Athenian citizens, had convened to consider a rumoured threat against Athens by the king of Persia. Demosthenes' tightly reasoned oration helped persuade the Athenians to build up their naval strength quietly to show the Persians that, though Athens would not launch an attack, it was ready to fight. He pointed out that, while Athens would have no allies if it attacked first, every other Greek city-state would join Athens if the Persians were the first to attack. Here, for the first time, Demosthenes sounded a theme that was to run through his whole public career—the policy that Athens could best keep its democratic freedom by remaining independent of all other cities while, on the other hand, being ready to make temporary alliances whenever danger threatened. In the same speech, revealing his penchant for careful fiscal planning, he proposed an elaborate revision of the method used to tax the wealthy to raise money for ships.

First Assembly speech

Leader of the democratic faction. From this point on (354), Demosthenes' career is virtually the history of Athenian foreign policy. It was not very long before his oratorical skill made him, in effect, the leader of what today might be called the democratic party. Some interests, especially the wealthy, would have preferred an oligarchy instead of a democracy; many merchants would have preferred peace at almost any price. While they agreed that the Macedonians were barbarians, most Athenian citizens distrusted other Greek city-states, such as Thebes and Sparta. The Athenian Assembly was a loosely organized, often tumultuous body of up to 6,000 male citizens; it was capable of shouting down a speaker it did not like or of routing him with laughter. Any citizen could speak, but the criteria were so high that only the best orators survived for long. In this turbulent arena Demosthenes stood out. Contemporaries refer to him as "a water drinker"; that is, a severe and perhaps forbidding personality. Although name-calling was common practice in the Assembly, Demosthenes' wit was exceptionally caustic; when defending himself in his speech "On the Crown" against the attacks of his lifelong rival, Aeschines, he did not scruple to call him "sly beast," "idle babbler," "court hack," and "polluted." Demosthenes was not merely better at abuse than most; he also realized the advantage of making an audience lose respect for his opponent.

Oratorical style

He was an assiduous student of Greek history, using detailed historical parallels in almost all his public speeches, and reportedly copied out Thucydides' *History of the Peloponnesian War* eight times in order to improve his command of language and to absorb its history. He constantly asked the Athenians to recall their own history, to remember their past belief in democracy, and to remind themselves how much they hated tyrants. His love of democracy gives his speeches a humanistic breadth that makes them interesting even today. Demosthenes was also extremely industrious. Plutarch says that it was his habit to sit down at night and go over the conversations and speeches he had heard during the day, experimenting with various replies or speeches that could have been made. He excelled whenever he could prepare his speeches carefully in advance, but the nature of Athenian political life must often have forced him to reply to an opponent on the spur of the moment. Unfortunately, because all of the surviving speeches are

carefully edited texts, it cannot be established how often Demosthenes spoke extemporaneously.

His famous speech in 354 "On the Navy Boards" was addressed to the threat from the East. Meanwhile, in Macedonia, to the north, the young King Philip, almost the same age as Demosthenes, was gradually annexing Greek cities south of his borders. In 356 Philip had captured an Athenian possession in Thrace, after hoodwinking the Athenians with promises to protect the city, and in 354 he took another Athenian possession. By 353 both Sparta and Arcadia were asking Athens for military assistance against Philip. When he continued to move south, employing bribery and threat as well as military force, the Athenians sent a small force to close off the pass at Thermopylae. Although Philip turned aside to the coast of Thrace, avoiding a direct confrontation with Athens, his intentions were clear. Yet many Athenians continued to believe that Philip's threat was transitory.

The Philippics. Early in 351 Demosthenes delivered a speech against Philip, the so-called "First Philippic," that established him as the leader of the opposition to Macedonian imperial ambitions. For the next 29 years Demosthenes never wavered; as Plutarch says, "The object which he chose for himself in the commonwealth was noble and just, the defense of the Grecians against Philip." In the "First Philippic" he reminded the Athenians that they had once defeated the Spartans, who were as strong as Philip, and sarcastically pointed out that Philip would never have conquered their territories if he had been as timid as the Athenians seemed to be. He concluded by challenging his countrymen to take their affairs in their own hands rather than let Philip win by default.

Defense of the Grecians against Philip

This goading speech nonetheless failed to rouse the Athenians. Philip advanced into Chalcidice, threatening the city of Olynthus, which appealed to Athens. In 349 Demosthenes delivered three stirring speeches (the "Olynthiacs") to elicit aid for Olynthus, but the city fell the following year without significant help from Athens. Finally, Philip and the Athenians agreed in April 346 to the Peace of Philocrates; Demosthenes, partly to gain time to prepare for the long struggle he saw ahead, agreed to the peace and went as one of the ambassadors to negotiate the treaty with Philip. During the negotiations, Philip, recognizing Demosthenes' eloquence as a threat to his plans, ignored him and addressed his fellow ambassador Aeschines instead. The two men returned from the embassy bitter foes, Demosthenes denouncing Aeschines and Aeschines assuring everyone of Philip's good intentions.

In his oration "On the Peace" late in 346 Demosthenes, though condemning the terms of the treaty of Philocrates, argued that it had to be honoured. Meanwhile, Philip continued his tactic of setting the Greek city-states, such as Thebes and Sparta, against each other. Demosthenes was one of several ambassadors sent out on a futile tour of the Peloponnesus to enlist support against Philip. In retaliation Philip protested to Athens about certain statements made by these ambassadors. Demosthenes' "Second Philippic," in 344, retorted that he would never have agreed to the Peace of Philocrates if he had known that Philip would not honour his word; moreover, he asserted, Aeschines and others had lulled the Athenians into a false sense of security. The issue came to a public trial in the autumn of 343, when Demosthenes, in his speech "The False Legation," accused Aeschines of rendering false reports, giving bad counsel, disobeying instructions, and being susceptible to bribery. The court, however, acquitted Aeschines.

Demosthenes' attack on Aeschines

The tangled pattern of threat and counterthreat continued into 341, until an Athenian general incurred Philip's wrath for operating too near one of his towns in the Chersonese. Philip demanded his recall, but Demosthenes replied in a speech, "On the Chersonese," that the motive behind the Macedonian's "scheming and contriving" was to weaken the Athenians' will to oppose Philip's conquests. "Philip is at war with us," he declared, "and has broken the peace." Shortly afterward, Demosthenes delivered his "Third Philippic," perhaps the most successful single speech in his long campaign against Philip. As a result, Demosthenes became controller of the navy and could thus carry out the naval reforms he had proposed in 354. In addition, a grand alliance was formed against Philip, including Byzantium and former enemies of Athens, such as Thebes. Indecisive warfare followed, with Athens strong at sea but Philip nearly irresistible on land. The Macedonian army was well organized under a single brilliant commander who used cavalry in coordination with highly disciplined infantry, while the Greek alliance depended upon what was essentially a group of citizens' militia.

Disaster came in 338, when Philip defeated the allies in a climactic battle at Chaeronea in north central Greece. According to Plutarch, Demosthenes was in the battle but fled after dropping his arms. Whether or not he disgraced himself in this way, it was Demosthenes whom the people chose to deliver the funeral oration over the bodies of those slain in the battle. After the peace concluded by the Athenian orator and diplomat Demades, Philip acted with restraint; and, though the pro-Macedonian faction was naturally greatly strengthened by his victory, he refrained from occupying Athens. Demosthenes came under several forms of subtle legislative attack by Aeschines and others.

Alexander's campaign. In 336 Greece was stunned by the news that Philip had been assassinated. When his son Alexander succeeded him, many Greeks believed that freedom was about to be restored. But within a year Alexander proved that he was an even more implacable foe than his father—for, when the city of Thebes rebelled against him in 335, he destroyed it. A string of victories emboldened Alexander to demand that Athens surrender Demosthenes and seven other orators who had opposed his father and himself; only a special embassy to Alexander succeeded in having that order rescinded. Shortly thereafter, Alexander began his invasion of Asia that took him as far as India and left Athens free of direct military threat from him.

In 330, nevertheless, judging that the pro-Alexandrian faction was still strong in Athens, Aeschines pressed his charges of impropriety against Ctesiphon—first made six years earlier—for proposing that Demosthenes be awarded a gold crown for his services to the state. The real target was, of course, Demosthenes, for Aeschines accused Ctesiphon of making a false statement when he praised the orator's patriotism and public service. The resulting oratorical confrontation between Aeschines and Demosthenes aroused interest throughout Greece, because not only Demosthenes but also Athenian policy of the past 20 years was on trial. A jury of 500 citizens was the minimum required in such cases, but a large crowd of other Athenians and even foreigners flocked to the debate.

Aeschines' attack on Demosthenes

The oration "On the Crown," Demosthenes' reply to Aeschines' charges of vacillating in his policy, accepting bribes, and displaying cowardice in battle, is universally acknowledged as a masterpiece of rhetorical art. It covers the entire two decades of Greek involvement with Philip and Alexander, contrasting Demosthenes' policies in every case with what he terms the treachery of Aeschines as an agent of the Macedonians. As always, his command of historical detail is impressive. Over and over again he asks his audience what needed to be done in a crisis and who did it. Addressing Aeschines directly, he says, "Your policies supported our enemy, mine, our country's." His scathing epithets picture Aeschines as a contemptible turncoat, a hireling of Philip. The jury's verdict was resoundingly clear—Aeschines failed to receive even one-fifth of the votes and was thus obliged to go into exile. Demosthenes and his policies had received a massive vote of popular approval.

Delivery of "On the Crown"

Imprisonment and exile. Six years later, however, he was convicted of a grave crime and forced to flee from prison and himself go into exile. He was accused of taking 20 talents deposited in Athens by Harpalus, a refugee from Alexander. Demosthenes was found guilty, fined 50 talents, and imprisoned. The circumstances of the case are still unclear. Demosthenes may well have intended to

use the money for civic purposes, and it is perhaps significant that the court fined him only two and one-half times the amount involved instead of the ten times usually levied in such cases. His escape from prison made it impossible for him to return to Athens to raise money for the fine. The onetime leader of the Athenians was now a refugee from his own people.

Another dramatic reversal occurred the very next year, however, when Alexander died. The power of the Macedonians seemed finally broken; a new alliance was concluded against them. The Athenians recalled Demosthenes from exile and provided money to pay his fine. But at the approach of Antipater, Alexander's successor, Demosthenes and other orators again fled the city. His former friend Demades then persuaded the Athenians to sentence Demosthenes to death. While fleeing Antipater's soldiers, he killed himself by taking poison at Calauria on October 12, 322. Following his long service to the state, which nonetheless ended in abandonment by the fickle Athenian citizenry, Demosthenes' death can be viewed as a symbol of the decline of Athenian democracy.

Influence and assessment. After his death, according to Plutarch, the Athenians erected a brass statue to his memory. In the following century the scholars at the library in Alexandria carefully edited the manuscripts of his famous speeches. The Roman orator and politician Cicero was so impressed with his speech "On the Crown" that he translated it into Latin and wrote a prologue for it. When Cicero delivered a series of speeches in 44 BC opposing Antony, in circumstances not unlike those in which Demosthenes opposed Philip, Cicero's speeches were called Philippics too. Roman schoolboys studied Demosthenes' speeches as part of their own oratorical training. During the Middle Ages and Renaissance his name was a synonym for eloquence; Queen Elizabeth I was so impressed with Demosthenes that she studied the Greek texts of his speeches with the scholar Roger Ascham. Some 19th-century scholars regarded the Macedonian conquest of the bickering city-states of Greece as a natural political evolution and thus condemned Demosthenes as an opponent of progress. A curious revival of interest in Demosthenes occurred in Europe during World War I: French writers such as the statesman Georges Clemenceau admired him for his resistance to the Macedonian invaders of Greece; some German writers, on the other hand, praised him for his skill in mobilizing political power. Modern scholars such as Werner Jaeger present a more dispassionate view by pointing to the highly complex political issues Demosthenes handled with his oratorical skill. Whatever the interpretations of his personality and work, he has in every age been regarded as one of the world's greatest orator–statesmen.

Influence on Cicero

BIBLIOGRAPHY. "Demosthenes," in *Plutarch, the Lives of the Noble Grecians and Romans*, trans. by JOHN DRYDEN and rev. by ARTHUR HUGH CLOUGH, pp. 1024–1040 (1955); WERNER W. JAEGER, *Demosthenes: The Origin and Growth of His Policy* (1938, reprinted 1963), a balanced view of his life and policies; JAMES J. MURPHY (ed.), *Demosthenes' on the Crown: A Critical Case Study of a Masterpiece of Ancient Oratory* (1967), includes Plutarch's biography, an analysis by GEORGE KENNEDY of Demosthenes' oratorical career, a translation of his speech "On the Crown," and analyses of his use of style, argument, and historical detail.

(J.J.M.)

Denmark

The ancient European kingdom of Denmark comprises the northern part of the low-lying Jutland Peninsula, an archipelago at the entrance to the Baltic Sea, as well as the tiny Faeroe Islands and the immense mass of Greenland, both in the northern Atlantic Ocean. It is the smallest of the three Scandinavian kingdoms, with an area (excluding Greenland) of 16,629 square miles (43,069 square kilometres). Denmark's situation in the Northern Hemisphere, in an area containing more than one-ninth of the world's population and an even greater part of its productive capacity, and especially its location on the North Sea, have brought it the advantage of close proximity and accessibility to the densely populated industrial states of western Europe and have made it a physical, cultural, and commercial bridge between Scandinavia and central Europe. The country was the home of almost 5,000,000 people in the early 1970s.

Division of the country into numerous islands has given it an exceptionally long coastline for its area. The configuration of much of this coast into fjords has, since ancient times, provided natural harbours that encouraged the growth of fishing, shipping, and associated industries. Denmark's economic and political life has long centred around these outlets to the sea. Periods of great political and economic power in the country's history have been based on strong naval power, as in the Viking period, the medieval expansion in the Baltic, the mercantile expansion and colonization of the 16th century, the thriving trade of the 17th century, and the shipping and associated activities of today.

Of Denmark proper the peninsula of Jutland covers 11,449 square miles. To the east, between the Skagerrak and the Baltic, lies the main archipelago of 483 islands, of which 97 are inhabited. Sjælland (Zealand), the largest island in the group, covers 2,708 square miles and is separated from southern Sweden by The Sound (Øresund). Fyn, the second largest island (1,152 square miles), is separated from Sjælland by the Great Belt (Store Baelt) and from Jutland by the Little Belt (Lille Baelt). In the Baltic, farther east of Sjælland, is the island of Bornholm. The Faeroe Islands, with an area of 540 square miles, have a special status within the kingdom as a self-governing region. Greenland (*q.v.*), a Danish colony until integrated into the kingdom in 1953, is the largest island in the world, covering 840,000 square miles, with 132,000 square miles along the coast ice-free. Denmark's 42-mile-long frontier with West Germany extends across the Jutland peninsula from Flensborg Fjord (Flensburger Förde) to a point between the islands of Rømø and Sylt.

The nation whose Viking ancestors terrorized the Atlantic Europe of 1,000 years ago was, in the early 1970s, a small but thriving community enjoying the 10th highest standard of living in the world, yet vulnerably dependent on foreign trade and looking for its future prosperity to the achievement of a state of full European economic cooperation. (For related information, see COPENHAGEN; SCANDINAVIA, HISTORY OF; and GREENLAND.)

THE LAND

The Danish landscape derives its character from the glaciations that covered northern Europe during the Quaternary Period (within the last 2,500,000 years). Erosion and sedimentary deposits in uplands, plains, and valleys are the predominant features.

Denmark proper is a lowland area, on average not more than 100 feet above sea level; its highest point, Yding Skovhøj in central Jutland, reaches only 568 feet.

Division of the country by glacial boundaries

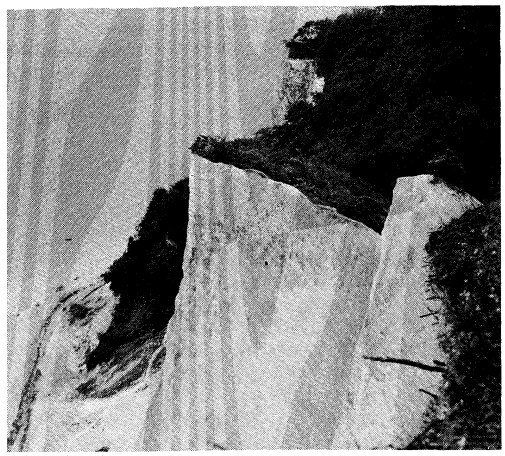
Although the whole of Denmark is relatively low-lying, a remarkable scenic boundary can be traced from Nissum Fjord on the west coast of Jutland eastward toward Viborg, thence swinging sharply south down the spine of the peninsula toward Åbenrå and the German city of Flensburg. This boundary represents the extreme limit reached by the Scandinavian and Baltic ice sheets during the last glaciation, which began about 600,000 years ago.

The ice front is clearly marked in the contrast between the flat west Jutland region, composed of sands and gravels strewn by the meltwaters that poured west from the shrinking ice sheet, and the fertile loam plains and hills of east and north Denmark, which become markedly sandier toward the ice front. In the northwest, around the Limfjorden area, there are numerous landscapes of flat, sand-and-gravel tracts created by marine deposits. Where drainage is difficult, some of these are overlaid by bogs, the peat being exploited and the land then drained for grazing.

In north Jutland and along the coast of south Jutland where there is a perceptible tidal range, salt marshes have been formed by evaporation of what was, in the Late Permian Period (around 225,000,000 years ago), an inland sea. These deposits are commercially exploited.

Senonian chalk, deposited around 100,000,000 years ago, is exposed at the base of the impressive cliffs (Stevns Klint) in southeast Sjælland, at Bulbjerg, and again at Møns Klint. Younger Danian limestone (around 65,000,000 years of age) is extensively quarried in southeast Sjælland.

George Whiteley—Photo Researchers



Møns Klint, chalk cliffs on Møn Island, which rise sharply to a height of over 400 feet (122 metres).

On the island of Bornholm the contrasting solid rock reveals close affinities with that of southern Sweden. Exposure of Precambrian granites (among the oldest on the planet's surface, more than 570,000,000 years old) cover extensive areas of the northern half of the island, which is overlaid to the south by Cambrian sandstone and shales.

About 85 percent of Greenland lies under an ice cap, the thickness of which is as great as 8,900 feet and averages 4,971 feet. The topography of the island is mountainous, consisting largely of Precambrian gneiss and granites. Its highest point, Gunnbjørns Fjæld, reaches 12,140 feet above sea level. The Precambrian bedrock is exposed along the whole west coast and on the east coast from Kap Farvel (Cape Farewell) to Kangerdlugssuaq. The north and east coast to Scoresby Sound consists of very thick continental and marine sediments.

The Faeroe Islands

The high and rugged Faeroe Islands, 17 of which are inhabited, are situated between Scotland and Iceland in the North Atlantic (between 6°15′ and 7°41′ W longitude and 61°26′ and 62°24′ N latitude) and consist of sheets of weather-resistant basaltic lavas from 30 to 100 feet thick (known locally as "hammers"), which originated from volcanic eruptions during the Tertiary Period (from 65,000,000 to 2,500,000 years ago). These hammers are topped with a reddish stone (tuff) about three feet thick. The coasts are indented with deep-cut cliffs and the mountain sides stepped because of the varying resistance of the strata to the ice sheets, which once covered them. Ice action has gouged the valleys into troughshaped hollows.

In Denmark proper the largest stream, the Gudenå, rises in east central Jutland and flows 98 miles to its mouth in the Randers Fjord on the east coast. Small lakes are numerous; the largest is 15.7-square-mile Arresø, on Sjælland. Large lagoons have been formed in places in the west such as Ringkøbing and Nissum fjords.

The substratum on which the Danish soils have developed are chiefly moraines of glacially deposited debris and meltwater sand. By admixture with the subterranean limestones the moraine became calcareous, a factor of vital importance to agriculture. Through centuries of cultivation the Danish soil has been radically improved; the result is an agricultural area comprising 70 percent of the land surface, with maximum potential cropping.

Climate. Being in the Temperate Zone at the meeting point of extremely diverse air masses from the Atlantic, the Arctic, and eastern Europe, Denmark experiences very changeable weather. Although the west coast faces the inhospitable North Sea, the climate—particularly in winter—derives great advantage from the warm North Atlantic Drift (the terminal section of the Gulf Stream). The mean temperature in the coldest month (just under 32° F [−0.1° C] in February) is 12 degrees higher than the average for its latitude. In summer, continental heat may be experienced. The mean temperature in July, the warmest month, is 61° F (16° C). The number of days when frost occurs ranges annually from 70 on the west coast to 120 in the interior.

The warm North Atlantic Drift

The climate is moist; the mean annual precipitation of 25 inches (639 millimetres) ranges from about 32 inches (800 millimetres) in southwest Jutland to about 16 inches on Sprogø in the Great Belt. Rain falls all year-round but is relatively low in winter and spring. The largest monthly volumes of rain fall in August and October.

Climatic variations between north and south on Greenland, situated partly within the Arctic Circle, are considerable. At Ivigtut in the south the mean temperature in July is 50° F (9.9° C) and in February is 18° F (−7.9° C). The corresponding figures for Upernavik in northern Greenland are 41° F (4.9° C) and −9° F (−23° C). Rainfall diminishes from 40 inches per year in the south to eight inches in the north.

The climate of the Faeroes is typically oceanic, with frequent fogs and heavy rain (63 inches annually). The mean temperature in January is 38° F (3.2° C), and in July 51° F (10.8° C). Periods of frost seldom last long and harbours are rarely icebound because of the North Atlantic Drift.

Vegetation and animal life. Denmark's natural vegetation is deciduous forest; but as the country borders the coniferous belt, plantations of spruce and fir thrive. Natural deciduous forests survive in only a few places and consist of oak, elm, beech, and lime. About 10 percent of the land is forest, but almost all of it has been planted and is cultivated. Plants include dune vegetations and heathers.

Tundra vegetation of Greenland

Greenland's vegetation consists chiefly of tundra types —heather, birch, willow, and alder scrub, together with lichens, sedge, and cotton grass. In the south, vegetables such as lettuce and radish are raised, and hay is grown for sheep.

Because of frequent gales, strong west winds, and poor soil, forests do not thrive in the Faeroes and vegetation consists mainly of grass, moss, and mountain-bog flora, affording a basis for large-scale sheep raising.

As Denmark is a densely populated and intensively cultivated country, its original stock of large mammals has been heavily reduced. The largest wild species is the red deer, which inhabits Jutland's forests and plantations. Three hundred and thirty-three bird species have been observed in Denmark, 163 of which breed in the country. Livestock, chiefly cattle and swine, number 12,000,000. Horses number around 40,000 but are in fast decline. The fresh water fauna have been seriously affected by pollution; the least disturbed animal communities are the marine fauna, which form the basis of a large fishing industry.

The rich animal life of Greenland includes reindeer, musk-oxen, polar bears, ermines, snow hares, lemmings, and Arctic foxes, as well as many types of sea birds—eiders, guillemots, auks, wild geese, ducks, and gulls. Land birds include ravens, ptarmigan, falcons, white-tailed eagles, snowy owls, snow buntings, and longspurs. Salmon and trout are found in the rivers; and flounder, capelins, and halibut are important saltwater fish.

On the Faeroes, rats and mice have come with ships, and polar hares have been imported, but there are no indigenous land mammals. Millions of sea birds, including eider ducks, guillemots, and puffins, nest in the high cliffs, and there are abundant trout in the several lakes.

Pattern of land settlement. The original form of habitation in Denmark was the nucleated village with a common-field system. This system was abolished around 1800 and today the landscape is characterized by dispersed farms. West of the ice front is a region of scattered farms, low population density, and rapid population growth; east of the front there is village habitation,

high population density, and a declining rural population. Oats, rye, turnips, and potatoes dominate the fields of west Jutland, while barley, wheat, and sugarbeet are characteristic of east Denmark. As a result of industrialization there has been an increased migration from country to town over the past century. By 1970, 55 percent of the population were living in towns and urbanized districts.

THE PEOPLE

There is little archaeological evidence of human habitation in Denmark earlier than 100,000 years ago. Nomadic hunters were in evidence from about 10,000 BC, living by fishing and catching birds and large animals from the forests. (The skeleton of an auroch with arrowheads embedded in it was found during peat cutting in 1900.) Agriculture was introduced around 3000 BC.

Although the present Danish population is the result of a racial mixture that occurred in the New Stone Age, along with the additional integration of various small groups that have immigrated since, the Nordic characteristics of blond hair and blue eyes do predominate.

The only non-Danish minority in Denmark consists of a German colony in south Jutland, numbering 40,000. Copenhagen (København), a principal transit port of northern Europe, with its suburbs, contained 1,384,000 inhabitants, or just over 25 percent of the total population, in 1971. Other large cities were Århus with 238,000, Odense with 166,000, and Ålborg with 155,000. The Greenland population (47,000 in 1970) is partly Eskimo but consists chiefly of Greenlanders, a Mongoloid–Caucasian intermixture of Eskimos and Danes. Godthåb (population 8,000) is the capital and an important fishing centre. The Faeroese population (39,000 in 1970) is descended from 9th-century Norse settlers. The capital and port is Thorshavn (10,000) on the largest island, Streymoy (144 square miles).

Principal urban places

Language and religion. The Danish language, too, is Nordic, forming a subdivision of the Germanic group. It is closely related to Norwegian and Swedish, though less melodic; it was first distinguishable from Swedish only around 1100 AD. Outside Denmark, Danish is spoken in southern Schleswig-Holstein in northern West Germany.

The Faeroese language resembles modern Norwegian and Icelandic; and, of course, there is a vast difference between Danish and the Eskimo languages spoken in Greenland. Both Greenlandish and Faeroese are recognized as the principal languages of their respective regions, but a thorough grounding in Danish is still compulsory in the islands' schools.

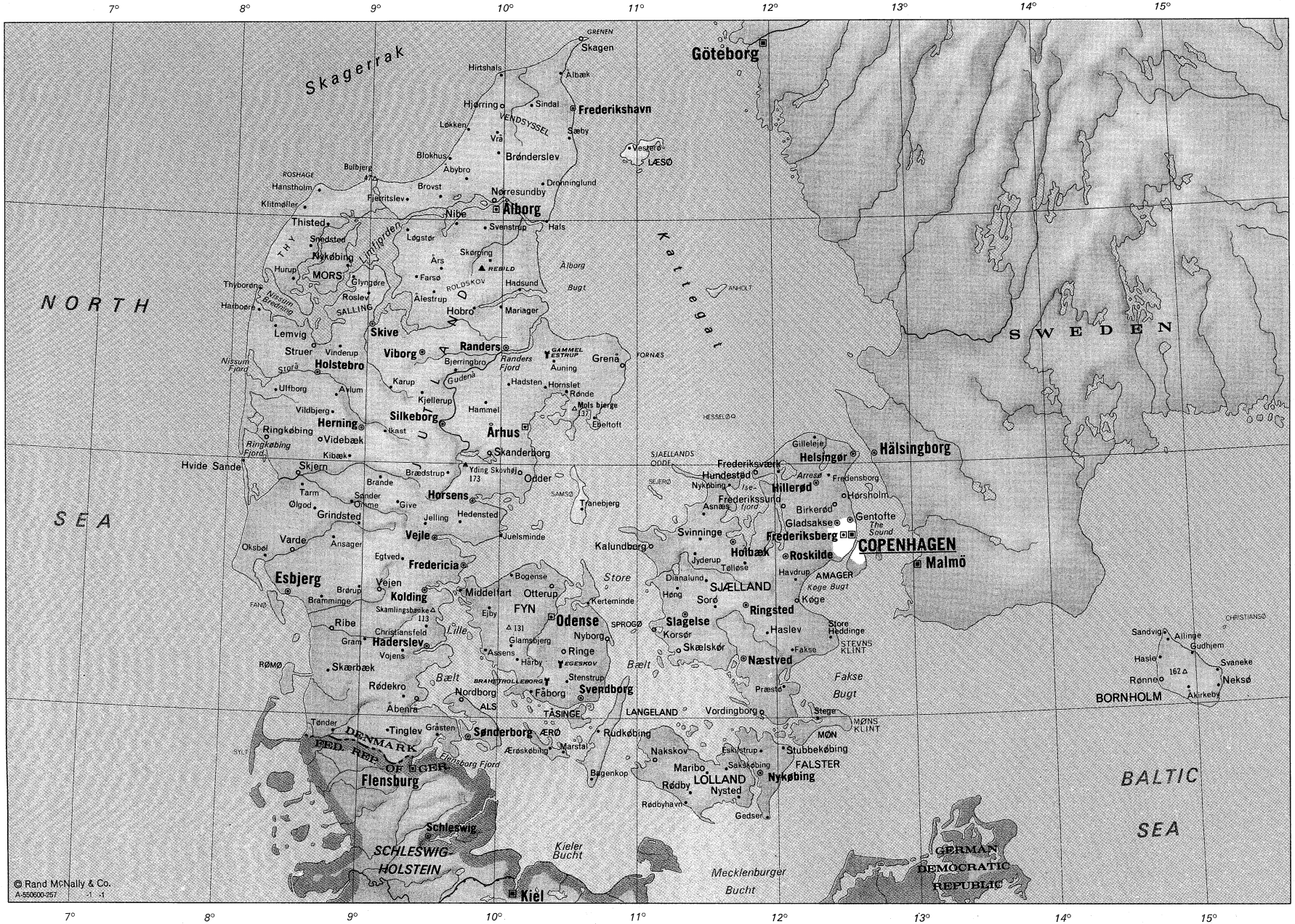
MAP INDEX

Cities and towns

Åbenrå 55·02n 9·26e
Åbybro 57·09n 9·45e
Ærøskøbing 54·53n 10·25e
Åkirkeby 55·04n 14·56e
Ålbæk 57·36n 10·25e
Ålborg 57·03n 9·56e
Ålestrup 56·42n 9·30e
Allinge 55·16n 14·49e
Ansager 55·42n 8·45e
Århus 56·09n 10·13e
Års 56·48n 9·32e
Asnæs 55·49n 11·31e
Assens 55·16n 9·55e
Auning 56·26n 10·23e
Avlum 56·16n 8·48e
Bagenkop 54·45n 10·41e
Birkerød 55·50n 12·26e
Bjerringbro 56·23n 9·40e
Blokhus 57·15n 9·35e
Bogense 55·34n 10·06e
Brædstrup 55·58n 9·37e
Bramminge 55·28n 8·42e
Brande 55·57n 9·07e
Brønderslev 57·16n 9·58e
Brørup 55·29n 9·01e
Brovst 57·06n 9·32e
Christiansfeld 55·21n 9·29e
Copenhagen (København) 55·40n 12·35e
Dianalund 55·32n 11·30e
Dronninglund 57·09n 10·18e
Ebeltoft 56·12n 10·41e
Egtved 55·37n 9·18e
Ejby 55·26n 9·57e
Esbjerg 55·28n 8·27e
Eskilstrup 54·51n 11·54e
Fåborg 55·06n 10·15e
Fakse 55·15n 12·08e
Farsø 56·47n 9·21e
Fjerritslev 57·05n 9·16e
Fredensborg 55·58n 12·24e
Fredericia 55·35n 9·46e
Frederiksberg 55·25n 11·34e
Frederikshavn 57·26n 10·32e
Frederikssund 55·50n 12·04e
Frederiksværk 55·58n 12·02e
Gedser 54·35n 11·57e
Gentofte 55·45n 12·33e
Gilleleje 56·07n 12·19e
Give 55·51n 9·15e
Gladsakse 55·44n 12·29e
Glamsbjerg 55·16n 10·07e
Glyngøre 56·46n 8·52e
Gram 55·17n 9·04e
Gråsten 54·55n 9·36e
Grenå 56·25n 10·53e
Grindsted 55·45n 8·56e
Gudhjem 55·13n 14·59e
Haderslev 55·15n 9·30e
Hadsten 56·20n 10·03e
Hadsund 56·43n 10·07e
Hals 57·00n 10·19e
Hammel 56·15n 9·52e
Hanstholm 57·07n 8·38e
Harboøre 56·37n 8·12e
Hårby 55·13n 10·07e
Hasle 55·11n 14·43e
Haslev 55·20n 11·58e
Havdrup 55·32n 12·08e
Hedensted 55·46n 9·42e
Helsingør 56·02n 12·37e
Herning 56·08n 8·59e
Hillerød 55·56n 12·19e
Hirtshals 57·35n 9·58e
Hjørring 57·28n 9·59e
Hobro 56·38n 9·48e
Holbæk 55·43n 11·43e
Holstebro 56·21n 8·38e
Høng 55·31n 11·18e
Hornslet 56·19n 10·20e
Horsens 55·52n 9·52e
Hørsholm 55·53n 12·30e
Hundested 55·58n 11·52e
Hurup 56·45n 8·25e
Hvide Sande 55·59n 8·08e
Ikast 56·08n 9·10e
Jelling 55·45n 9·26e
Juelsminde 55·43n 10·01e
Jyderup 55·40n 11·26e
Kalundborg 55·41n 11·06e
Karup 56·18n 9·10e
Kerteminde 55·27n 10·40e
Kibæk 56·02n 8·51e
Kjellerup 56·17n 9·26e
Klitmøller 57·02n 8·31e
København, see Copenhagen
Køge 55·27n 12·11e
Kolding 55·31n 9·29e
Korsør 55·20n 11·09e
Lemvig 56·32n 8·18e
Løgstør 56·58n 9·15e
Løkken 57·22n 9·43e
Mariager 56·39n 10·00e
Maribo 54·46n 11·31e
Marstal 54·51n 10·31e
Middelfart 55·30n 9·45e
Næstved 55·14n 11·46e
Nakskov 54·50n 11·09e
Neksø 55·04n 15·09e
Nibe 56·59n 9·38e
Nordborg 55·03n 9·45e
Nørresundby 57·04n 9·55e
Nyborg 55·19n 10·48e
Nykøbing 54·46n 11·53e
Nykøbing 55·55n 11·41e
Nykøbing 56·48n 8·52e
Nysted 54·40n 11·45e
Odder 55·58n 10·10e
Odense 55·24n 10·23e
Oksbøl 55·38n 8·17e
Ølgod 55·49n 8·37e
Otterup 55·31n 10·24e
Præstø 55·07n 12·03e
Randers 56·28n 10·03e
Ribe 55·21n 8·46e
Ringe 55·14n 10·29e
Ringkøbing 56·05n 8·15e
Ringsted 55·27n 11·49e
Rødby 54·42n 11·24e
Rødbyhavn 54·39n 11·21e
Rødekro 55·04n 9·21e
Rønde 56·18n 10·29e
Rønne 55·06n 14·42e
Roskilde 55·39n 12·05e
Roslev 56·42n 8·59e
Rudkøbing 54·56n 10·43e
Sæby 57·20n 10·32e
Sakskøbing 54·48n 11·39e
Sandvig 55·17n 14·47e
Silkeborg 56·10n 9·34e
Sindal 57·28n 10·13e
Skælskør 55·15n 11·19e
Skærbæk 55·09n 8·46e
Skagen 57·44n 10·36e
Skanderborg 56·02n 9·56e
Skive 56·34n 9·02e
Skjern 55·57n 8·30e
Skørping 56·50n 9·53e
Slagelse 55·24n 11·22e
Snedsted 56·54n 8·32e
Sønderborg 54·55n 9·47e
Sønder Omme 55·50n 8·54e
Sorø 55·26n 11·34e
Stege 54·59n 12·18e
Stenstrup 55·07n 10·31e
Store Heddinge 55·19n 12·25e
Struer 56·29n 8·37e
Stubrekøbing 54·53n 12·03e
Svaneke 55·08n 15·09e
Svendborg 55·03n 10·37e
Svenstrup 56·57n 9·52e
Svinninge 55·43n 11·28e
Tarm 55·55n 8·32e
Thisted 56·57n 8·42e
Thyborøn 56·42n 8·13e
Tinglev 54·56n 9·15e
Tølløse 55·37n 11·45e
Tønder 54·56n 8·54e
Tranebjerg 55·50n 10·36e
Ulfborg 56·16n 8·20e
Varde 55·38n 8·29e
Vejen 55·29n 9·09e
Vejle 55·42n 9·32e
Vesterø 57·18n 10·56e
Viborg 56·26n 9·24e
Videbæk 56·05n 8·38e
Vildbjerg 56·12n 8·46e
Vinderup 56·29n 8·47e
Vojens 55·15n 9·19e
Vordingborg 55·01n 11·55e
Vrå 57·21n 9·57e

Physical features and points of interest

Ærø, *island* 54·53n 10·20e
Ålborg Bugt, *bay* 56·45n 10·30e
Als, *island* 54·59n 9·55e
Amager, *island* 55·37n 12·37e
Anholt, *island* 56·42n 11·34e
Arresø, *lake* 55·58n 12·08e
Baltic Sea 54·40n 15·00e
Bornholm, *island* 55·10n 15·00e
Brahetrolleborg, *castle* 55·09n 10·22e
Bulbjerg, *hill* 57·09n 9·02e
Christiansø, *island* 55·19n 15·12e
Egeskov, *castle* 55·10n 10·30e
Fakse Bugt, *bay* 55·10n 12·15e
Falster, *island* 54·48n 11·58e
Fanø, *island* 55·25n 8·25e
Flensborg Fjord (Flensburger Förde), *bay* 54·59n 9·45e
Fornæs, *cape* 56·27n 10·58e
Fyn (Fünen), *island* 55·20n 10·30e
Gammel Estrup, *castle* 56·26n 10·21e
Great Belt, see Store Bælt
Grenen, *spit* 57·44n 10·40e
Gudenå, *river* 56·29n 10·13e
Hesselø, *island* 56·12n 11·43e
Isefjord, *bay* 55·52n 11·49e
Jutland (Jylland), *peninsula* 56·00n 9·15e
Kattegat, *strait* 57·00n 11·00e
Køge Bugt, *bay* 55·30n 12·20e
Læsø, *island* 57·16n 11·01e
Langeland, *island* 55·00n 10·50e
Lille Bælt (Little Belt), *strait* 55·20n 9·45e
Limfjorden, *channel* 56·55n 9·10e
Little Belt, see Little Bælt
Lolland, *island* 54·46n 11·30e
Mols Bjerge, *hill* 56·13n 10·32e
Møn, *island* 55·00n 12·20e
Møns Klint, *cliff* 54·58n 12·33e
Mors, *island* 56·50n 8·45e
Nissum Bredning, *bay* 56·38n 8·22e
Nissum Fjord 56·21n 8·14e
North Sea 56·00n 7·00e
Øresund, see Sound, The
Randers Fjord, *bay* 56·36n 10·20e
Rebild, *museum* 56·50n 9·51e
Ringkøbing Fjord, *bay* 56·00n 8·15e
Roldskov, *forest* 56·48n 9·50e
Rømø, *island* 55·08n 8·31e
Roshage, *point* 57·07n 8·38e
Salling, *physical region* 56·40n 9·00e
Samsø, *island* 55·52n 10·37e
Sejerø, *island* 55·53n 11·09e
Sjælland (Zealand), *island* 55·30n 11·45e
Sjællands Odde, *point* 55·58n 11·22e
Skagerrak, *strait* 57·30n 9·00e
Skamlingsbanke, *hill* 55·25n 9·34e
Sound, The (Øresund) 55·50n 12·40e
Sprogø, *island* 55·20n 10·58e
Stevns Klint, *cliff* 55·18n 12·27e
Storå, *river* 56·19n 8·19e
Store Bælt (Great Belt), *strait* 55·30n 11·00e
Tåsinge, *island* 55·00n 10·36e
Thy, *physical region* 57·00n 8·30e
Vendsyssel, *physical region* 57·20n 10·00e
Yding Skovhøj, *hill* 56·00n 9·48e
Zealand, see Sjælland



DENMARK

Size of symbol indicates relative size of town • ○ ◉ ⊡ ▣

Elevations in metres

0 25 50 100 150 km

0 25 50 100 mi

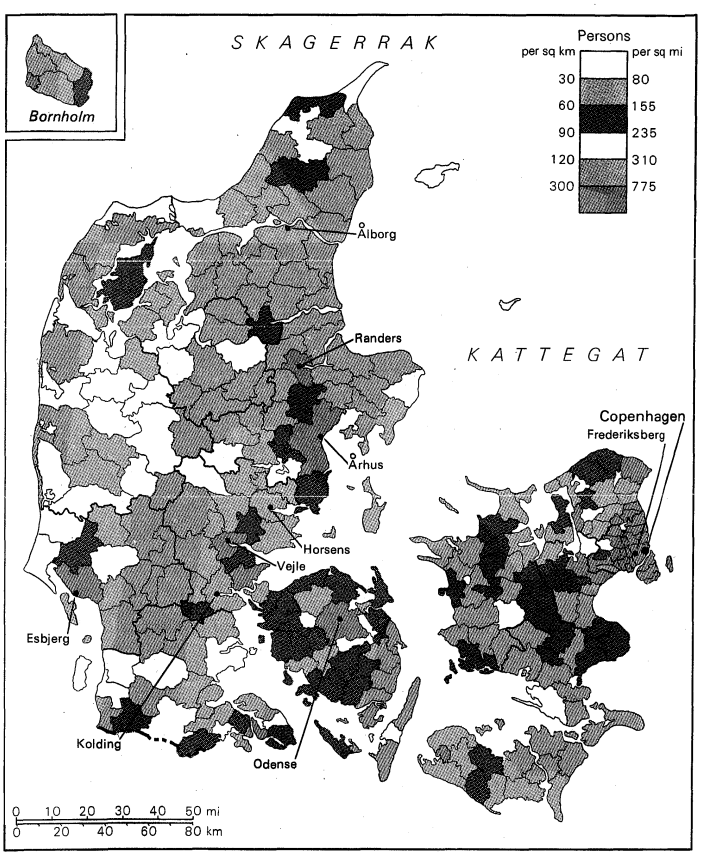
Denmark, Area and Population

	area		population	
	sq mi	sq km	1965 census	1971 estimate
Amtskommuner				
Århus	1,764	4,570	498,000	534,000
Bornholms	227	588	49,000	47,000
Frederiksborg	520	1,346	211,000	261,000
Fyns	1,346	3,486	425,000	434,000
København	201	520	558,000	617,000
Nordjyllands	2,383	6,171	451,000	457,000
Ribe	1,210	3,135	191,000	198,000
Ringkøbing	1,872	4,849	231,000	242,000
Roskilde	344	890	119,000	154,000
Sønderjyllands	1,517	3,929	231,000	239,000
Storstrøms	1,311	3,396	252,000	253,000
Vejle	1,155	2,991	294,000	307,000
Vestsjællands	1,152	2,983	248,000	259,000
Viborg	1,591	4,120	220,000	221,000
Kommuner				
Frederiksberg	3	9	111,000	102,000
Copenhagen (København)	33	85	678,000	626,000
Total Denmark	16,629	43,069*	4,768,000*	4,951,000

*Figures do not add to total given because of rounding.
Source: Official government figures.

The religion to which the bulk of the population belongs is Evangelical Lutheran.

Demography. Population growth in Denmark is determined primarily by the annual increase in births over deaths rather than by the effect of net migration, which is normally insignificant in comparison. In the mid-1960s, immigration and emigration figures cancelled each other out, with an approximate movement of 30,000 in either direction. By the early 1970s, however, immigration of short-time workers from countries of southern Europe was increasing. In 1968 the overall increase in population amounted to 27,000 annually, but the introduction of the contraceptive pill resulted in a rapid fall in the birth rate, even though the average woman giving birth was younger. The death rate among infants in Denmark over the past half century was cut from more than 10 percent to less than 2 percent by 1971. Average life expectancy, put at 50 years at the beginning of the century, by 1971 was just over 70 years for men and 75 years for women. The most frequent causes of death in Denmark were heart diseases (34 percent) and cancer (23 percent). About 5.25 percent of deaths were caused by accidents. The suicide rate in Denmark (2 percent) was high: it was 24 per 100,000 for men and 12 per 100,000 for women.



Population density of Denmark.

Changing pattern of age distribution

Until World War I the distribution of the population in terms of age groups (children, young people, adults, and old people) was very stable. Since then, considerable changes have taken place, partly because of the declining birthrate. The youngest age groups were comparatively smaller and the oldest comparatively larger. In 1969, 12 percent of the population was under seven; 11 percent was aged between seven and 13; 9 percent between 14 and 19; 59 percent between 20 and 66; and 10 percent was 67 and over. (Estimates based on present rates of increase give a total population of 6,000,000 by 2000 AD, slightly more than 800,000 of whom will be older than 67.)

In Greenland the birth rate was two and one-half times as high as in Denmark proper. A halving of the death rate during the 1950s, primarily from a marked decrease in fatalities from tuberculosis, led to a 1971 deathrate of 0.7 percent, considerably lower than in the rest of Denmark (0.98 percent in 1970). Children under 14 in Greenland made up more than 40 percent of the population, however, as compared with 20 percent for the rest of the country; and the age group above 60 years constituted only 4 percent of Greenland's population, compared with 14 percent in the rest of the country. Life expectancy in Greenland was 57 years for males and 63 for females—13 and 12 years lower, respectively, than in the rest of Denmark. The difference was partly accounted for by an infant mortality rate that, despite a marked reduction, was still two or three times as high as in the rest of the country. The most serious diseases reported in Greenland were gonorrhea, influenza, and quinsy; and the most frequent causes of death were accidents, particularly drowning.

There was a distinct tendency among the Danes to marry younger. Over the past century the average marrying age for those marrying for the first time dropped by 7 years: for men, from 32 to 25 years, and for women, from 29 to 22 years. Just over 25 percent of marriages contracted annually were solemnized by a secular authority, not in a religious ceremony.

Between 1901 and 1905 there were, on average, 500 divorces annually; thus, 4 percent of all marriages were dissolved. In 1969 there were 9,000 divorces; thus, almost 25 percent of marriages were dissolved that year, by death or divorce.

THE NATIONAL ECONOMY

Standard of living

Resources and production. Despite rather poor natural resources, the Danish population enjoys a standard of living equalled by few other countries in the world. The 1970 gross national product (GNP) represented the 10th highest average income per capita in the world. In proportion to its size Denmark has an extremely large volume of foreign trade (exporting 30 percent of GNP). Its chief trading partners are Great Britain, the Federal Republic of Germany, Sweden, and the United States.

Natural deposits of importance to industry as fuels or raw materials are found to a very limited extent in Denmark. There is a particular scarcity of metals, coal, and oil. Oil and gas prospecting is indeed going on in the Danish sector of the North Sea, with good expectations of early exploitation, especially of gas resources. Nevertheless, next to its labour force, the soil is still Denmark's most important raw material, for, as has been mentioned, about 70 percent of the nation's land is devoted to agriculture, and only a small percentage of this is permanently in grass. The soil is intensively exploited and extensively fertilized. More than half of the cultivated land is devoted to cereals, barley accounting for about 70 percent of the total grain harvest. Dairy farming is important, with annual production of milk around

5,000,000 tons. About 12,000,000 pigs are killed each year. Chickens numbered about 18,000,000 in 1970, producing 90,000 tons of eggs.

Farms are generally small or medium sized, with holdings of between 25 and 150 acres shouldering the main burden of production. There is an unusually large number of freehold farms—98 percent against 2 percent rented. Cooperatives help to make small farm operations profitable, and the dominance of animal husbandry also lends itself to individual enterprise. There were 140,000 agricultural properties by 1970, although the number was decreasing by around 7,000 annually.

Of the total value of Danish animal and vegetable output, domestic animal products account for 90 percent. Total agricultural production is sufficient to meet the requirements of more than 16,000,000 people and is steadily increasing despite the fall in agricultural manpower.

The fishing industry employs approximately 13,000 fishermen in over 8,000 vessels with a total tonnage of just over 100,000 gross registered tons. Herring, cod, and plaice (or flatfish) account for about half the total catch, which is made up of about 50 species in all; other important species are salmon, mackerel, eels, and deepwater shrimp.

Denmark's small extractive industry relies on the supply of granite (for roads and house building) and kaolin (for ceramic and paper manufacture) found on Bornholm. At brickworks all over the country boulder clays are molded and baked to make bricks and tiles; moler (marine diatomaceous earth) is used in insulating materials for the building industry and white chalk in the important cement industry.

In Greenland, marble is mined at Marmorilik and coal on Disko Island. Cryolite mining at Ivigtut ceased in 1962 although stockpiles were expected to last for 20 to 30 years. There is intensive exploration for minerals such as uranium and thorium. Fairly well developed coal deposits are mined on Sudhuroy in the Faeroe Islands.

Organization, trade, and finance. Imports of raw materials and fuel used to be paid for largely by exports of agricultural products, supplemented by income from Danish shipping and the tourist trade. But since the early 1960s the overseas trade pattern has changed, and industrial products were accounting for 66 percent of total exports by 1970, against only 40 percent in 1959. Agriculture's share fell in this period from over 50 percent to around 30 percent. The largest employers were the iron and metal industries (28 percent of all workers); followed by the food industry (18 percent); the paper and graphic industries and transport (both 9 percent); and then footwear and clothing, wood and furniture, and extraction industries. By mutual agreement between workers' and employers' organizations, wages are automatically adjusted to the cost-of-living index every six months and wage agreements negotiated biennially.

Power plants are owned by cooperatives or municipal authorities and mainly burn imported oil or coal. Some power is imported from Swedish hydroelectric plants.

The National Bank of Denmark is the only bank of issue and has a special status as a self-governing institution under government supervision. Profits revert to the state treasury. There are three large private banks and about 110 smaller ones. Besides these commercial banks, there are approximately 370 nonprofit-making savings banks.

Trading partners

Denmark's trade is almost entirely within Europe. Its trade within the European Free Trade Association, of which it became a member in 1960, increased over the 1960s, although trade with the European Economic Community (EEC) was also considerable. Before EEC entry, EFTA countries accounted for 50 percent of Denmark's exports and EEC countries for 23 percent of exports.

Great Britain is the chief buyer of Danish agricultural produce, with West Germany second. Those two countries jointly absorb almost 60 percent of all agricultural exports. Sales to West Germany, however, dropped at a considerable rate as a result of the EEC's self-supplying policy. Sweden is Denmark's largest market for manufactured goods and its second largest market overall (after Great Britain).

Denmark's is a mixed welfare-state economy, with private sector expenditure accounting for 55 percent of net national income. The very high share of public expenditure in the net national income is the result of a considerable expansion in higher education, national defense, social services, and, through subsidization, agriculture. Neither state nor local government authorities own capital nor have they any significant commercial or industrial income. The greater part of public income is thus derived from taxation: on real estate, on personal income and capital, and through customs and excise duties.

Taxes on real estate amounted to about 5 percent of total receipts in 1971, by far the greatest part being received by local authorities. Personal income tax—levied by both state and municipal authorities—represents 45 percent of total tax receipts. Local taxes are always proportional to income, and there is a state surtax on incomes above a certain level. Income tax is deducted on a PAYE (pay-as-you-earn) system. Danish company tax is relatively low.

PAYE and VAT tax systems

The chief of the indirect taxes, which go to the state, is the VAT (value-added-tax) of 15 percent. In addition to these and to customs duties, there are high indirect taxes on selected goods such as wines and spirits, tobaccos, petroleum, and motor vehicles.

Both employers and employees are well organized in Denmark. Workers are normally organized according to their particular skills, no matter what industry they happen to be in. The dominant national employees' organization is the National Trades Union Centre (Landsorganisationen i Danmark, De Samvirkende Fagforbund). The principal employers' organization is the Danish Employers' Confederation (Dansk Arbejdsgiverforening).

The change that Denmark underwent in the 1960s from an agricultural to an industrial country was financed by foreign borrowings and resulted in increasing inflation. That inflation and the accompanying deficit in the balance of payments, which had grown from near zero in 1960 to U.S. $1,550,000,000 by 1970, were the chief economic problems facing contemporary Danish governments. Their solution was seen, in part, in participation in the expanded market that EEC entry (effective January 1, 1973) would bring.

Transportation. Communications by land, sea, and air are conditioned principally by comparatively high population density, a small and low-lying geographical area, the considerable amount of overseas trade engaged in, and the fact that Denmark is one of the most dissected landmasses in Europe.

Highways total 5,400 miles; of these 1,400 miles are arterial roads. Secondary roads maintained by local district authorities total 29,000 miles. Road networks in cities and towns account for a further 4,000 miles. The number of motor vehicles in Denmark has risen rapidly in recent years. In December 1968 there were 1,140,000 automobiles, of which 870,000 were private cars, just over 65,000 motorcycles and 140,000 certified tractors. (The number of private cars exceeded 1,000,000 by 1970.) While the number of motorcycles was declining, that of cars had virtually doubled since 1966. By 1971 there was one private car to every six persons. Bicycles were estimated to number 2,000,000.

Throughout the country in 1969 there were 1,100 bus and coach routes, the majority operated by the Danish State Railways. The routes totalled about 19,000 miles.

Denmark's comparatively large railway network, built mainly during the last half of the 19th century, totals 2,199 route miles; two-thirds of that trackage is state-operated, and the rest is run by companies the shares of which are normally owned by local government authorities and the state.

Ferry services and bridges

An important supplement to the railways and a characteristic feature of the Danish communications system are the ferries and many bridges. The principal ferry services are the Great Belt Ferry between Sjælland and Fyn and the ferries linking Denmark and Sweden. Two bridges connect Fyn and Jutland. Denmark's longest bridge, the Storstrøm (two miles), connects Sjælland and Falster.

Denmark's considerable overseas trade and many good harbours provide favourable conditions for shipping. In 1970 the Danish merchant fleet comprised 1,433 ships of 100 gross registered tons or more, making a total tonnage of 3,220,000 gross registered tons; 85 of them were tankers with a combined weight of 1,200,000 tons.

Copenhagen, with one of the busiest airports in Europe (at Kastrup), serves inland routes to Jutland and Bornholm and is a centre of international traffic. The Scandinavian Airlines System (SAS), a joint Danish-Norwegian-Swedish enterprise, serves European and intercontinental lines.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Constitution and politics. Constitutional monarchy was introduced in Denmark on June 5, 1849. Legislative authority, based on the constitution of June 5, 1953, rests jointly with the crown and with parliament, which consists of one chamber, the Folketing, with 179 members (including two from the Faeroe Islands and two from Greenland) elected for four years. Acts passed by the Folketing must be signed by the king and countersigned by at least one minister before becoming law. Furthermore, the monarch, as chief executive, is responsible for carrying legislation into effect, seeing that laws are observed and violations prosecuted. The king exercises his executive powers through the 19-minister State Council, headed by the prime minister. No minister can remain in office if the Folketing passes a vote of no confidence in him.

Apart from the Faeroes and Greenland, where special administrative arrangements are in force, the country is divided into 14 counties and 277 municipalities. All persons over 20 are entitled to vote in both the Folketing and local council elections, and the ballot is voluntary and secret. All voters (except those found guilty of a criminal offense) are also eligible to stand for election. In an attempt to achieve the equitable representation for all prevailing political outlooks, as guaranteed in the constitution, elections in individual constituencies are combined with election by proportional representation.

At the 1971 general election the Social Democratic Party, mainly representing the workers, won 70 seats; the Conservative People's Party, representing the more prosperous urban population, gained 31 seats; the Liberal Party, representing the farmers, held 30 seats; the Radical Liberal Party, which obtained most of its support from intellectuals, the urban middle class, and independent smallholders, gained 27 seats; and the small Socialist People's Party took 17 seats.

Role of the Ombudsman

The amendment to the Constitution passed in 1953 included the introduction of the Ombudsman (1954), whose office was, on behalf of the Folketing, to superintend the administrative services of the government and, to a certain extent, those of the municipalities. Every citizen could thus lodge a complaint in the event of arbitrary treatment from ministers, civil servants, or others employed by the public administration. Once a case has been investigated, the Ombudsman may express an opinion and, if necessary, bring the case before the courts. If a complaint concerns a minister or former minister the Ombudsman may report it to the Folketing.

Justice and police. The administration of justice is exercised by more than 100 lower courts, two high courts and a Supreme Court with its seat in Copenhagen. The independent judiciary is appointed by the king.

The police service is organized on a national basis under the control of the chief of state police; there are 72 police districts, each headed by a chief constable.

Defense. Danish national defense policy is based on membership in the North Atlantic Treaty Organization (NATO). It is the constitutional duty of all men capable of bearing arms to contribute to national defense. The normal period of compulsory national service is 12 months, but the armed forces also enlist many volunteers. Conscientious objectors to military service may serve in the Civil Defense Force, in special camps in the state forests, or in various positions in social work. Service in the Home Guard and Civil Defense Corps is voluntary.

There is a United States military base at Thule in Greenland, but no nuclear weapons are admitted on Danish soil or over Greenland's air space.

Social welfare. *Education.* Education in Denmark is compulsory for children between seven and 16 years of age. A growing number continue education after primary school, either in a grammar school or in various recently established nonacademic institutes. The majority of schools are publicly run and free. Private schools receive large grants from the state, as do youth schools, which offer vocational and other supplementary training. Denmark's most significant contribution to adult education is the folk high school, in which no entrance conditions are imposed; great freedom exists within the folk high school curriculum, and no examinations are held on conclusion of courses.

Further education in technical schools and universities

Further education is given at commercial and technical colleges, professional institutions, and at the universities of Copenhagen (founded in 1479), Århus (1928), and Odense (1964). A fourth university was scheduled to open in Roskilde in 1972. All universities and higher seats of learning receive their incomes from the state.

Social insurance. According to the Danish constitution, "any person unable to support himself or his dependents shall, where no other person is responsible for his or their maintenance, be entitled to receive public assistance."

A national old-age pension scheme is available for all persons over 67, with basic amounts being paid, irrespective of the recipient's financial position. The scheme is administered by the municipalities and financed largely by the state through taxation. In addition, a supplementary pension is financed by contributions from wage earners themselves and employers (although employers' contributions to social-insurance schemes in general are substantially lower than in some other European countries). There is, moreover, a disability pension, and a pension for widows over 55 years of age is also in effect. Membership in the public health insurance program is compulsory, and all persons are entitled to free hospital treatment and home nursing. The cost of medicine is largely subsidized.

Unemployment insurance is voluntary but widespread: worker and employer pay the insurance but get state repayments if unemployment rises or becomes protracted. Industrial injury insurance is paid by all employers to cover employees against the consequences of accidents; and the Rehabilitation Act provides for the physical, mental, and occupational training of vocationally handicapped persons. Children's allowances are paid until the age of 16, irrespective of a family's total income. All children are medically examined, free of charge, at regular intervals throughout infancy and schooling.

Housing. At the 1965 census of population and housing there were 1,600,000 dwellings (5,500,000 rooms) in Denmark. Of these, 12 percent were in agricultural properties, 35 percent in other single-family houses, nine percent in two-family houses, and 44 percent in blocks of flats. The average size of dwellings was probably one of the highest in Europe. Under a 1966 act tenants receive rent rebates from both state and municipal authorities according to a graded scale dependent on rent level and family income. Annual housing production amounted to 50,000 dwelling units, or more than 10 dwellings per 1,000 inhabitants. In money terms, housing construction represented just over 4 percent of total national production.

Health and leisure. With free medicine available to all and a high standard of living (per capita consumption at three times the level of the turn of the century), health conditions are good. An increasing proportion of the Danes' income is spent on the enjoyment of leisure hours. The Danes are keen sportsmen. Football (soccer) and gymnastics are the most popular sports. There are an estimated 200,000 or more football players and 120,000 amateur gymnasts in the country. Other popular sports include sailing, rowing, skating, tennis, badminton, and swimming. All major provincial towns have their own sports hall, stadium, or ice rink.

Popularity of sports

Denmark is among the richest nations in the world, and, although there are great differences in living conditions between Denmark proper and Greenland and the Faeroes, the emphasis on high public-sector expenditure at the expense of private consumption has tended to narrow the income range and prohibit extremes of wealth or poverty.

CULTURAL LIFE

Just as Denmark is economically dependent on its more powerful neighbours, so too its culture reveals their influence. English, French, and German thought has often provided inspiration. German in particular has been a frequent channel, though not always a source, of ideas. It has been only in those periods when German intellectual life was stagnant, as, for example, under the Nazis, that cultural influences from Denmark's southern neighbour have dwindled.

The arts. *Literature.* There is an original Danish contribution to literature, however, and it dates back almost 1,500 years. Although the original texts produced during the great tribal migrations of the 5th century were lost, their contents are preserved in the Icelandic sagas, dating from around 1200, and in the great history *Gesta Danorum* by the Danish monk Saxo Grammaticus (died 1204). The richest source of medieval Danish literature is the folk ballad, its lyrical qualities stemming from the German influence; in the Reformation a homiletic literature flourished.

In the 18th century, Danish literature was dominated by Norwegian-born Ludvig Holberg, whose philosophic writings were inspired by John Locke and the French Encyclopaedists, his contemporary satire by Jonathan Swift and his comedy by Molière and the Italian commedia dell'arte. Another notable figure of the period was poet and hymn writer Thomas King.

Among the 19th-century writers were the religious philosopher Søren Kierkegaard; N.F.S. Grundtvig, who was the founder of the folk high school; the critic Georg Brandes; the lyrical novelist Jens Peter Jacobsen; and Hans Christian Andersen. All of these writers produced works valued by the world at large.

The Nobel Prize for Literature was awarded to the Danish novelist Henrik Pontoppidan in 1917 and, in 1944, to Johannes Vilhelm Jensen for his novel *Den lange rejse* (*The Long Journey*). Other 20th-century writers of note include story writer Karen Blixen (Isak Dinesen; 1885–1962); novelist Martin A. Hansen; and poet and novelist Klaus Rifbjerg.

From the late 1960s pornography was the one area of the Danish "literary" scene that fell under the spotlight of international notoriety. The laxity of Denmark's censorship laws, the subsequent increase in production of pornographic material, and the claimed drop in public interest and sex crimes gave the rest of the European world much matter for contemplation.

The theatre. The Danish theatre, like others in Europe, stemmed from medieval morality plays and court entertainments. The first Danish-speaking theatre was opened in Copenhagen in 1722 in Lille Grønnegade, followed in 1748 by the establishment of the Royal Theatre at Kongens Nytorv, which remained under court patronage for a century. In 1848 it was taken over by the state and is now administered by the Ministry for Cultural Affairs. An annex, the so-called New Stage, was built in 1931. These two theatres are subsidized by the state and produce drama, ballet, and opera.

Since World War II the number of private commercial theatres has been declining. In 1970 there were eight left in Copenhagen. The three large provincial towns of Århus, Odense, and Ålborg have theatres subsidized by the local authorities. Besides a relatively large number of classical and modern Danish plays, the theatrical repertoire includes much that is current in England, the United States, Germany, and France.

Ballet

The Danish ballet goes back more than 200 years, but it is only through its youngest generation of dancers in the style of choreographer August Bournonville that it has become internationally famous as the Royal Danish Ballet. An important contribution to its renown has been the regular annual Copenhagen Summer Festival held since 1950.

Music. Music in Denmark has been practiced since ancient times. Medieval sagas and chronicles tell of harp playing and horn blowing. Splendid relics of Bronze Age instruments have been found in the form of elaborate bronze lurs (trumpets of mammoth tusk shape). Denmark supports 10 symphony orchestras; two of the more important are the Danish Radio Symphony Orchestra and the Royal Orchestra. Musicians and singers are trained at the Royal Danish Conservatory in Copenhagen, at four other conservatories, and at the Opera Academy.

Fine arts. The Royal Danish Academy of Fine Arts was established in 1754 and has produced men of international repute, especially among its sculptors and architects. Among them are the sculptor Bertel Thorvaldsen and more recently the sculptor Robert Jacobsen and the architects Arne Jacobsen and Kaj Gottlob. Danish schools of arts and crafts have also produced talented young designers, and Danish industrial and handicraft design enjoys a high international reputation. Famous craft concerns include the Royal Copenhagen Porcelain Factory, Bing and Grøndahl, the glassworks Holmegard and Kastrup, and the furniture manufacturers Fritz Hansens Eftf.

Communications. The Danish press dates back to 1666. The guarantee of its freedom was secured in the 1849 constitution, and by 1914 each region had newspapers representing the four main political parties. A large number of these publications were forced into liquidation after World War II. In 1970 there were about 60 independent newspapers with a total circulation of nearly 2,000,000.

Radio Denmark, established in 1925, is an independent public institution exercising a monopoly on all radio and television broadcasting in Denmark. No private individuals own shares in the institution. There are three radio channels and one television channel covering the whole country and all owners of radios and television sets (248 per 1,000 in 1968) pay license fees to Radio Denmark.

OUTLOOK AND PROSPECTS

The traditional dependence of Denmark's 5,000,000 people on foreign markets had, before the agreement (approved 1972) to take the nation into the EEC the following January, placed the nation in a somewhat cruel dilemma and pointedly underlined the inherent tension between its Nordic and European interests.

Though a member of NATO, Denmark has traditionally sought neutrality; and though a member of EFTA, it had for some time cast longing glances at the EEC. The change from its traditional agricultural base to an industrial one has given greater impetus to Denmark's search for new agricultural and industrial markets. Thus, its sponsorship of the ill-fated Nordic Economic Union (Nordek) and its anxiety to join the EEC were based largely on the need to diversify its outlets, to lessen the dependence of its agricultural exporters on Great Britain and that of its industrial exporters on Nordic neighbours such as Sweden.

It seemed that much of the 1970s would be devoted to working out these complex problems.

BIBLIOGRAPHY. *Denmark: An Official Handbook*, edited in cooperation with the Royal Danish Ministry of Foreign Affairs (1970), gives concise and accurate information about Denmark: its geography, history, institutions and occupations, science, art, and general culture. This book also contains a bibliography of 850 titles, covering all major publications and periodicals about the subjects. Other works particularly recommended are NIELS NIELSEN (ed.), *Atlas of Denmark*, 2 vol. (1949–62); V.E. KAARIS, *Do You Know Denmark?*, 12th rev. ed. (1968), Denmark in pictures with descriptive text in English and Danish; E.T. APPLETON, *Denmark* (1968), a tourist guide; JOHANNES BRONDSTED, *Vikingerne* (1960; Eng. trans., *The Vikings*, 1965); PALLE LAURING, *A History of the Kingdom of Denmark*, 3rd ed. (1968); BENT FURSTNOW-SORENSEN, *Social Conditions in Denmark*, 6 vol. (1968, revised periodically); COPENHAGEN, THE MINISTRY OF

HOUSING, *Housing in the Nordic Countries*, 2nd ed. (1968); and TOBIAS FABER, *The New Danish Architecture* (1968). *Facts About Denmark* (1969, revised periodically), provides a brief account of most aspects of Danish society. See also the *Statistisk Årbog* (annual), the statistical yearbook of Denmark, with English subtitles and summaries; the *Yearbook of Nordic Statistics*, edited by the Nordic Council; *Danish Journal* (quarterly), and the *Economic Survey of Denmark* (annual), both issued by the Royal Danish Ministry of Foreign Affairs; and *Monetary Review* (quarterly), issued by Danmarks National bank (text in English).

(A.M.Sl.)

Density Currents

Density currents are currents in liquids or gases that are kept in motion by the force of gravity acting upon relatively small differences in density. A density difference can exist between two fluids (or gases) or between different parts of the same fluid and is caused by a difference in temperature, salinity, or sediment concentration. Density currents in nature include large-scale ocean currents, currents in lakes and reservoirs that are entered by rivers, and currents in meteorological frontal systems (see WINDS AND STORMS). In each case that part of the fluid that is either colder, more saline, or contains more suspended sediment than the other will be more dense and under the effect of gravity will flow beneath the less dense fluid. A cold air mass will flow beneath a warm air mass when the two converge, and a river will sometimes plunge beneath the surface of a lake and flow along its bottom in accord with this principle. The term density current is not used, however, when the difference in densities involved is a very large one. A river, for example, contains water that is about 800 times as dense as the air above it, but the river is not termed a density current relative to the air.

Density currents in which the density difference is caused by suspended sediment are called turbidity currents. Small turbidity currents have been studied in the laboratory and have been observed directly in lakes. Indirect evidence suggests that very large-scale turbidity currents occur in ocean basins. These large currents are thought to be caused by the slumping of sediment that has piled up at the top of the continental slope, particularly at the heads of submarine canyons. Slumping of large masses of sediment creates a dense sediment–water mixture, or slurry, that then flows down the canyon to spread out over the ocean floor and deposit a layer of sand in deep water. Repeated deposition forms submarine fans, which are analogous to the alluvial fans found at the mouths of many river canyons. Sedimentary rocks that are thought to have originated from ancient turbidity currents are called turbidites and are common in the geological record.

Density currents are of considerable practical importance. Deposition of sediment from turbidity currents in lakes may lead to a rapid decrease of reservoir capacity. Industries discharging large volumes of hot or polluted water may create density currents that have many adverse effects on neighbouring human or animal communities. On the other hand, proper design and placing of sewage outlets in the deeper, cooler water below the warm surface layer of the sea may permit formation of a sewage deepwater mixture that will not rise to the surface because its density is greater than that of the surface waters.

Because of the practical importance of density currents many experimental studies of their properties have been undertaken. Wilhelm Schmidt, a Viennese meteorologist, first studied small saltwater currents in the laboratory in 1910 in an attempt to understand the movement of cold fronts. Much of the knowledge of the hydraulics of density currents, however, stems from more sophisticated experiments since 1950.

HYDRAULICS OF DENSITY CURRENTS

If a dyed salt solution is released from a lock into a horizontal channel filled with pure water to the same depth as the lock, then the salt solution will flow beneath the pure water because it is more dense, and the pure water will flow into the lock replacing the solution (Figure 1). A notable feature of such flows is the formation of a distinct head at the front of the density current. After the current has fully developed, the head has a depth or thickness (d_2) about twice that of the current behind the head (d_1). It has been shown both theoretically and experimentally that the velocity of the head, v, may be expressed in terms of the density of the overlying fluid, ρ, the density difference between the two fluids, $\Delta\rho$, and the acceleration due to gravity, g:

$$v = 0.7 \sqrt{(\Delta\rho/\rho\, g d_2}.$$

From B.J. Daly and W.E. Pracht, *Physics of Fluids*, vol. 11, no. 1 (1968), p. 23

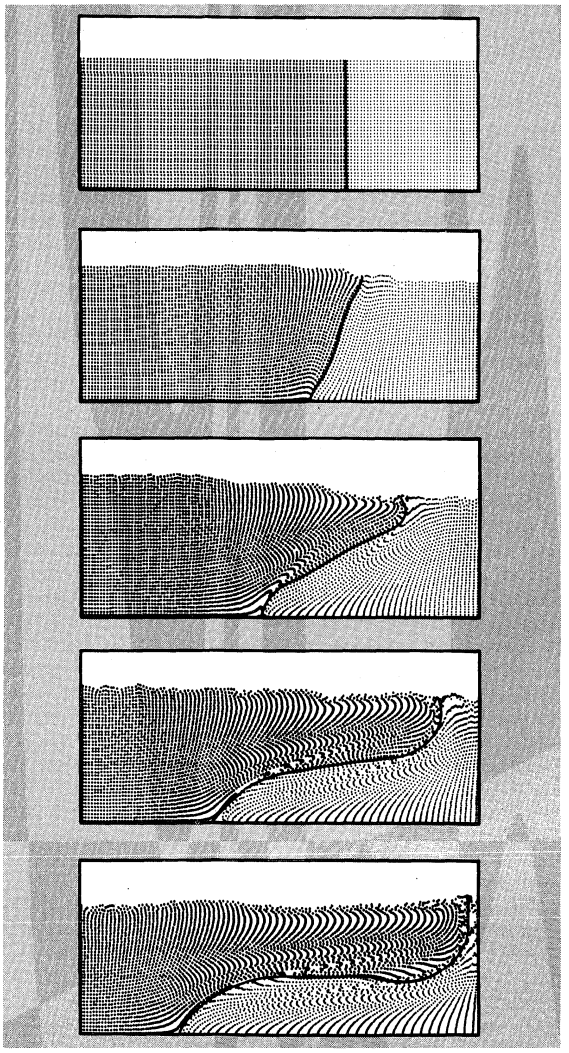


Figure 1: Motion of fluid particles in a density current surge. The denser fluid is on the right and the flow begins when the partition between the denser and the lighter fluids is removed.

Density current surges of this type may be created by the opening of river locks in estuaries or they may be brought about by catastrophic events in nature such as the slumping of an unstable mass of sediment down the continental slope.

If the density current is not a transient surge, then steady flow may be established. This type of flow has been studied by introducing a constant discharge of denser fluid at the upstream end of a tilted channel that is initially filled with lighter fluid. At first a head forms that is very similar in shape and behaviour to the head formed in density current surges. Behind the head, a thinner current is established that is steady in character, almost uniform in thickness, and has an average velocity, u. In small-scale laminar flows (nonturbulent flow, in which motion is best represented by parallel streamlines), the boundary between the denser and lighter fluids may be sharp and

Flow boundaries and mixing

planar, or waves may form and travel along the interface downstream. In large-scale turbulent flows (which are most common in nature), the boundary between the denser and lighter fluid is not sharp and distinct but is a zone of mixing between the two fluids. It has been established that in such a density current, below the velocity maximum, the velocity profile is similar to that observed in a river channel (see FLUVIAL PROCESSES). The exact nature of the profile will depend on the roughness of the bottom. In the part of the underflow above the velocity maximum, however, the velocity decreases due to mixing, in the same manner as at the boundary of a fluid jet, and the velocity profile is Gaussian in character. Above the zone of mixing, the lighter fluid is not at rest but is pulled along or entrained by the underflow (Figure 2).

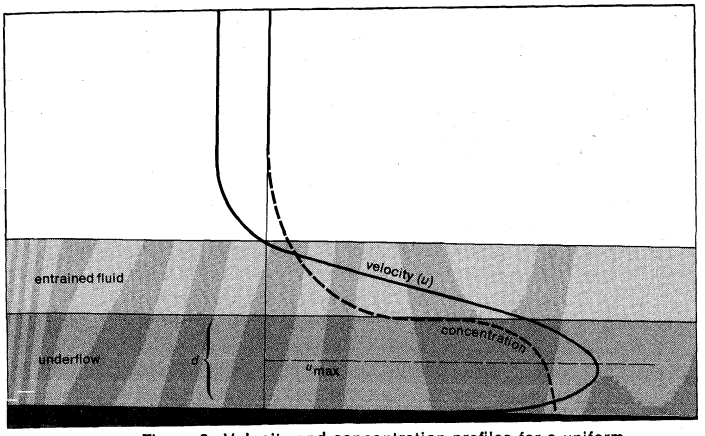


Figure 2: Velocity and concentration profiles for a uniform, steady density underflow of thickness (d). Concentration may be of salt or suspended sediment.

The degree of mixing that takes place at the interface can be related to the densiometric Froude number, a dimensionless parameter involving flow velocity and density, which may be expressed:

$$F = u/\sqrt{(\Delta\rho/\rho)\, gd},$$

in which u is average velocity of flow, d is depth, and $\Delta\rho/\rho$ and g are as previously given. An alternative commonly used by meteorologists is the Richardson number, R_i, which also is dimensionless and is expressed

$$R_i = 1/F^2.$$

For a density current flowing down a slope, the Froude number depends mainly on the degree of slope. There is maximum mixing at the interface on a steep slope and very little mixing where the slope is slight (less than about half a degree). There is also some evidence that the degree of mixing depends on the scale of the flow, with mixing being smaller for large-scale flows. These results make it possible to understand how density currents can preserve their identity, with very little dissipation of the density difference by mixing, even for flows that travel very long distances.

The average velocity, u, of a uniform density current is given by an equation similar to that for rivers (see FLUVIAL PROCESSES):

$$u = C\sqrt{(RS)},$$

in which S is the slope, R is the hydraulic radius (approximately equal to the depth, d, for a wide current), and C is a coefficient that depends on the density contrast and a friction factor. The friction factor, in turn, depends not only on the roughness of the bottom but also on the degree of mixing at the upper interface, and therefore on the Froude number, or slope.

Sediment transport

It has been postulated that there must be a condition for which the turbulence induced by the flow is just sufficient to maintain in suspension all of the sediment in a turbidity current, so that no sediment is deposited on or eroded from the bottom. This condition has been called auto-suspension, but the criteria for auto-suspension are not yet fully understood. Indirect evidence suggests, however, that large turbidity currents can transport clay, silt, and fine sand on slight slopes for very long distances, so that the conditions of auto-suspension must be approached very closely in these flows.

By courtesy of R.G. Walker

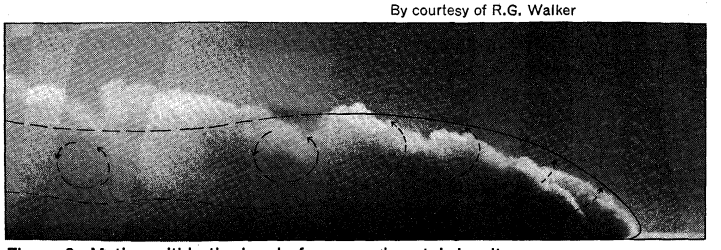


Figure 3: Motion within the head of an experimental density current, showing characteristic shape of head, upward diverging flow, breaking away of eddies from the back of the head. Depth of water is about 30 centimetres.

Observation of tracer particles in experimental flows indicates that the pattern of flow in the head of a turbidity current is similar to that shown in Figure 3. The head of the current is a region of diverging flow, where upward components tend to sweep material off the bottom and circulate it around inside the head. It might be expected, therefore, that the head of the current is not a region of deposition of particles but, rather, a region in which there may be some erosion of the bottom, even in flows in which there is deposition of particles taking place behind the head. Based on experimental observations, this conclusion is consistent with structures observed in beds of sediment thought to have been deposited by natural turbidity currents.

The sediment in turbidity currents cannot be held in suspension indefinitely. It is deposited, even in fast, eroding currents, from the tail of the current, where the thickness and velocity of the current begin to diminish as the supply of suspended sediment from upstream is reduced. Moreover, the current itself must begin to decelerate as the bottom slope becomes reduced. The first sediment to be deposited on the sea bottom is the coarsest; and the grain size decreases upward, producing what is termed a graded bed. This grading is caused by the gradual decay of turbulent eddying, a decrease of velocity from the head to the tail of a given flow, and the transfer of the coarsest material to the head of the current during the early stages of a flow.

DENSITY CURRENTS IN LAKES AND RESERVOIRS

It was a Swiss engineer who first observed that where the glacial meltwaters of the Rhône River enter Lake Geneva, they are denser than the lake water because they are both cold and highly charged with fine-grained sediment (glacial flour). As a result, the river water does not mix with the lake water but plunges below the surface and flows to a depth of over 300 metres (about 1,000 feet) down a subaqueous channel that extends for ten kilometres (about six miles) down the front of the Rhône Delta. The channel has a depth below its rim of almost 60 metres and the rim has subaqueous natural levees, which rise over five metres above the surrounding part of the lake floor. Two hypotheses for the origin of the channel were first considered: that it was produced by building up of the sides by deposition or that it was produced by subaqueous erosion of the delta slope. A similar phenomenon was also described for the Rhine Delta in Lake Constance.

Findings at Lake Mead

In 1935 the Hoover Dam was completed on the Colorado River in Nevada and Arizona, and Lake Mead began to form as the canyon was inundated upstream from the dam. In the first year of operation, when the reservoir was 110 to 145 kilometres (70 to 90 miles) long, it was observed that at three different periods turbid density currents formed at the head of the lake and flowed in the old river channel along the bottom of the lake all the way to the dam, where they were discharged through tunnels near the base of the dam. Analyses of the sediment and dissolved salts in the underflow at the dam indicated that

very little mixing with the lake water had taken place, even over this great distance of travel. Records of turbidity levels above and below the dam showed that the density currents flowed through the reservoir in about one week.

Subsequently, the tunnels near the base of the dam were closed, but further investigations were carried out in the lake, including the measurement of temperature, suspended sediment, and velocity profiles near the bottom. From these surveys it was established that underflows were formed whenever the density of the inflow was greater than the density of the lower lake levels: density differences of as little as one part per thousand were sufficient. Underflow was almost continuous during the fall and winter, and was intermittent at other times. The downward plunge of the underflow could clearly be seen at the head of the lake, where it was marked by an accumulation of floating logs and debris at the convergence of the underflow and the reverse circulation, which was induced in the upper parts of the lake water. Most underflows did not travel the full length of the lake, presumably because of the loss of density due to the deposition of the coarser silt in suspension. Underflows that did travel through the lake carried only very fine sediment in suspension (over 90 percent of the sediment being finer than 20 microns or 0.02 millimetre, as determined after dispersion). Velocities of the underflows ranged from less than 9 centimetres per second (cm/sec) to more than 30 centimeters per second. Thickness of the underflows was generally only a few metres, so that the flows were restricted to the old submerged channel of the Colorado River. Density of the underflows was generally less than 1.05 grams per cubic centimetre (g/cc), and the contact with the clear water above was quite sharp. The slope of the bottom was about one to two per thousand.

Density currents in Le Sautet Reservoir on the Drac River, in France, arise from concentrations of about 0.01 to 0.05 grams per cubic centimetre suspended sediment. The currents are two to ten metres thick and flow at velocities of about 5 to 15 centimetres per second down slopes of about .01 or half a degree.

Temperature contrast

Density currents have been recorded and studied in many other reservoirs. Density currents caused by temperature differences have proved to be of economic importance in some reservoirs. Flow into reservoirs on the Cumberland and Clinch rivers in Tennessee is controlled from storage dams located upstream, which release cold water during the summer. This cold water forms density currents when it enters the reservoirs because of a temperature contrast of 15° C (59° F) or more. Structures have been designed which permit the diversion of the density underflow and preferential withdrawal of the denser water for use as a coolant in power plants.

Turbidity currents in lakes and reservoirs are of relatively small size and low density, travel at low velocities, and carry only fine-grained sediment in suspension. They are generated by inflow of muddy river water and continue to flow over long periods of time, whereas most oceanic turbidity currents are probably transient phenomena generated by relatively short-lived episodes of submarine slumping.

DENSITY CURRENTS IN THE SEA

Salinity and temperature currents. The composition of seawater is remarkably constant, at least with respect to the relative proportions of the major dissolved salts. Locally, the concentration of salts may be changed by the discharge of large rivers, by melting of ice, or, more importantly, by evaporation of water from the sea surface. There are also large variations in temperature, particularly in the surface layer of the oceans.

These salinity and temperature variations produce stratification in the oceans (see OCEANS AND SEAS). Below the surface layer, which is disturbed by waves and is lighter than the deeper waters because it is warmer or less saline, the oceans are composed of layers of water that have distinctive chemical and physical characteristics, which move more or less independently of each other and which do not lose their individuality by mixing even after they have flowed for hundreds of miles from their point of origin.

Salinity currents in the Strait of Gibraltar

An example of this type of density current, or stratified flow, is provided by the water of the Mediterranean Sea as it flows through the Strait of Gibraltar out into the North Atlantic. Because the Mediterranean Sea is enclosed in a basin that is relatively small compared with the ocean basins and because it is located in a relatively arid climate, evaporation exceeds the supply of fresh water from rivers. The result is that the Mediterranean contains water that is both warmer and more saline than normal deep-sea water, the temperature ranging from 12.7 to 14.5° C (55 to 58° F) and the salinity from 38.4 to 39.0°/₀₀ (parts per thousand). Because of these characteristics, the Mediterranean water is considerably denser than the water in the upper parts of the North Atlantic, which has a salinity of about 36°/₀₀ and a temperature of about 13° C (55° F). The density contrast causes the lighter Atlantic water to flow into the Mediterranean in the upper part of the Strait of Gibraltar (down to a depth of about 200 metres) and the denser Mediterranean water to flow out into the Atlantic in the lower part of the strait (from about 200 metres to the top of the sill separating the Mediterranean from the Atlantic, at a depth of 320 metres). Because the strait is only 20 kilometres wide, both inflow and outflow achieve relatively high speeds. Near the surface the inflow may have speeds as high as two metres per second and the outflow reaches speeds of over one metre per second at a depth of about 275 metres. One result of the high current speeds in the strait is to cause a considerable amount of mixing, which reduces the salinity of the outflowing Mediterranean water to about 37°/₀₀. The outflowing water sinks to a depth of over 1,000 metres, where it encounters colder, denser Atlantic water. It then spreads out as a layer of more saline water between two Atlantic water masses and extends as far as the Azores before the salinity is reduced to about 35°/₀₀ by mixing. During World War II, submarines commonly used the salinity currents to traverse the Strait of Gibraltar with engines silent.

Turbidity currents. Density currents caused by suspended sediment concentrations in the oceans are relatively short-lived, transient phenomena. They occur at great depths and consequently are much more difficult to observe than turbidity currents in reservoirs. In fact, large scale turbidity underflows have never been directly observed in the oceans, and their very existence is somewhat hypothetical. There exists, nevertheless, much evidence that indicates that very large turbidity currents are formed in the present oceans and that, although their occurrence may be infrequent, they produce important physiographic effects.

Evidence for currents

The evidence for the present and past existence of turbidity currents in the oceans may be briefly summarized: (1) Telegraph cables have been broken in the deep sea in a sequence that indicates some disturbance at the bottom moving from shallow to deep water at speeds of the order of 10 to 40 knots (11 to 46 miles per hour). The trigger for this phenomenon is commonly, though not exclusively, an earthquake near the edge of the continental slope. The only disturbance that seems capable of being transmitted downslope at the required speed is a large turbidity current. The best known example of such a series of cable breaks took place in the North Atlantic following the 1929 earthquake under the Grand Banks of Newfoundland (Figure 4), but other examples have been described from the Magdalena River Delta (Colombia), the Congo Delta, the Mediterranean Sea north of Orléansville and south of the Straits of Messina, and Kandavu Passage, Fiji. (2) Cores taken from the sea bottom in the area downslope from cable breaks reveal layers of sand interbedded with normal deep-sea pelagic or hemipelagic oozes (sediments formed in the deep sea by quiet settling of fine particles). In the case of the cable breaks south of the Grand Banks, a large diameter core taken from the axis of a submarine canyon in the continental slope contained one centimetre (about ¼ inch) of gray

clay underlain by at least 20 centimetres of gray pebble and cobble gravel. Cores farther south showed a graded layer, about one metre thick, of coarse silt and fine sand. The presence of these gravel and sand layers is consistent with the hypothesis that they were deposited by the turbidity current that broke the cables. (3) Coring

Ocean floor sediments

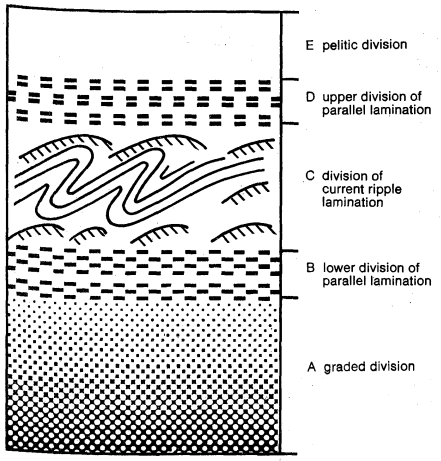
has revealed layers of fine-grained sand or coarse silt at many other localities in the abyssal plains of the oceans. These layers are generally moderately well sorted and contain microfossils characteristic of shallow water that are also size sorted. In some cases the layers are laminated and arranged in a definite sequence (Figure 5). It is clear that the sand forming these layers has been moved down from shallow water, and in many cases the only plausible mechanism appears to be a turbidity current. (4) At the base of many submarine canyons there are very large submarine fans. Deep-sea channels on the fan surfaces extend for many tens of kilometres and have depths of over a hundred metres and widths of a kilometre or more. Submarine levees are a prominent feature, and these project above the surrounding fan surface to elevations of 50 metres or more. The gross characteristics of such channels suggest that they were formed by a combination of erosion and deposition by turbidity currents.

Turbidites

(5) Thick deposits of interbedded graded sandstones and fine-grained shales are common in the geological record. In some cases there is good fossil evidence that the shales were deposited in relatively deep water, perhaps as much as several thousand metres deep. Relatively deepwater deposition is also suggested by the absence of sedimentary structures characteristic of shallow water. The interbedded sandstones, however, contain shallow-water fossils that are sorted by size, have a sharp basal contact with the shale below and a transitional contact with the shale above, and display a characteristic sequence of sedimentary structures (Figure 5). The structures include erosional marks made originally on the mud surface but now preserved as casts on the base of the sandstone bed (sole marks) and internal structures including some or all of the following: massive graded unit, parallel lamination, ripple cross-lamination or convolute lamination, and an upper unit of parallel lamination.

This combination of textural and structural features can be explained by deposition from a current that slightly erodes the bottom and then deposits sand that becomes finer grained as the velocity gradually wanes. The properties inferred from these ancient sandstone deposits are consistent with the properties of turbidity currents inferred from laboratory experiments.

In spite of the convincing nature of the evidence, there



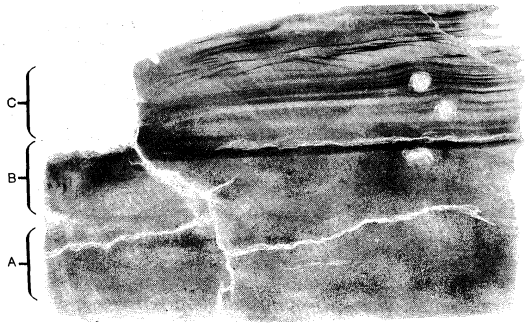


Figure 5: *Turbidity currents.*
(Top) Ideal sequence of sedimentary structures deposited in a bed by a turbidity current. (Bottom) Radiograph of a turbite off southern California showing sequence of sedimentary structures. (A) Graded division. (B) Lower division of parallel lamination. (C) Division of current ripple lamination.
By courtesy of (bottom) R.G. Walker; adapted from (top) A.H. Bouma, *Sedimentology of Some Flysch Deposits*

From H.W. Menard, *Marine Geology of the Pacific* (© 1964); used with permission of McGraw Hill Book Company

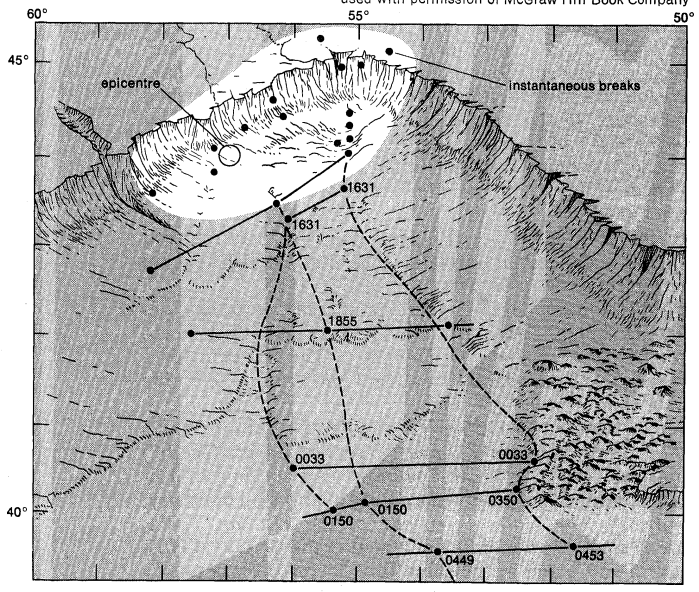


Figure 4: Location of cables (heavy lines) and times of breaks following the Grand Banks earthquake of 1929.

Summary of views on turbidity currents

are still some objections to the turbidity current hypothesis. Most geologists and oceanographers accept that such currents exist and that they are important agents of erosion and deposition, both in modern and ancient seas, but some believe that the turbidity current hypothesis has been overworked. Many geologists have suggested, for example, that large submarine canyons (*q.v.*) have been eroded by turbidity currents. Direct proof is lacking, however, and attempts to trigger turbidity flows by underwater detonation have thus far been unsuccessful.

Moreover, a growing body of evidence suggests that currents flowing parallel to submarine contours exist in many ocean basins. These bottom currents have been observed in a few cases, and velocities from 20 to 50 centimetres per second have been recorded. These currents can produce some of the features that previously had been attributed to turbidity current action.

Finally, nearly all features of sands that are produced by turbidity currents can be formed by shallow water action, such as fluvial processes (*q.v.*). Hence the problem of discriminating between deposits of turbidity currents and deposits of other current types is quite complex and requires a careful assessment of all lines of evidence in each case. Some ancient sandstones have been interpreted as "fluxoturbidites" because the sedimentary structures and other properties suggest a transporting agent intermediate between turbidity currents and large-scale slumping and sliding of sediment. Future studies will no doubt resolve many of the problems connected with turbidity current deposits and will provide further insight to the hydrodynamics of density flows and their possible economic utilization.

BIBLIOGRAPHY. D.R.F. HARLEMAN, "Stratified Flow," in V.L. STREETER (ed.), *Handbook of Fluid Dynamics* (1961), a

review article on the hydraulics of density currents; H.W. MENARD, *Marine Geology of the Pacific*, ch. 9 (1964), a review of turbidity currents, written from the point of view of marine geology; "Symposium on Density Currents," *Proc. Int. Ass. Hydraul. Res., 8th Congr.*, vol. 2 (1959), many technical papers on the hydraulics of density currents; A.H. BOUMA and A. BROUWER (eds.), *Developments in Sedimentology*, vol. 3, *Turbidities* (1964), symposium with extensive bibliography on sediments deposited by turbidity currents; CHIA-SHUN YIH, *Dynamics of Nonhomogenous Fluids* (1965), an advanced text on theoretical aspects of density currents.

(G.V.M.)

Dentistry

Dentistry is that profession concerned with the prevention and treatment of oral disease, particularly disease of the teeth and supporting tissues. In addition to general practice, dentistry includes such specialties as oral surgery, prosthodontics, periodontics, orthodontics, pedodontics and public health. These specialties are defined and considered later in this article.

The article includes sections on the history of dentistry, dental education, the practice of dentistry, dental specialties, and dental auxiliaries.

International organizations

Present status of dentistry. Associations of dentists, dental journals, and dental schools now exist in almost every country of the world. The Fédération Dentaire Internationale (International Dental Federation) was founded in 1900 and has met annually except in times of war. It has sponsored international dental congresses that are planned to meet every five years. Other international organizations include the International Association for Dental Research (Association Internationale pour la Recherche Dentaire) and the Association pour les Recherches sur les Paradentopathies (Association for Research into Periodontal Diseases), which was organized in 1932. *The International Dental Journal*, published by the Fédération Dentaire Internationale, was founded in 1950.

Within the general framework of the World Health Organization, which has now existed for about 25 years, the dental health program has progressed steadily from the beginning. A proposal for a joint review of stomatology (study of the structures, functions, and diseases of the mouth) and dental hygiene in collaboration with the International Dental Federation was made at the first World Health Assembly in 1948.

Between 1956 and 1961 the World Health Organization provided a number of short-term dental consultants to developing countries. The regional offices' dental health activities in the developing countries expanded, but the dental profession was not represented on their staff except in the regional office for the Americas, where a post of dental health officer was established in 1954. Between 1963 and 1967 a technical unit in the Division of Health Protection and Promotion was developed, a long-term program was set up, and new posts of dental epidemiologist and technical officer were established.

In 1963 dentistry was being practiced in 72 countries with a total population of 2,422,623,000. There were 371 dental schools, including 7 in Africa, 136 in the Americas, 59 in Asia, 129 in Europe, 33 in the Union of Soviet Socialist Republics and 7 in Oceania. These schools served a population of 6,530,000 per school and graduated 15,290 dentists. There were 432,237 practicing dentists with an average of 5,600 population per dentist. The disparity between the areas was as follows: in Africa, one dentist per 81,000 population; in the Americas, one dentist per 2,620; in Asia, one dentist per 17,000; in Europe, one dentist per 3,080; in the Union of Soviet Socialist Republics, one dentist per 3,900; in Oceania, one dentist per 3,160.

The prevalence of dental disease throughout much of the underdeveloped areas of the world, its interrelationship with general health, the paucity of dental manpower at the professional level, and the competing demands of other needs render the acceptance, training, and use of dental auxiliaries an urgent necessity.

Statistically valid dental health surveys of many countries do not exist, but scattered evidence from Thailand, Jamaica, Guatemala, Indonesia, Kenya, and many other countries, indicates that dental caries, periodontal disease, and malocclusions are widespread, and cancers of the soft and supporting tissues of the mouth fairly frequent.

Certain organizations, including the World Health Organization and Fédération Dentaire Internationale, and countries including New Zealand and the United States offer direct and financial assistance to many developing countries in the development of health educational and dental care services. For example, New Zealand has been active over the past 14 years in providing developing countries with the benefit of its experience in the use of dental auxiliaries or what is commonly known as school dental nurses. Direct assistance is being provided in the development of other public-health dental services to Ceylon, Malaysia, Singapore, Brunei, Thailand, Indonesia, Hong Kong, and Papua New Guinea. A number of dentists from the aforementioned countries have had the opportunity of studying the New Zealand system; a sizeable number of school dental nurses (63 in 1966) have received their training, others have received advanced training, enabling them to assist the developing countries in the establishment of their own training facilities.

Dentistry in Latin America is developing at a steady pace. Certain countries have quite adequate systems of dental education, and some dental schools have excellent training facilities. Government still has a great deal to say in the control of dental education, including admission of students, but this situation is changing. The dental profession is beginning to become an autonomous health profession, as in Canada and the United States.

The history of dentistry. The foundations of scientific medicine were laid in the 16th and 17th centuries, the same centuries that saw the beginnings of a separate dental literature and a recognized specialization in the practice of dentistry.

Barbers as dentists

In France during the early 1500s barbers treated teeth and performed almost all surgical procedures, although there was a Faculty of Medicine at the University of Paris at the time. During the century, surgeons were first licensed, and surgery began to pass out of the hands of the barbers, although they still continued to perform dental services. In England in 1544, Henry VIII granted a charter to the Barber-Surgeons' Company of England, to which women were admitted. During the same period some consideration was being given to the improvement of dental service. The first textbook in dentistry was published in Leipzig in 1530, and 50 years later students of dentistry were admitted to the University of France. By 1622 a number of men had been granted the title of surgeon-dentist, although the title was not fully established for a number of years after that. During the reign of Louis XIV, the surgeon-dentists formed a separate subdivision of the surgeons' guild and, the year after the subdivision was formed, it became law that those who wished to practice in the field of mouth surgery and artificial restoration had to pass prescribed examinations. At this time some women were permitted to practice dentistry in France, although the privilege was revoked during the mid-1700s.

In the 17th century in England dentistry was referred to as an independent specialty. The first English text on the subject of dentistry was published in 1685.

Publication of three highly significant textbooks occurred during the 18th century. The French dentist Pierre Fauchard's *Le Chirurgien Dentiste* ("The Surgeon Dentist") in 1728 put dental treatment on a more scientific plane than ever before and advocated a broader education for dentists. In 1756 a German dentist Philipp Pfaff's *Abhandlung von den Zähnen* ("Treatise on the Teeth") appeared, and in 1771 an anatomist and surgeon of England, John Hunter, who was giving lectures on dentistry, published *The Natural History of the Human Teeth.* Joseph Fox was appointed dental surgeon at Guy's Hospital in 1799, and at the same time lectures on dentistry were set up in Guy's and in a number of other areas.

No formal dental schools were established in England

until 1858, although dental hospitals had been established earlier. These dental hospitals were created as service centres for the poor. They were founded and supported by dentists. For the most part they were independent of medical schools and general hospitals. They did accept some students, however, who partly provided cheap labour to operate the hospital and whose fees were used to support the charitable work undertaken. The honorary dental surgeons were responsible for any teaching that was done. In 1858 the first dental school in the United Kingdom was established by the Odontological Society of London and the second school the following year by the College of Dentists of England. Both were private schools.

Early dental schools

At about the same time the Royal College of Surgeons arranged to hold examinations for licensure for dental surgery. It remained the examining body for dentistry in Great Britain for about 20 years. In 1878 the first Dentists' Act in the United Kingdom was passed and the General Medical Council established a register of those qualified to hold the title of dental surgeon. The act did not, however, prohibit individuals without these qualifications from practicing. The General Medical Council also prescribed a curriculum for the training of dentists. It required two years of preceptorship to learn dental mechanics and three years in a medical school and dental hospital. An inspector of the examination procedures at the colleges in 1897 recommended that the one examination be replaced by three—in preliminary science and dental mechanics and a final examination. These recommendations were put into effect and remained without much alteration until 1922.

While schools were being established and standards for licensure created in England, rapid development had also been going on in North America. Dentistry was being practiced in the United States by the late 1790s, and the first textbooks appeared in the early 1800s. By 1834 dental services were being provided by three distinct classes of people. First, there were those who had qualified by a course of study in the principles of medicine and surgery. Second, there were individuals who had obtained a preparatory course of medical study and then commenced practicing dentistry without having studied the mechanics of dental operations. These people were held in comparative respect since in due course they did obtain some degree of skill in their operations, although they were criticized for not having obtained this skill before actually inflicting their services on the public. The third group included a great number of charlatans—anyone who decided he would like to practice dentistry. It is this group that brought dentistry into disrepute.

Early dentistry in the United States

The first meeting of a dental society in the United States was held in 1834. Shortly thereafter two unsuccessful attempts to establish dental schools were made, the first in Kentucky and the second in New York City. The latter failed to get under way because of staff problems. No money was available at that time for salaries, and local dentists evidently felt that they could not spare time from their practice for teaching.

The first dental school in North America was established in Baltimore in 1840. Subsequently many other schools were started, many of which were operated privately or commercially. All of the early schools were separate from universities. The first, and for many years the only, school of dentistry in the United States associated with a university was opened at Harvard in 1867.

This was the era during which dentistry became organized in Canada. British schools had just started, and in the United States requirements for qualification and examination had only recently been drawn up; some standards for practice were being established and a number of schools begun.

The first dental act in Canada was passed by the legislature of the Province of Ontario in 1868. This act incorporated the Royal College of Dental Surgeons of Ontario and gave it the dual function of teaching and licensing. License requirements stipulated five years of practice in a dental office for registration. Except for a simple act that was passed in Alabama in 1841, the Ontario Act was the first law respecting dental practice on this continent. The regulations for registration were fairly simple by today's standards, but with the large number of itinerant dentists, who had no formal training whatever, moving back and forth across the Canadian–American border at that time, even this five-year requirement proved useful.

Dental education. *Predental programs.* There are 72 countries in the world in which undergraduate training in dentistry is available. Of these, 34 require predental training prior to acceptance into a school of dentistry. The predental training is in addition to primary and secondary education, which usually takes from 10 to 12 years. The required number of years in predental education varies from one to seven (such countries as Portugal, Italy, Spain, and Austria require from five to seven years of medical education before entering dentistry). In the two-year predental course training usually includes such studies as biology, chemistry, physics, and mathematics. Certain faculties of dentistry in Canada and the United States require a bachelor's degree in arts or science as a prerequisite for admission into a dental faculty.

Dental school and training. After the two-year predental course, training consists of four years in a faculty of dentistry to qualify as a doctor of dental surgery (D.D.S.) or doctor of dental medicine (D.M.D.). The program of studies during the four-year course includes the following biological sciences: human anatomy, biochemistry, bacteriology, histology, pathology, pharmacology, microbiology, and physiology, upon which the succeeding studies of the theory and techniques of dental practice are based. Studies required with respect to dental practice include restorative dentistry, prosthetics, orthodontics, surgery, preventive dentistry, medicine, dental public health, pedodontics, periodontics, radiology, clinical practice, and anesthesia.

THE PRACTICE OF DENTISTRY

Licensure requirements. The practice of dentistry is now well controlled, and in all countries of the world in which dentistry is practiced there is a licensing requirement. Licensing authority may be government, national dental organizations, state boards, or provincial boards. Some examples follow.

Canada. In Canada each province has its own licensing authority. This can be a college such as the Royal College of Dental Surgeons of Ontario, or an association, such as the Manitoba Dental Association. There is also a national authority, the Dental Examining Board.

Requirements in Canada

The university degree (doctor of dental surgery or doctor of dental medicine) does not in itself entitle the holder to practice but is an academic qualification for presentation to the licensing board under whose jurisdiction the holder desires to obtain a license to practice. The regulations of the provincial licensing boards vary but usually require an examination for licensing.

In some provinces the provincial board accepts the university examination results of the final year for licensing graduates of a dental school in that province. The national Dental Examining Board conducts an examination for persons who have been graduated prior to 1971 from approved dental schools, and who are Canadian citizens but does not require an examination for graduates of such schools in 1971 or later. The certificate granted by the board is valid for licensing in all provinces except one, where a clinical examination is required.

The national board still requires examinations of graduates of schools outside the United States and Canada.

United States. Licensing authority in the United States is vested in state boards of dental examiners. Thirty-three states require United States citizenship as a prerequisite; 12 states and the District of Columbia require the declaration of intent to become a citizen or the first papers as a requirement for admission to the dental licensure examination; six states and Puerto Rico do not require citizenship for such admission.

Requirements in the United States and Britain

Nationals with foreign diplomas may be admitted to practice if their diplomas were issued by a school approved by the American Dental Association and if they pass the state licensure examination.

The United Kingdom. Permission to practice is granted by the General Dental Council to those holding (1) a degree or diploma in dentistry or dental surgery conferred in Great Britain or Ireland, (2) a degree or diploma in dentistry or dental surgery granted elsewhere that has been recognized by the General Dental Council, or (3) a degree or diploma approved by the General Dental Council provided that these graduates have passed the statutory examination written under arrangements made by the General Dental Council.

Union of Soviet Socialist Republics. The status of the dentist of the type called stomatologist is on a par with that of the physician in the Union of Soviet Socialist Republics. He receives training equal to that of the physician and, in addition, is trained in operative dentistry, crown and bridge dentistry, prosthetics, exodontia, and anesthesia, both general and local. The stomatologist attends professional school for five years. His degree entitles him to operate on the hard as well as all soft tissues in the mouth and throat.

The stomatologist receives his education free of charge and is, in fact, paid to go to school. His income is sufficient to maintain himself and his family during his entire school year.

At present there are three classifications of dentists in the Soviet Union, known as the two-, three- and five-year dentists. The two-year dentist is a dental technician who studies for two years beyond high school, after which he is eligible to work in a dental laboratory. He cannot treat patients. The three-year dentist can treat patients, but his work is limited to restorative, prosthetics, and prophylactic dentistry. The five-year dentist is the stomatologist.

Requirements in West Germany, Italy, and Japan

Federal Republic of Germany. Dentists must hold a dental surgeon's diploma, which authorizes private practice without further examination. They must be registered by the local dental boards and by the health authorities. Persons holding foreign qualifications may obtain registration without taking a further examination by presentation and approval of their credentials.

Italy. The diploma in dentistry, which allows the use of the title of Specialist in Diseases of the Mouth, Teeth, and Jaws, constitutes license to practice. Holders of the diploma of Doctor of Medicine have passed examinations in dentistry and for this reason may also practice dentistry but do not have the specialist title.

Citizens of foreign countries, both those who have qualified in Italy and those who hold foreign degrees or diplomas, are authorized to practice if the relevant reciprocity exists between Italy and their country of origin.

Japan. Persons wishing to practice dentistry or dental surgery must be recognized by the Ministry of Health and Welfare (Dental Health Section, Medical Affairs Bureau, Ministry of Health and Welfare, Japanese Government). Applicants for registration must pass the national examination for dentists and obtain license to practice from the Ministry of Health and Welfare. These requirements must also be fulfilled by registered medical practitioners wishing to practice dentistry, by Japanese citizens, and by foreigners who have qualified in Japan.

Western medicine was first introduced into Japan from The Netherlands about the middle of the 17th century. In the 18th century German medicine was introduced, and from then on Japan made rapid and significant progress in the field of medicine. The introduction of modern American dentistry into Japan is thought to have occurred about 1903. Since then Japanese dentistry has been mainly patterned after that of the United States.

The first dental schools were established in the 1890s, and Japanese dentistry progressed thereafter independently, in co-operation with Japanese general medicine, though dentistry has never been considered a branch of medicine. Japanese dentistry has achieved success through its own basic research in clinical science techniques.

Types of practice. *Private practice.* In such countries as Canada, the United States, the United Kingdom, France, the Federal Republic of Germany, and Australia, dentists in private practice constitute the vast majority of all licensed dentists. It is expected that this type of practice will continue to be the prevalent type in most countries of the world, although dentists may shift even further to group practice or make more extensive use of auxiliaries.

In such countries as the Union of Soviet Socialist Republics dentists are employed by the government, and there is little private practice.

Shift from individual to group practice

Dental practice has changed markedly since 1920. It has changed significantly in the last quarter century, without a concurrent change in the basic dental curriculum as developed in 1935. Dental procedures shifted from the repair of a few teeth and extraction of teeth for the relief of pain in 1920 to an era of disease prevention in 1967. Dental practice has changed in the larger centres from the isolated private practice common in 1920 to a complex grouping of professional persons in a central location.

In Canada, for example, especially in the western provinces, group practices are developing whereby patients are examined on their first visit by one dentist who then allocates the patient to a suitable dentist within the group.

Extensive use is being made of dental hygienists, who receive the patient from the examining dentist. This type of practice is conducted like any other business with a business manager and competent clerical staff. There are sufficient dental hygienists to provide services such as preventive procedures, prophylaxis, scaling, X-rays, and dental-health education, and there is a supporting group of dental assistants, enabling the practice to run as a competent, smooth-working operation.

Another development that has occurred in dental-health-care services and could completely change the present concept of dental practice is the extension of the duties currently carried out by dental auxiliaries. New Zealand pioneered in the field with the creation of the dental nurse, an auxiliary trained to provide dental care for children and works without dental supervision. The United Kingdom has also developed the "dental auxiliary," who performs somewhat similar duties but under closer supervision.

In Canada and the United States pilot projects are being conducted to test the feasibility of using dental auxiliaries for certain operative procedures to increase productivity, quality, and general service to the public.

France may be taken as an example of the development of the practice of dentistry in continental Europe. The practice of dentistry in France as a chirurgien dentiste has, since 1892, been restricted to persons (1) of French nationality, (2) who hold a state diploma, and (3) who are registered with the Order of Dentists. The Order of Dentists is responsible for registration and discipline but is not concerned with dental education, which is controlled by the state through the common state diploma.

There are two types of dentists practicing in France, the chirurgien dentiste and the stomatologist. Stomatologists are practitioners who possess a diploma in medicine together with either the diploma in dental surgery or a certificate of special studies (two years) in stomatology. Specialization within the field of dentistry is not encouraged. There are no rules laid down for it nor are there any special courses or diplomas or titles.

Hospital dental practice. Three types of dental care are normally carried out in the hospital environment: (1) clinical procedures normally provided in a dental office, for ambulatory inpatients and outpatients; (2) bedside care for persons admitted for medical reasons; and (3) inpatient care for patients admitted to the hospital for purely dental conditions.

Dentists may treat patients in hospitals either privately, on a fee-for-service basis, or under some form of government program, such as the National Health Service in the United Kingdom or the Provincial Medicare Plan (surgery only) in Canada. Hospital dental services have for years been an integral part of dental-health care and dental education in the United Kingdom, and such services by hospital dental departments are expanding steadily in the United States and Canada.

Hospital dental departments are normally established in the same manner as any other hospital department and are headed by a chief of service, who has the same status as other chiefs of service within the hospital. In some instances, the chief of the dental department may be responsible to the chief of surgery. There are two types of hospital dental departments—one that is established in a teaching hospital, and the other in a general hospital with no teaching component. In the teaching hospital the dental department is associated with a faculty of dentistry and forms an integral part of the undergraduate curriculum and, if they exist, of the graduate and postgraduate programs. One of the chief purposes of hospital dental departments is to make available the service of consultants to other hospital departments and general practitioners. This service is most highly developed in teaching hospitals. Usually certain general dental treatment is provided for inpatients and outpatients.

Under the present regulations of the National Medicare Bill in Canada necessary surgical services must be performed in a hospital in order to qualify for payment, to the dentist, under the plan. Surgical services which are performed in a private office must be paid for personally by the patient. It follows that many more surgical services are now being performed in a hospital facility than ever before. This regulation can bring about the development of an ever-increasing number of hospital dental departments. In Canada in order for a hospital dental department to be recognized, it must be accredited by the Canadian Dental Association. A similar situation is in effect in the United States, where the 176 oral-surgery training programs are assessed and, if up to standard, approved by the American Dental Association's Council on Education.

Hospital dental services or departments are most prevalent in western Europe and the Union of Soviet Socialist Republics.

Public health practice. This section is concerned with governmentally supported health education and disease prevention, as in Canada and the United States. The provision of dental care for all, as in the Soviet Union, is discussed in the section *Governmental practice.*

Generally typical of dental public-health practice in Canada and in many areas of the United States is the program carried on in the Province of Ontario. In the early 1970s about 20 dentists trained in public health, about 70 hygienists, and numerous dental assistants carried out a preventive and educational program basically concerned with examination of children, recording basic dental conditions, and providing dental-health education. Cards sent home to parents advise them of the dental conditions of their child and stress any need for dental care. The care itself, except for families receiving social assistance, is provided by the family dentist. In addition, a detailed examination of a random sample of the children is carried out in conjunction with the screening to provide the data to develop statistical reports, required for a yearly evaluation of the program. The program also includes the application of fluorides to the teeth and, in certain areas, special efforts to encourage brushing. Dental-health education is carried out directly in the schools. The cooperation and active aid of the school principal and teachers is solicited in an attempt to have dental health stressed throughout the school year. A dental-health manual is made available to the teachers; pamphlets and posters are used extensively. Outside the school system hygienists regularly attend dental child-health conferences or well baby clinics, organized by the local health unit, where the importance of dental health is discussed with the mothers of young children. Lectures are also given to teachers' colleges and to any organization or group interested in the health of children.

Ontario program

Military practice. Most countries of the world provide dental-care service for their armed forces. The organization of such a service varies extensively. In Canada the Royal Canadian Dental Corps has the same status as the Royal Canadian Medical Corps, with a brigadier-general as the director. Military service for dentists in the United States is under the United States Public Health Service, the chief of service being an assistant surgeon general. In the United Kingdom dental care is provided by three separate dental branches—Navy, Army, and Air Force.

Governmental practice. In certain countries, such as the Union of Soviet Socialist Republics, all dentists are employed by the government. In certain other countries dentists are required to work a stated number of years for the government before being considered private practitioners of the type known in Canada and the United States. This requirement may be based on the fulfillment of an obligation for government financial support during undergraduate training, or there may be a government regulation that all dental graduates must work for the state a prescribed number of years. Another example of government practice is in the United Kingdom, where dentists are employed by local authorities to provide dental care under the Maternal and Child Welfare Services and the School Dental Service.

The employment of dentists on a salary basis for the general practice of dentistry is not extensive in the United States or Canada. At the national level it may be the provision of dental care for eligible Indians and Eskimos, care provided for war veterans, or inmates of penitentiaries. At a municipal level, dentists may be employed in a school dental service. Dentists in both Canada and the United States commonly agree to provide service for families who qualify for social assistance. They are paid on a fee-for-service basis; the fee schedule is usually set, normally after consultation with the profession, by the agency responsible for the social-service plan.

Government medical care was introduced in Japan in the late 1930s. This system was expanded until by 1962 almost the entire population was covered. There are limitations to the services offered by government medical care, as in orthodontics or in preventive dentistry.

Dental specialities and subspecialities. A specialist may be defined as a person who, because of graduate or postgraduate training, possesses special knowledge and skills of an expert nature, by reason of which he is capable of rendering a service above the average. The term therefore implies not only the qualities developed through limiting of practice but in addition the academic qualifications linking this special field with all related studies.

The status of the specialist

In most countries that recognize specialties in dentistry, the specialist is limited to practice in the specialty and must not carry out the practice of general dentistry. Where the specialty is thus limited the general dentist may refer patients, and a specialist's practice is mainly on a referral basis. In certain provinces in Canada specialists may now conduct a general practice.

Orthodontics. Orthodontics has for its objective the prevention and correction of malocclusion of the teeth and associated dentofacial incongruities. The orthodontist is commonly thought of as a dentist who specializes in straightening crooked teeth.

Orthodontics has been in existence since the days of the early Egyptians, but the more elaborate methods of treatment involving the use of bands and removable appliances have become prominent only since the beginning of the 20th century. The United States gave great impetus to this development and orthodontics was recognized as a specialty with the formation of the American Society of Orthodontists in 1900.

In practice the orthodontist may deal directly or indirectly with the alleviation or elimination of any one or more of the following: impairment of ability to chew; susceptibility to dental caries; periodontal disease (disease in the structures around the teeth) and other disturbances of the oral tissues; an unaesthetic facial appearance; dentofacial abnormalities of genetic, congenital, and environmental origin, including those resulting from surgery; shifted teeth and abnormal jaw relationship (these the orthodontist may correct prior to the construction of partial dentures or any other dental restorations); abnormal respiratory habits; and abnormal mental attitudes in relation to dentofacial aesthetics.

Currently the demand for this service has extended from the child patient to the mature adult, although human bone responds to tooth movement best in the person

under 18, and it is logical that the child will benefit from treatment more than the adult. In general, oral health and physical appearance are the two most important reasons for undertaking a course of orthodontic care. There is a great need for enlarged training facilities in this field at the postgraduate level.

Training in orthodontics is provided by a two-year postgraduate course open only to dentists who have had two years of general practice. A diploma is awarded upon completion of the course.

Dental care for children

Pedodontics. Pedodontics, analogous to pediatrics in medicine, is concerned with the dental care of children and adolescents.

Much of the routine of practice is centred about caries (tooth decay) control and involves the use of fluoride and dietary and hygienic instruction. The need to influence tooth positions presents the next most frequently encountered problem. The correction of incipient abnormalities in tooth alignment may obviate the necessity for lengthy treatment. Many pedodontists use growth-influencing techniques to correct jaw alignments. A working knowledge of children's behaviour patterns, patience, a knowledge of childhood physical and mental diseases and their oral ramifications, and a facility in the hospital environment round out the qualifications of the pedodontist. The postgraduate course in pedodontics is of two years' duration, leading to a diploma in pedodontics.

Periodontics. Periodontics is concerned with the prevention, diagnosis, and treatment of diseases of the periodontal tissues—the tissues that surround and support the teeth. These tissues consist mainly of the gums and the jaw bones and their related contiguous structures.

The most prevalent periodontal disease is periodontitis, commonly called pyorrhea, an inflammatory condition usually produced by local irritants. Periodontitis, if untreated, destroys the periodontal tissues and is a major cause for the loss of teeth in adults.

The advances of periodontics in the past quarter century have been mostly in techniques of treatment. It is believed that bacterial plaque, a hard layer of substances rich in bacteria that adheres to the teeth, is the factor responsible for most destruction of the gums and the tissues surrounding the teeth. Periodontists advocate removal of such plaque in a treated or healthy mouth by a specific regimen of controlled hygiene.

Prosthodontics. Prosthodontics is concerned with the restoration and maintenance of oral function, comfort, appearance, and health by the replacement of missing teeth and contiguous tissues with artificial substitutes, or prostheses.

Branches of prosthodontics

There are three main branches of the specialty, concerned, respectively, with removable prostheses, fixed prostheses, and maxillofacial prostheses. Maxillofacial prostheses are supplied to persons who have suffered congenital, traumatic, or surgical defects of the mouth, jaws, or associated facial structures.

A prosthodontist must be competent to make necessary judgments concerning, among other matters, social and psychological complications, nutritional imbalances, appropriate medications, the diagnosis and early detection of oral cancer, the diagnosis and treatment of many varieties of oral lesions, and the restoration of a normal appearance and of a normal ability to chew one's food and to speak. The substitutes for missing teeth should function in harmony with the remaining teeth to prevent their premature loss.

The proper fitting of oral prostheses requires a detailed knowledge of the anatomy of the head and neck, of the physiology of the neuromuscular system, and of the science of occlusion and jaw movements. It also requires skill in planning, mouth preparation, impression making, registration of jaw relations, try-in procedures, placement of the prostheses, and follow-up care.

The usual course in prosthodontics is of two years' duration, leads to a master's degree in prosthodontics, and is restricted to candidates holding a D.D.S. or equivalent with good standing from an approved institution.

Oral pathology. Oral pathology is the study of the causes, processes, and effects of oral disease, together with the resultant alterations of oral structure and functions. The oral pathologist need not treat a disease directly but through knowledge of the disease guides his associates to more effective therapy.

Oral surgery. Oral surgery deals with the diagnosis and with the surgery required by the diseases, injuries, and defects of the human jaws and associated structures.

Both dentists and physicians refer a wide variety of special dental problems to the oral surgeon. These may include the removal of impacted and infected teeth, the treatment of cysts, tumours, lesions, and infections of the mouth and jaws. In addition there are the more complex problems such as jaw and facial injuries, cleft palate, and cleft lip.

The program leading to the diploma in oral surgery is of three years' duration after the candidate has qualified as a dentist at an accredited faculty of dentistry.

Public health dentistry. Dental public health is recognized as a specialty in three Canadian provinces and is expected to be recognized by the Canadian Dental Association in the near future. In the United States and in Canada postgraduate courses in dental public health of one-year's duration lead to a master's degree in public health, in the United States, and a diploma in dental public health in Canada. The American Dental Association recognizes dental public health as a specialty if the holder of the master's degree proceeds to a further year of study in training and passes the examination of the American Board of Dental Public Health. Training in dental public health is available in the United Kingdom, but the specialty is not emphasized to the same degree in the rest of the world.

ANCILLARY DENTAL FIELDS

Dental hygienists. The hygienist is becoming an important figure in the campaign to reduce tooth decay and to improve physical well-being by promoting better care of the mouth.

Chief function of hygienists

The prevention of oral disease through education and treatment is the chief function of hygienists. The specific duties and services that they are allowed to perform depend on the bylaws of the provincial licensing bodies, the requirements of the dental offices in which they are employed, and the aims and objectives of the public-health programs in which they are engaged. At all times hygienists work under the effective supervision of a qualified dentist, and they are not permitted to establish their own practice.

Hygienists employed in dental offices remove deposits and stains from the patient's teeth and apply fluorides. In addition, they observe and record conditions of decay and disease for the dentist's information. Further duties may include the taking of X-ray photographs of parts of the mouth, which the hygienist develops and mounts. Another function of the hygienist is to promote dental health by advising on diet and nutrition and encouraging oral hygiene.

Hygienists employed by educational authorities assist school dentists by examining children's teeth and performing other duties as required. They may also visit classrooms to explain the importance of oral hygiene and to give instruction in the proper care of the teeth and gums.

In hospitals they perform mainly the same duties as for private practitioners.

In Canada five universities offer a two-year diploma course in dental hygiene. Plans are also proceeding to establish a four-year course in dental hygiene leading to a bachelor of science degree. In the United States diploma courses in dental hygiene are available at the university or junior college level. Programs are also available whereby a holder of a certificate in dental hygiene may proceed to a bachelor's or master's degree. In the United Kingdom and certain other European countries schools of dental hygiene have been established; these have a two-year course with entrance requirements below university level.

Dental nurses and dental auxiliaries. In New Zealand an auxiliary known as the dental nurse has been carrying

Training and functions of dental nurses

out a dental-care program for children for a number of years. The dental nurse receives minimal supervision but has proven to be equipped to provide a quite adequate dental-care program for children up to 13 years of age. The dental nurse is trained for two years in a special course for dental nurses with entrance requirements below the university level. There is a similar auxiliary in the United Kingdom who, after two years of training, provides, under more direct supervision than the New Zealand dental nurse, a dental-care program for children. The training is again a two-year course with entrance requirements below university level.

Dental assistants. The majority of dentists in private practice employ one or more dental assistants to provide such services as reception of patients, the keeping of records and accounts, chairside assistance for the dentist while he is treating patients, general upkeep of the office, developing of dental X-rays, and sterilization of instruments. The general trend is toward a greater utilization of auxiliary services.

Dental technicians and dental mechanics. The dental technicians, also called dental mechanics, make artificial crowns, bridges, dentures, and other dental appliances according to dentists' specifications. Work orders, accompanied by models or impressions of patients' mouths, state the exact requirements for each particular job. In large laboratories the various stages of manufacture are often divided and the technicians employed may specialize. Sometimes partially skilled persons are hired to work in limited aspects of production on an assembly line basis.

(R.A.Co.)

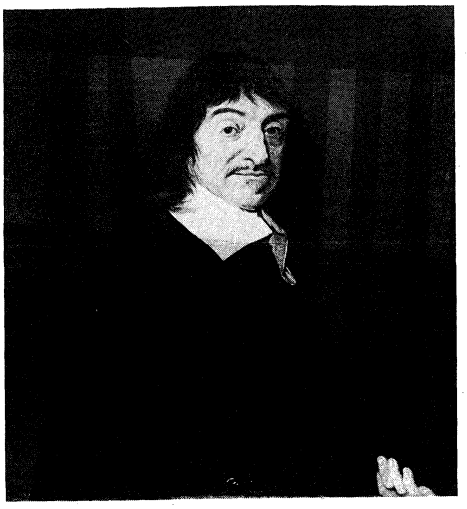
BIBLIOGRAPHY. Dental literature is extensive and international. Of some 1,958 dental periodicals that have existed since 1839, when the first dental periodical appeared, 770, of all sizes and kinds in at least 29 languages, were extant in the early 1970s. The United States accounts for 511 of these. The *Index to Dental Literature* (quarterly and cumulative) offers the most comprehensive coverage of the world dental literature. *Oral Research Abstracts* (monthly), and the *Advances* series in caries, oral surgery, orthodontics, pedodontics, periodontics, and prosthodontics derived from it, are comprehensive compilations of abstracts of current articles related to all phases of dentistry published in hundreds of journals throughout the world. *Dental Abstracts* (monthly) is a more selective compilation intended for general distribution but based on the same body of literature. *The Year Book of Dentistry* is an annual review of selected articles in the literature published since 1936. To select a bibliography from the tremendous number of works published on this subject is difficult. The bibliography that follows attempts to be representative only and to show the various facets of dentistry. PIERRE FAUCHARD, *Le Chirurgien Dentiste*, 2nd ed. (1746; Eng. trans., *The Surgeon Dentist*, 1946), is the all-time dental classic. H.M. GOLDMAN *et al.* (eds.), *Current Therapy in Dentistry*, 4 vol. (1964–70), correlates the latest research information with a reasonable immediacy in the practice of dentistry. LESTER W. BURKET, *Oral Medicine: Diagnosis and Treatment*, 6th ed. (1971); and DAVID F. MITCHELL, S. MILES STANDISH, and THOMAS B. FAST, *Oral Diagnosis/Oral Medicine*, 2nd ed. (1971), expose the foundation upon which the dental curriculum and dental practice rest. G.V. BLACK, *Operative Dentistry*, 9th ed., 2 vol. (1955), is a long-time classic; while B.H. BELL and D.A. GRAINGER, *Basic Operative Dentistry Procedures*, 2nd ed. (1971), provides a fine example of a general recent work. H.M. GOLDMAN and D.W. COHEN, *Periodontal Therapy*, 4th ed. (1968); and IRVING GLICKMAN, *Clinical Periodontology*, 4th ed. (1972), are the standard works on the subject. A.J. LAZARE (ed.), *Periodontal Therapy: A Review* (1967), reprints a number of classics that have appeared in the periodical literature. KURT H. THOMA, *Oral Surgery*, 5th ed., 2 vol. (1969), is probably the most comprehensive book on this subject; other outstanding works are W. HARRY ARCHER, *Oral Surgery*, 4th ed. (1966); and GUSTAVE O. KRUGER (ed.), *Textbook of Oral Surgery*, 2nd ed. (1968). NORMAN L. ROWE and H.C. KILLEY, *Fractures of the Facial Skeleton*, 2nd ed. (1968), is a classic text on the specific area of facial trauma. ROY G. ELLIS and KEITH W. DAVEY, *The Classification and Treatment of Injuries to the Teeth of Children*, 5th ed. (1970); DAVID B. LAW, THOMPSON M. LEWIS, and JOHN M. DAVIS, *An Atlas of Pedodontics* (1969); and RALPH E. MACDONALD, *Dentistry for the Child and Adolescent* (1969), are fine representatives of the numerous works in this area. S.N. BHASKAR, *Synopsis of Oral Pathology*, 3rd ed. (1969); W.G. SHAFER, M.K. HINE, and B.M. LEVY, *A Textbook of Oral Pathology*, 2nd ed. (1963); KURT H. THOMA, *Thoma's Oral Pathology*, 6th ed. (1970); and R.W. TIECKE, *Oral Pathology* (1965), give excellent coverage of the field of oral pathology. On preventive dentistry, see J.L. BERNIER and J.C. MUHLER, *Improving Dental Practice Through Preventive Measures*, 2nd ed. (1970); and A.E. NIZEL, *Nutrition in Preventive Dentistry: Science and Practice* (1972). T.M. GRABER (ed.), *Current Orthodontic Concepts and Techniques* (1969); P.R. BEGG and P.C. KESLING, *Begg Orthodontic Theory and Technique*, 2nd ed. (1971); and C.P. ADAMS, *The Design and Construction of Removable Orthodontic Appliances*, 4th ed. (1970), are recent texts in the extensive orthodontic literature. L.I. GROSSMAN, *Endodontic Practice*, 7th ed. (1970); SAMUEL SELTZER, *Endodontology: Biologic Considerations in Endodontic Procedures* (1971); and R.F. SOMMER, F.D. OSTRANDER, and M.C. CROWLEY, *Clinical Endodontics*, 3rd ed. (1966), are excellent standard texts in this field. F.M. MCCARTHY, *Emergencies in Dental Practice*, 2nd ed. (1972), outlines ways of preventing, recognizing, and treating emergency problems that occur in every dental practice. Excellent standard texts in prosthodontics are R.W. DYKEMA, D.M. CUNNINGHAM, and J.F. JOHNSTON, *Modern Practice in Removable Partial Prosthodontics* (1969); C.O. BOUCHER (ed.), *Swenson's Complete Dentures*, 6th ed. (1970); S.D. TYLMAN, *Theory and Practice of Crown and Fixed Partial Prosthodontics*, 6th ed. (1970); and J.M. BUCHMAN, *An Atlas of Complete Denture Prosthesis* (1970). E.W. SKINNER and R.W. PHILLIPS, *The Science of Dental Materials*, 6th ed. (1967), is now a classic work with a noteworthy companion, F.A. PEYTON and ROBERT G. CRAIG, *Restorative Dental Materials* (1971). H. SICHER and E.L. DUBRUL, *Oral Anatomy*, 5th ed. (1970); I. SCHOUR, *Noyes' Oral Histology and Embryology*, 8th ed. (1960); and H. SICHER and S.N. BHASKAR (eds.), *Orban's Oral Histology and Embryology*, rev. 7th ed. (1972), are excellent representatives of their fields. VINCENZO GUERINI, *A History of Dentistry* (1909), has not been superseded in English although closely followed by B.W. WEINBERGER, *An Introduction to the History of Dentistry* (1948). M.D.K. BREMNER, *The Story of Dentistry*, 3rd ed. (1954), is written in a more popular vein. B.W. WEINBERGER, *Orthodontics* (1926); and JOHN ENNIS, *The Story of the Federation Dentaire Internationale, 1900–1962* (1967), are two examples of more specialized histories in dentistry; while C. PROSKAUER and F.H. WITT, *Bildgeschichte der Zahnheilkunde* (1962), is an excellent pictorial history with descriptive material in five languages. J.M. DUNNING, *Principles of Dental Public Health*, 2nd ed. (1970); W.O. YOUNG and D.K. STRIFFLER, *The Dentist, His Practice, and His Community*, 2nd ed. (1969); R.K. STINAFF, *Dental Practice Administration* (1968); and S.H. WILLIG, *Legal Considerations in Dentistry* (1971), provide examples of the relationship between the dentist and society. G. GUSTAFSON, *Forensic Odontology* (1966); and T. FURUHATA and K. YAMAMOTO, *Forensic Odontology* (1967), illustrate the developing field of dental identification. C.O. BOUCHER (ed.), *Current Clinical Dental Terminology* (1963), is the present standard dictionary, though it should be supplemented by more recent glossaries for various areas in dentistry, such as oral surgery and periodontics.

Descartes, René

René Descartes, French mathematician and philosopher, was the first to liberate philosophical thought from the confines of tradition-bound Scholasticism. Hegel and many others have called him the father of modern philosophy. The proponent of a thoroughgoing dualistic philosophy, Descartes radically distinguished between the mind, perceived as indubitable in his famous statement, "I think, therefore I am," and the body (or matter in general), for which he was the first to provide a comprehensive explanation on the basis of purely mechanistic principles.

Early life. Descartes was born on March 31, 1596, in the village of La Haye (now called La Haye-Descartes) in the province of Touraine. His father was a counsellor in another province, Brittany, at the local parliament of Rennes; in modern terms he was a lawyer as well as a judge. Both he and his wife belonged to the same social class, the *noblesse de robe*, below the nobility itself, yet above the bourgeoisie, able to maintain its independence and privileges because the functions of lawyers were largely hereditary.

René was the fourth child of the marriage. His mother was to die a year later in childbirth. Little is known about his childhood. He was probably brought up by his grandmother and only saw his father, who remarried, during



Descartes, oil painting by Frans Hals, 1649. In the Louvre, Paris.
Cliché Musées Nationaux, Paris

the six months of the year when the courts at Rennes were on vacation. A delicate child, he also had a nurse to whom he was devoted and to whom he paid a pension up to the day of his death. His apparent want of robustness later allowed him indulgent exceptions to the rules in his collegiate days, but from his 20th year, Descartes enjoyed fairly good health. Based chiefly on the character and actions of the mature man, biographers have drawn the picture of a solitary and exceptionally intelligent boy. Undoubtedly, introspective detachment dominated his relations with his closest family. He did not, for instance, attend the marriage of either his brother or sister, nor was he present at his father's death.

All other evidence indicates a normal and happy upbringing. It is known that his father was amused by the precocious curiosity of his small son who plied him with questions about "the reasons of things and their causes" and that he called him "his little philosopher." Some 30 years later, Joachim Descartes was to declare that, of all his children, René was the only one who had given him cause for discontent and spoke scornfully of him as "only fit to be bound in calf-skin." It was the contemporary comment of a father, disappointed that his son had relinquished the quasi-noble profession of law and taken up the socially doubtful status of a writer.

Jesuit education

In 1604 Joachim Descartes made a decision about the education of René, then eight years old, that had a crucial effect on his development. The Jesuits had opened that very year a college, the Royal College, endowed by Henry IV, in the small town of La Flèche, north of Touraine. In the charge of these very competent educators, La Flèche was to become, in Descartes's own words, "one of the most celebrated schools of Europe." René was sent there and confided to the special care of Father Charlet, a distant relative. Charlet was a distinguished scholar, who soon became rector of the college; Descartes was to call him "his second father." The effect of the approximately ten years that Descartes spent at La Flèche can be estimated from his reactions in his works, in which he relates some of his debts as well as doubts, his gratitude as well as dissatisfaction.

The syllabus of studies followed the pattern laid down for all Jesuit schools. The first five years were devoted to the humanities (that is, the study of Latin and Greek and the classical authors), the French language with prose and verse composition, music, acting, and the accepted accomplishments of a gentleman—riding and fencing. Descartes certainly acquired a remarkable ability to write a fluent and clear prose, both in Latin and French. He later composed a short work on the art of fencing (he fought a duel) and also a small treatise on music. He read the classics and poetry with pleasure during his school days, but this was a passing interest that he dropped and later condemned as wasteful. His character and natural bent were essentially opposed to those of the learned scholar, in the sense of one who merely acquires handed-down knowledge.

Philosophy was the generic name given to the three final years of the curriculum. The first year consisted of logic, based on the syllogistic system derived from Aristotle, and moral philosophy, which consisted of a detailed analysis of another Aristotelian work, the *Nicomachean Ethics*. The study of physics, mathematics, and astronomy filled the second year, but the science then taught was little more than the theories of Aristotle, elaborated and developed by medieval commentators. Only in mathematics and astronomy did his teachers impart some of the more recent advances. The third year consisted of metaphysics, chiefly the philosophy of Thomas Aquinas, and the glosses of Jesuit commentators.

On these vital formative years, one can compose, from the sources available, two school reports. The first, from his professors, would note him as intelligent, hard working and well behaved, introspective and disinclined to be competitive, with a marked inclination and ability for mathematics. Because of his delicate health, he was allowed to stay long in bed in the mornings and even to miss some lectures, but this did not affect his progress, as he was capable of working alone and inclined to "meditate." The second report, according to Descartes's own account in later years, was quite different. By his graduation ("admitted into the ranks of the learned"), he was bewildered and utterly disappointed by what he had been taught. The subtle dialectical arguments and inconclusive and contradictory theories of his textbooks had left him with nothing but doubts and uncertainties. Only in his mathematical studies had he found some consolation "because of the self-evidence of its reasonings." The inquisitive child, the "little philosopher," was not only dissatisfied with his own baggage of knowledge, but he equally rejected the erudite lumber of his textbooks. Instead of considering himself a learned man, he thought of himself as ignorant, because he had acquired no certainty about anything. All he had was his deep religious faith as a Catholic, which he retained to his dying day, and a resolute, passionate desire to discover the truth or some truths—if this was possible.

Rejection of textbooks

Less explicitly perhaps than his own actual account, the young Descartes made two major decisions that were to dictate the subsequent pattern of his life, a strange pattern of nomadic restlessness contrasted with periods of almost hermitic solitude. The first decision was "to abandon the study of letters": he rejected a life of erudition and scholarship. Instead he would "gain experience" in the study of "the book of the world": he would travel and observe. The second was "to make studies within myself" and thus attempt to fulfill "his extreme desire to distinguish truth from falsity": these words formulate the vocation of Descartes, the adult philosopher.

In 1616, after two years of study, Descartes took his degree in law at the University of Poitiers, as his elder brother had done before him. But henceforth he no longer followed the path traditional in his family. Without entirely breaking all familial ties, his visits became infrequent and short during the rest of his life. One visit, in 1623, was important, for it was then that he sold the properties in Poitou inherited from his mother, including a small estate at Perron. Descartes, in fact, was known in his family as M. du Perron, but he detested the title. He invested the monies from the sale in a form of bonds. With this moderate wealth he was thus able to live the comfortable life of a gentleman of leisure whose tastes and needs were simple.

Travel and study were the plan of the dissatisfied and restless student. Both of these objectives were to be fulfilled in an unusual and intricate way. It is possible to simplify the complicated story, because his travels between 1618 and 1628 are sparsely documented and little is known of his studies and thoughts, except for a few evocative incidents. The period of intensive study, or rather of writing and publishing, would then be allotted

to the period 1628–1649, when Descartes lived in Holland, although even there he was to change his abode 18 times. In making such a simplification, it should be noted that during his travels he was also studying "the book of the world" and even making notes and beginning drafts of his future works. During his stay in Holland, but to a much lesser extent, he also travelled, returning to France and paying visits to his friends.

Military service. In 1618 Descartes went to Holland and joined, as an unpaid officer, the army of Maurice of Nassau, prince of Orange, who at that moment was deploying his troops against the Spanish forces, which were trying to recover Spain's richest European province. It was a quite conventional move on Descartes's part, since military service and the practical study of warfare were complementary to a gentleman's education. That Descartes, a Catholic, should volunteer to serve in the army of a Protestant prince was in no way exceptional: there were many young Frenchmen in William's army. Eighteen months later, by contrast, Descartes was, in the early stages of the Thirty Years' War, in the army of the Catholic Duke of Bavaria, Maximilian, who was taking the field against Frederick V, the Elector Palatine and Protestant king of Bohemia. Frederick was to lose his throne in 1620, after the decisive battle of the White Mountain, near Prague. His daughter, Princess Elizabeth, became in 1643 one of the philosopher's closest friends and correspondents. It is improbable that Descartes ever participated in any real fighting. The camp life bored him: there was, in his own words, "too much idleness and dissipation." He continued to observe and to make notes, and above all, his fascination for the mathematical sciences was given impetus by a close friendship with the Dutch philosopher, doctor, and physicist Isaac Beeckman, who was then a professor at a nearby town. Beeckman was struck by the mathematical bent and ability of the young French officer who alone, so we are told, was able to solve in a short time an elaborate mathematical puzzle. The friendship was to continue for 20 years, with some upsets. Beeckman, at least, brought his young friend up-to-date with many recent developments in mathematics, including the work of the French mathematician Franciscus Vieta, who by using letters as symbols for constant quantities and for unknowns in an equation was one of the pioneers of modern algebra. An important part of the fame of Descartes was that in a treatise he applied algebraic formulation to geometrical problems, the basic concept of modern analytic geometry. This treatise, part of his first published work (*The Discourse on Method*, 1637), attempts to deal with abstract general qualities, in their relations and orders, instead of groups of geometric figures and separate theorems: it is essentially a method of simplification and unification.

Mathematics: coordinate geometry

Formulation of a rational scheme of knowledge. The meeting with Beeckman renewed Descartes's enthusiasm to pursue his chosen path. He noted, in March 1619, the possibility, which he hardly yet believed himself, "of a completely new science which would be capable of solving in a general manner all problems which could be proposed in all the quantitative fields, whether continuous or numerical." He had begun to envisage the possibility of a universal method to solve problems of a mathematical nature. Toward the end of the same year this possibility was to become even more far-reaching in its application.

Leaving the army of the Prince of Orange, Descartes travelled to Denmark, Danzig, Poland, and Germany and then joined the Bavarian Army in its winter quarters, near Ulm. According to his own account, he lived in a well-heated room, slept ten hours nightly, "occupying myself with my own thoughts." According to his first biographer, Adrien Baillet, who had Descartes's own detailed account before him, it was there on November 10, 1619, that Descartes "was filled with enthusiasm, he discovered the foundations of the Admirable Science, and at the same time his vocation was revealed to him in a dream." The notes in fact spoke of three separate parts of a dream. The symbolism, interpretation, and import of this dream episode have been the controversial subject of many articles and even books, as its cause has been variously attributed to overwork, indigestion, fever, or a mystical crisis. Whatever the meaning or cause, Descartes definitely had the idea, or even vision, of some kind of unitary universal science that would link all possible human knowledge together into an all-embracing wisdom. The mystical element, completely out of context with the almost ruthless reasonableness of Descartes's character, had certainly the psychological effect of confirming his intent to pursue his search for truth and of convincing him of his ultimate success. This conviction may even account for the arrogance with which he defended impatiently his published works against his critics.

Vision of a unitary universal science

The ardent belief about his real vocation in life, the conviction that he had discovered the key to final success in his search for certainty and true knowledge, did not outwardly alter the vagabond pattern of his life in the next seven years. In 1621, he was—for the last time—experimenting with the military life as an officer in the Imperial Army in Hungary. More travels in Germany and Holland followed and then, in 1622, a stay in France, in Brittany, and Paris. From the autumn of 1623 to the spring of 1625, he wandered around Italy, returning thence to France, where he lived mainly in Paris. There he made new friends among the savants and renewed old acquaintances, especially with Father Marin Mersenne, a great polymath, who was to act as his trusted correspondent and adviser in future years. Mersenne was in touch with all of the famous intellectuals in Europe and was thus in a unique position to introduce the work of Descartes to them and report back their comments and criticisms. Discussions with friends, private study, and reflection were the pattern of Descartes's life in Paris. One noteworthy incident is reported. At the house of the Papal Nuncio, after someone expounded "a new philosophy," Descartes in a succinct argumentation, based on reasoning akin to the mathematical methods of proof, confounded and refuted his opponent. All present were deeply impressed: the name of the young philosopher began to be repeated. Cardinal de Bérulle, the leader of the Catholic reaction against Calvinism, who was present, sent for him and insisted that Descartes had the duty to utilize his talents to the full and complete the design that he had outlined to his audience.

Philosophical works. This advice corresponded exactly with Descartes's innermost feelings. But to carry out this work, he needed peace and quiet, and Paris was too distracting. In the autumn of 1628 he went for a few months to the north of France and then to Holland, where he was to live, except for short absences, until 1649. The travels as such were over: it was time to put in writing the results of his experience of "the book of the world" and of his own meditations. If in Holland he was to change his place of residence almost yearly, except for the last five years at Egmond-Binnen, this apparent restlessness enabled him to enjoy periods of solitary work and yet, but only at his own choice, to keep contact with congenial friends by visits and correspondence. Descartes was not, despite appearances, antisocial but merely, for the sake of his comfort and work, opposed to an indiscriminate social life. As his fame in Europe grew, this frequent moving around was to prove a prudent precaution. From 1628 onward, Descartes composed the works that made his name famous in his own lifetime, but he also kept in touch with old and new friends in a voluminous correspondence, of which a great deal is extant. All these show abundantly that the years of travel had also been years of intensive study and reflection and that Descartes had accumulated a wealth of ideas, which he was ready to put in writing.

Posthumous publication of *Regulae*

In 1701 a posthumous collection of Descartes's works was published, which contained a treatise on methodology called "Rules for the Direction of the Mind" (abbreviated generally as *Regulae*). This treatise, incomplete and roughly drafted with repetitions and inconsistencies, was composed by Descartes during the winter months of 1629 and the following year. Possibly never intended

for publication, it may have been used by Descartes as a kind of intellectual balance sheet for future reference. The new method itself was expounded in simple terms, with less mathematical emphasis, in Descartes's first published work on philosophy, the *Discourse on Method*, in 1637. But the fundamental novelty of his approach was already sketched in the first rules of the *Regulae*. There he asserts that all knowledge is of one kind only, since its acquisition is entirely dependent on the use of the human mind. He rejects the Scholastic view that there is a distinction between various kinds of knowledge based on the diversity of knowable objects. To him the mind, the "power of knowing," is always the same, whatever the objects to which it is applied. Well applied, it can attain truth and certainty: misapplied it will fall into error or doubt. All sane men have this natural ability of discerning the true from the false, a "natural light of reason." Only by discovering the nature and limits of this power can one determine the correct way of using this ability. This implies in the first instance the elimination of any factors that may constitute a hindrance, such as preconceived opinions of any kind, and secondly the strict practice of an orderly method, such as is found, for instance, in the mathematical sciences. Thus one must start from self-evident data, which is known to be clearly and distinctly true, and make doubly sure that every step in the deductive progress from the data is itself self-evident. The syllogistic reasoning, on which Scholastic philosophy was based, must be rejected as useless for the discovery of truth. The unity of all the sciences and the analytic method of reasoning were the germinal ideas that had been fermenting in Descartes's mind during his years of travel: they are the basis of the Cartesian revolution in science and philosophy.

The *Regulae* did not suppose naively that anybody and everybody can discover new truths merely by following a set of rules. Apart from inborn mental ability, which is obviously unequal among persons, a long apprenticeship of self-training was indispensable to acquire what Descartes called sagacity and perspicacity. This stage, Descartes considered, with justification, he had reached.

Research studies for *Le Monde*

For four years, from 1630 onward, he spent his time mainly in the study of different sciences, which, unified by his new method, would lead to a universal scheme of knowledge. This study, which nowadays would be regarded as research into the mathematical and physical sciences, was carried out, occasionally with a friend, often in letters to other savants, especially Father Mersenne, but mainly alone in different houses in Holland. The research covered many fields: optics, the nature of light, the laws of refraction; meteorology, the nature and structure of material bodies, air, water, earth; mathematics, especially geometry; and the physiological and anatomical sciences, in which he engaged in dissection of different organs, obtaining his specimens from local slaughterhouses, inventing the term of embryogeny for what is now called embryology.

It was Descartes's ambition to publish a comprehensive work, which he had entitled *Le Monde* (*The World*). By 1633, he had almost completed his draft. Then, in a letter from Mersenne, he learned that the astronomer Galileo had been condemned in Rome by the Catholic Church for his advocacy of the Copernican system. From Beeckman he borrowed Galileo's book, in which he recognized many of his own conclusions, particularly his endorsement of the Copernican theory of the earth's movement around the sun. Although he risked no physical danger in Holland, he was sufficiently prudent not to publish his work: he did not even send the manuscript to Mersenne.

This decision left him unshaken in his own self-certainty about the truth of his conclusions, and later criticism was never to affect his convictions. But Descartes realized that his scientific work was a set of interconnected propositions based on a unitary strand of premises and argumentation and therefore stood or fell as a whole. He therefore undertook, in the next three years, to demonstrate that his new method, and some instances of its application, was essentially based on a firm philosophical foundation, on a metaphysical doctrine that he believed was also acceptable to the Catholic theologians, including his former teachers, the Jesuits.

Publication of the *Discourse on Method*

At Leyden, in 1637, Descartes published anonymously a volume entitled: *Discours de la méthode* (1637; *Discourse on Method*). It was, quite exceptionally, written in French, when Latin was the common language of all learned works. He had salvaged three sections of his ill-fated *World* and composed, in fact very hastily, as an introduction, a semi-autobiographical account of his method and philosophical doctrine: the efficacity of his methodological principles was demonstrated by successful examples of the application of these principles to three scientific problems. These essays as such are of interest mainly to historians of science, who point out that the first essay, on optics, formulated the law of refraction; the second, attempting a scientific explanation of the weather, included an explanation of the rainbow; the third, in Descartes's view, and correctly so, the clearest vindication of his method, introduced the famous Cartesian coordinates (so named by the philosopher G.W. Leibniz) and, using algebraic notation to deal with geometrical problems, laid the foundations of modern analytic geometry.

"Discourse on Method." The introduction, merely a seventh of the original volume in number of pages, is the famous *Discourse on Method*, the most widely read of his works. Simply and elegantly written, the polished brevity of its text tends to conceal the difficulties of Descartes's doctrine, which philosophers, from its publication to the present day, have more often criticized—or even rejected—than praised. Yet, in praise or blame, it is generally agreed that the *Discourse* has radically influenced the trend of modern philosophical thought. It is a philosophical classic: it introduced a new mentality and, of course, new problems.

The opening phrases of the *Discourse* have become a commonplace:

> Good sense is, of all things in the world, the most equally shared: for everyone thinks himself so well provided with it that even those who are most difficult to satisfy in every other way, do not usually desire more of it than they already have.

The courteous irony hides the fundamental assertion that the human mind is basically sound and the only means of attaining truth. The first three sections of the *Discourse* tell how the author was deceived by the knowledge he acquired from education and books. His answer was twofold: to pull down the house before rebuilding it and, as expressed in the first four rules of method, here condensed from the *Regulae*, "never to accept anything as true which I did not clearly and distinctly see to be so." Descartes thus implies the rejection of all accepted ideas and opinions, the determination to doubt until convinced of the contrary by self-evident facts. In the *Discourse*, written in French for the general public, conventional moral principles are provisionally excepted since life must go on during this radical revision. The second rule is an instruction to analyze the problem to be solved. Once cleared of its prejudices, the mind, using the example set by mathematicians, "must divide each of the difficulties under examination into as many parts as possible"; that is, discover what is relevant to the problem and reduce it as far as possible to its simplest data. The third rule is "to conduct my thoughts in order, beginning with objects that are the simplest and easiest to know and so proceed, gradually, to knowledge of the more complex." The fourth rule is a warning to recapitulate the "chains of reasoning" to be certain that there are no omissions. These simple rules are not to be considered a mere automatic formula; they are to be regarded as a mental discipline, based on the example of mathematical practice.

To find the self-evident datum on which to build his science, Descartes, in the fourth section of the *Discourse*, outlined what he called the metaphysical foundations of his doctrine, since the science of the physical world must, in his view, be based on the absolute certainty of the

real world. The method of doubt is universally applied; the arguments of the Sceptics pushed to their limits, the uncertain evidence of the senses rejected, even the existence of my own body doubted as a possible illusion, a dream. But I who doubt, I who am deceived, at least while I doubt, I must exist, and, as doubting is thinking, it is indubitable that while I think, I am. (The Latin formulation is *Cogito, ergo sum*, "I think, therefore I am.") From this cornerstone of self-evident certainty, Descartes derived all philosophical propositions. The idea of God, he argued, implies the real existence of God. No finite or imperfect being like man could have produced the idea of an infinite or perfect being—hence only God could have revealed it to man. Inasmuch as the body can be so easily doubted, whereas the existence of the mind cannot be doubted, it is clear that the two are radically distinct. The body, as such, is subject to mechanistic explanation. Animals, being without rational minds, are strictly to be considered as machines, automatons subject to instinctive reflexes. The universe consists of two different substances: minds, or thinking substance, and matter, the latter being basically quantitative, theoretically explicable in scientific laws and mathematical formulas. Only in man are the two joined in a substantial union.

"I think, therefore I am"

Despite the anonymity of the *Discourse*, the author's name and theories were soon known in the learned circles of Europe. It was the scientific essays that attracted attention and provoked much controversy, among others with the French mathematician Pierre de Fermat. But there were also enthusiastic disciples who spread the Cartesian doctrine in the Dutch universities, provoking again bitter opposition. As far as was possible, except for a few visits and letters, from 1638 to 1640, Descartes was busy writing in the small town of Santpoort. He was not entirely alone. With him was a Dutch servant girl, Helen, who had borne him a daughter in 1635. The child, to his bitter grief, died in 1640.

"Meditations." The work he was writing was in Latin, a more elaborate development of the fourth part of the *Discourse*, his metaphysical doctrine. The final manuscript was sent to Mersenne, who was entrusted with the task of getting the formal approval of the Sorbonne and also "the opinions of the learned." Mersenne collected these critical opinions, including those of the English philosopher Thomas Hobbes and the French mathematician and philosopher Pierre Gassendi, and forwarded them to Descartes, who drafted replies, somewhat irritably and reluctantly. Ultimately, in 1640, the text, with the objections and replies, was published in Paris. The work is known as the *Meditations*: it was translated into French in 1647. Despite differences of detail and some omissions, the doctrines expounded in the *Discourse* and in the *Meditations* are essentially similar. The most striking difference is that the methodological doubt is pushed even further, to include the hypothesis of an omnipotent and malignant demon who could make everything that one thinks to exist into nothing more than one great cosmic deception. The answer is the same: at least I, who am deceived, exist. The *Meditations*, constructed as a single argument, each link dependent on the previous in an elaborate "order of reasons," is the major exposition of Descartes's philosophical doctrine.

Impact of *Meditations*

If the publication of the *Meditations* brought Descartes renown as one of the most famous philosophers, it also involved him, directly or indirectly, in bitter controversies carrying theological overtones. In Holland, the president of the University of Utrecht accused him of atheism, and Descartes was, in fact, condemned by the local authorities in 1642 and again in 1643. The intervention of the French ambassador prevented worse consequences, and in 1645 a decree of the University of Utrecht created an armistice by forbidding the publication of any works for or against the Cartesian doctrine. At Leiden, in 1647, another attack, including an accusation of Pelagianism—the belief that the will is equally free to choose to do good and to do evil—produced a similar decree of neutral censorship. In France, the Jesuits, with a few exceptions among the younger fathers, had given a cool reception to the work of their former pupil. In the midst of these disputes, Descartes began work on a more formalized account of his whole thought, philosophical and scientific, which he thought might receive support in Catholic circles, especially among the Jesuits. But his hope was vain: eventually the Jesuits officially rejected Cartesianism, and his works were placed on the *Index*, the Catholic Church's list of forbidden books.

Other works. *Principles of Philosophy* appeared in Amsterdam in 1644, while Descartes was on a short visit to France: it was translated from Latin into French in 1647. His philosophical doctrines are formally repeated in the first part. The other three parts are a comprehensive attempt to give a logical account of all natural phenomena in one single system of mechanical principles, through the whole field of physics, chemistry, and physiology. Historically, the importance of these three parts lies in the total rejection of all "spiritual" or qualitative notions in scientific explanation, as well as the refusal to include teleological or purposive causes. On the positive side, the expressed determination to explain all physical phenomena in mechanical terms and to relate these terms to geometrical ideas and the use of hypotheses to aid generalizations led the way to the modern approach to scientific theory. On the other hand, Descartes's own views were often wrong or quickly made obsolete.

Principles of Philosophy

Descartes visited Paris again in 1647 and 1648, always for a few months at a time. He met the French mathematician and theologian Blaise Pascal on the first visit and suggested to him the famous experiment proving that air exerts pressure on all objects. The later visit enabled him to meet some of his famous contemporaries, and critics, Gassendi and Hobbes and, of course, Mersenne, who was to die soon. The political situation was unsettled, and Descartes detested the noise and bustle of the city as well as its social life: he longed for "the innocence of the desert." He seemed to feel an urgency to continue his studies: he had already begun to rewrite another part of the *World*, a treatise on physiology, *On Man* (*de Homine*). He returned to Holland at the end of the summer.

The *Principles* were dedicated to Princess Elizabeth of Bohemia. They had met in 1643, and an affectionate friendship had developed between Descartes and the intelligent young woman. Visits were rare, but there was a voluminous correspondence, with a documentary value scarcely less than that of the correspondence with Mersenne and the French theologian Antoine Arnauld. It ranged over a varied field, from geometry to political science, from medicine to metaphysics, but concerned itself especially with the problems of the interaction of body and soul. In dealing with this last problem, both in a practical and theoretical manner, Descartes unsystematically sketched the outlines of his views on ethics. He summarized these views in a small book, printed in 1649, the *Treatise on the Passions*. Descartes described his approach to the subject not as that of a moral philosopher but of a physician. The mind, in many of its activities, is dependent on the body: a passion (*i.e.*, that which is felt) is often an action in the body. Physiologically, Descartes placed the centre of interaction in the pineal gland, a small body centrally located at the base of the brain. The body itself, the most perfect of machines, works by what is now called conditioned reflexes, but the effects of these automatic instincts and desires can be controlled or modified by the mind, by rational willpower. Bodily hygiene is important, but there is equally a need of a mental hygiene, which is based on true knowledge of the psychophysiological factors that condition behaviour. But it is also based on the training of the "good sense" and the acquiring of "wisdom," which depends upon the knowledge of the truths of metaphysics, which in turn include knowledge of God; that is to say, the supreme Good. Descartes thus concludes that moral activity is based on a true knowledge of the relative value of things.

Treatise on the Passions

A manuscript copy of the *Passions* had gone to Queen Christina of Sweden, who since 1647, through the French

ambassador, had obtained the works of Descartes and begun to write to him. An ambitious patron of the arts and collector of learned men for her court, she was anxious to meet "the celebrated M. Descartes." Despite pressing invitations, even the sending of a naval vessel, Descartes was extremely loath to leave Egmond and offered excuses of all kinds, suggesting that it was sufficient to read his books. Finally he accepted and, as he wrote, "born in the gardens of Touraine," he went to the "land of bears between rock and ice," arriving in Stockholm in October 1649. He was welcomed with great ceremony and was impressed by the eagerness and energy of the 23-year-old queen; he was less impressed by her devotion to classical studies and her ignorance in philosophy. Excused from most court ceremonial, except from writing French verses for a ballet, his chief obligation was to instruct the queen in philosophy. Tutorial time was at five in the morning. In the rigorous climate, where, in Descartes's words, "men's thoughts freeze during the winter months," his health deteriorated. On February 1, 1650, he caught a chill that developed into pneumonia. Ten days later, after receiving the last sacraments, he died in the confession "of the faith of his nurse."

Death in Sweden

In Protestant Sweden he was, as a Catholic, buried in the cemetery reserved for unbaptized children. In 1667, his remains were conveyed to Paris and buried in the church of St. Geneviève-du-Mont. Disinterred during the French Revolution for burial among the illustrious French thinkers in the Panthéon, his tomb is now in the church of St. Germain-des-Près.

MAJOR WORKS

Compendium musicae, 1618; *Regulae ad Directionem Ingenii*, written 1628, published posthumously 1701; *Traité du monde ou de la lumière*, written 1633; *Discours de la méthode . . .*, 1637; *Meditationes de prima philosophia, . . .* 1st ed. 1641; *Meditationes de prima philosophia, . . .* 2nd ed. 1642; *Principia philosophiae*, 1644; *Les passions de l'âme*, 1649.

For translations into English of the above works and others, see below *Bibliography*.

BIBLIOGRAPHY. GREGOR SEBBA, *Bibliographia Cartesiana: A Critical Guide to the Descartes Literature, 1800–1960* (1964), is a readable bibliography covering biographical and doctrinal books and articles. On Descartes's life, the basic sources are his own works and letters. The classical edition is *Oeuvres de Descartes* (1897–1913) in 12 volumes, the last of which is a biography by CHARLES ADAM, still the most comprehensive modern work. The letters were republished, to include many from Constantyn Huygens, in *Descartes: Correspondence*, 8 vol. by CHARLES ADAM and GÉRARD MILHAUD (1936–63). The other major source is ADRIEN BAILLET, *La Vie de Monsieur Descartes*, 2 vol. (1691). Despite apologetic bias, Baillet is still a unique source of contemporary information. An abridged version of 1692 (reprinted in 1950) is more readable.

In the English language, E.S. HALDANE, *Descartes: His Life and Times* (1905) and J.R. VROOMAN, *René Descartes: A Biography* (1970) are the only full biographies. As to the works, many English translations of the *Discourse* exist: E.S. HALDANE and G.R.T. ROSS, *The Philosophical Works of Descartes* (1931–34), still in print, includes the *Regulae*, the *Discourse*, the *Meditations*, with the *Objections* and *Replies* and the *Principles*. Another edition, published by Great Books of the Western World (1952), excludes the *Principles*, but includes the *Geometry*. Another translation by N. KEMP SMITH (1952) includes the three works, *Regulae*, *Discourse*, and *Meditations*, as well as sections of the *Dioptric*, and the *Passions*. There are also English translations in R.M. EATON, *Descartes: Selections* (1927); LOWELL BAIR, *Essential Works of Descartes* (1961); and A. KENNY (ed.), *Descartes: Philosophical Letters* (1970). More modern in style is the translation by E. ANSCOMBE and P.T. GEACH, *Descartes: Philosophical Writings* (1954). Most readers would find the Haldane-Ross edition (now in paperback) quite adequate.

Recent particular studies of Descartes include: L.J. BECK, *The Metaphysics of Descartes: A Study of the Meditations* (1965); F. BROADIE, *An Approach to Descartes' Meditations* (1970; and H.G. FRANKFURT, *Demons, Dreamers, and Madmen* (1970), all of which cover the philosophical doctrine. Collected essays on Descartes are contained in A. SESONSKE and B.N. FLEMING (eds.), *Meta-meditations* (1965); WILLIS DONEY (ed.), *Descartes* (1967); and *Cartesian Essays*, ed. by MAGNUS and WILBUR (1969). The best introduction to Descartes is to read the *Discourse on Method*.

(L.J.Be.)

Deserts

Deserts are arid areas of sparse to absent vegetation and very low population density that comprise more than one-third of the Earth's land surface, if semi-arid regions are included. Approximately 5 percent of the Earth's land area can be categorized as extremely arid; the regions involved are the central Sahara and the Namib Desert areas of Africa, the coastal areas of Ethiopia and Yemen (Ṣan'ā') near the southern end of the Red Sea, the Rub' al-Khali in Saudi Arabia, the Takla Makan in Central Asia, the Atacama Desert of Peru and Chile, and parts of the southwestern United States and northern Mexico. Despite the fact that the desert environment exists on parts of each of the continents, the nature and diversity of deserts are not widely understood.

The term desert is commonly associated with extremes of heat, the near absence of life, and the adventures of a few of the better known explorers and travellers of the past, such as H. St. John Philby, who traversed the Arabian Desert in a remarkable effort in the 1930s. These generalizations are not too wide of the mark, but the popular view of deserts as predominantly sandy areas that have existed in their present locations throughout geological time is an unfortunate misconception. Sands, for example, may cover about 10 percent of the surface area in the Sahara and, where dunes or sand sheets occur, are certainly prominent features of the Namib, Arabian, and some other deserts. In general, however, desert sands must be regarded as a minor portion of the deserts as a whole and are essentially absent over vast reaches of arid terrain. And with regard to the permanence of deserts, it is fair to say that the land areas subjected to arid conditions have varied widely throughout Earth history. This is a consequence both of continental drift—the shifting about of landmasses on the Earth's surface, thus bringing different regions into conjunction with arid climatic zones—and climatic changes per se. Moreover, it should be made clear at the outset that several decidedly different desert types occupy the arid environmental niche, and certain of these are not necessarily associated with elevated temperatures.

Polar, extra-terrestrial, and edaphic deserts

One desert type, for example, persists in the polar regions of the Earth. These areas are mantled by snow and ice or consist of barren tundra; in either case, extremes of heat clearly are alien. Nevertheless, they are true desert areas because lack of vegetation and low values of atmospheric water vapour are common characteristics. Although definitions of aridity are both numerous and subjective, a value of mean annual precipitation equal to 250 millimetres (10 inches) or less is one of the criteria commonly used to delimit desert areas. Variability of precipitation is another hallmark of arid climates, but it has been shown that variability is closely related to mean annual values, increasing as the total rainfall decreases. On this basis it need only be noted that the annual precipitation in the interior of Antarctica and over most of the Arctic region is about 125 to 200 millimetres (five to eight inches).

On similar grounds, such extra-terrestrial neighbours as the Moon and Mars represent a second desert type. The Moon has essentially lost its atmosphere through time, thus precluding the possibility of occurrence of precipitation or life-forms, and the entire atmosphere of Mars contains the equivalent of only three cubic miles of water. This water content is intimately related to the seasonal fluctuation of the polar caps of frost or ice on Mars. Some have argued that the melting of these caps, approximately once every 10,000 years, may involve the discharge of surface water. This is highly speculative, however, and it is fair to conclude at present that by any reasonable standard the prevailing environments on both the Moon and Mars are extremely arid.

In a third category are regions in various parts of the world that receive more than the requisite 250 millimetres of annual precipitation but that are nonetheless dry, barren, and relatively devoid of vegetation. Such regions are termed edaphic deserts, which signifies that the physical cause of aridity resides in the nature of the surface materials. The latter usually consist of very porous vol-

canic debris; as a consequence, moisture penetration to the subsurface is too rapid to permit plant sustenance regardless of the amount of rain that may fall. Some parts of Iceland, the Canarys and other volcanic islands, and the so-called cinder lakes on the Colorado Plateau of the western United States are examples of edaphic deserts.

Hot, arid regions and their variation with time

Finally, there are the relatively hot arid regions of the world that are more commonly intended when reference to the desert environment is made. These owe their existence largely to meteorological causes, being located along the Earth's two great subtropical belts of minimal precipitation, or otherwise removed from sources of rainfall; but, aside from some commonality of climate, considerable diversity exists and generalizations are poorly advised. Both physiographic and geological variations abound, and even the ages of these deserts are non-uniform; some have existed in their present state for millions of years, whereas others have changed drastically since the end of Pleistocene time—*i.e.*, during the last 10,000 years approximately.

Parts of the Sahara and Australian deserts, for example, were glaciated during the Paleozoic Era (570,000,000 to 225,000,000 years ago), and the ancient rocks that underlie all desert areas reflect many kinds of former nonarid environments or conditions of formation. These range from marine limestones and sandstones, many of which contain petroleum and natural-gas reserves in the Middle East and elsewhere, to coal deposits that indicate the existence of former swamps and lush vegetation, and crystalline rocks of many varieties that were formed by the cooling of molten silicate material (magma) above or beneath the Earth's surface. As shown in Figure 1 below, there is evidence that great trees once grew in currently arid Algeria.

Maximilien Bruggmann



Figure 1: Silicified tree trunks in the Sahara between Aoulef and In Salah, Algeria.

In the more recent past, during the Quaternary Period (the last 2,500,000 years), some rather profound climatic changes of uncertain synchroneity with the ice ages affected some desert areas. Artifacts such as those shown in Figure 2 indicate that early man hunted large mammals in areas that could not possibly sustain such life today; forests of oak and cedar grew in such highland areas as Tibesti, in the central Sahara; lakes and lake systems were at one time extensive in such regions as the Kalahari, the Iranian desert, and the western United States, where there is evidence of a former body of water 180–210 metres (600–700 feet) deep in Death Valley, for example. Indeed, relict surface features of many kinds, as well as some relict animals (specialized fishes and some crocodiles in oases), point toward the existence of former moist conditions in many of the world's desert areas in the relatively recent past. The same may be said of habitation in Roman times in North Africa. The theatre at Leptis Magna, Libya, was designed to seat many thousands (Figure 3), but the site today is reminiscent of a ghost town.

The arid lands were clearly of great importance to man during the beginnings of his historic record. The history

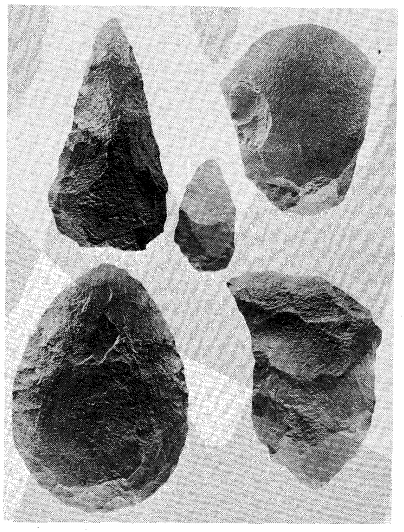


Figure 2: Acheulean artifacts from Ijâfen, found in the region of the Algerian Sahara.
By courtesy of the Museum National d'Histoire Naturelle Prehistoire, Paris

of land use in the vicinity of Mt. Carmel, Israel, and Mt. Hermon on the Syria–Lebanon border and in the Jordan Valley dates back to 6000 BC, when cereals were cultivated, and the irrigation systems of this region and the use of the annual Nile floods for agricultural purposes in Egypt are well documented. Similarly, civilizations are known to have flourished in the fertile Crescent formed by the Tigris and Euphrates rivers and in the Indus Valley region of Baluchistan more than 5,000 years ago. Early settlement and land use also occurred in some of the Asian desert areas, and in Iran, where *qanats* (galleries of tunnels in alluvial material that are dug to provide irrigation waters) still form part of the water supply system. In North and South America the advent of man came later; absolute age dating by radioactive-carbon techniques has failed to provide evidence of man's presence prior to 12,000 to 15,000 years ago in the Western Hemisphere, although work in progress may ultimately yield a greater age for sites of early man near Mexico City. In any case, one of the most curious facets of the history of man's development in both the Old World and the New is his apparent attraction to harsh arid regions despite the availability of better watered neighbouring terrain.

The importance of the arid regions today resides largely in their great mineral wealth. The oil and gas resources of Algeria, Libya, and Egypt in the Sahara and those of Saudi Arabia, Kuwait, and Iran clearly provide a prominent fraction of the world's total proved reserves. Gold and diamonds in southern Africa, phosphates in Spanish Sahara, nitrates in Chile, and vast iron-ore de-

Rotalfoto

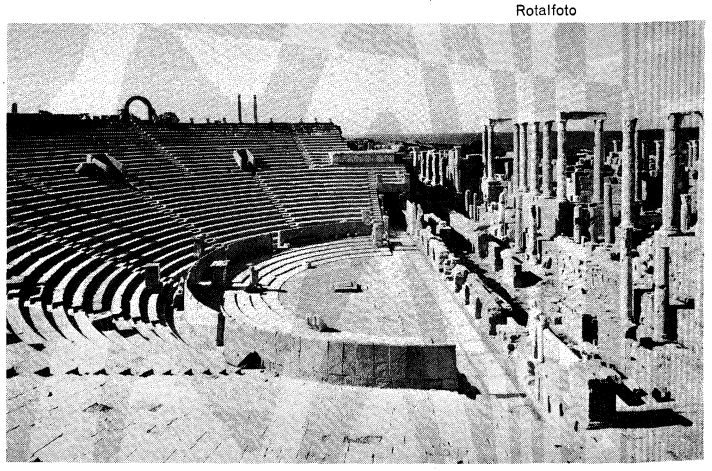
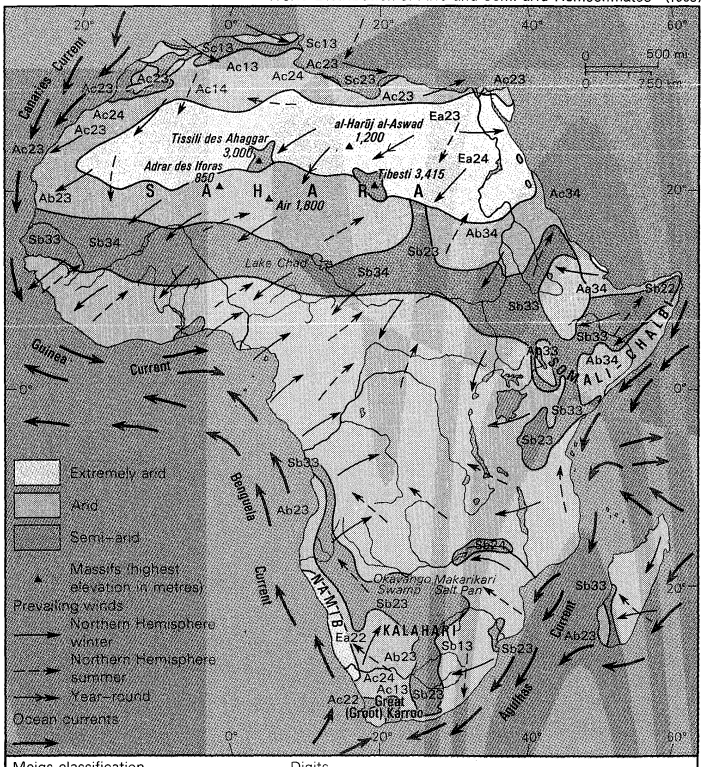


Figure 3: The theatre at Leptis Magna, Libya.

posits in Australia comprise but a few examples of other kinds of mineral resources. Given sufficient water for irrigation and the necessary fertilizers, the arid regions certainly can be rendered fertile, and the promise of augmentation of the world's food supply and relief of population pressures is implicit in the cliché of turning the deserts green. Although desalinization techniques to produce fresh water from seawater may prove a boon to desert coastal areas, the cost of constructing distributive systems to desert interiors is so prohibitive that it has never been seriously discussed. The widely quoted cost of water per 1,000 litres by means of desalinization is always the cost *in situ*—tomatoes can be grown along the Persian Gulf at a price within reason, but they would be precious as pearls if cultivated in the interior of the Arabian Desert. Hence, it is presently unrealistic to conceive of the world's desert areas as loci for vast population centres of the future.

This article treats the general climate, hydrology, and physical features of deserts and includes a section on the permanence of deserts through time. It concludes with coverage of desert flora and fauna. For detailed information on the characteristics of specific desert areas, see articles on those deserts—*e.g.*, THAR DESERT; GOBI (DESERT); NORTH AMERICAN DESERT. Some aspects of desert hydrology and climatology are covered in the more general articles CLIMATE; CLIMATIC CHANGE; and HYDROLOGIC CYCLE. The interested reader should see also appropriate sections of RIVERS AND RIVER SYSTEMS and SOILS, and those articles specifically covering desert features and processes—*i.e.*, ALLUVIAL FANS; PEDIMENTS; SAND SHEETS AND SAND DUNES; PLAYAS, PANS, AND SALINE FLATS; and WIND ACTION. An overview of the development of landforms is provided in LANDFORM EVOLUTION; and mineral formation in arid environments is covered in EVAPORITES.

Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)



Meigs classification		Digits	
E	extremely arid	First digit indicates mean temperature of coldest month	1 = 0° to 10° C 2 = 10° to 20° C 3 = 20° to 30° C 4 = more than 30° C
A	arid		
S	semi-arid		
a	no marked season of precipitation	Second digit indicates mean temperature of warmest month	
b	summer precipitation		
c	winter precipitation		

Figure 4: Deserts of Africa.

THE ARIDITY OF DESERTS

General cause of the arid zones. The general cause of the arid zones of the Earth and, hence, of the present location of the world's deserts resides largely in the dynamics of the atmospheric circulation. In the simplest sense, the engine that drives the planetary atmosphere is incoming solar radiation combined with the effects of the Earth's rotation. Much more incident heat is received in the equatorial region than at the poles, and this tends to lead to a poleward transfer of heat and to a meridional atmospheric circulation pattern. The rotation of the Earth produces latitudinal wind zonation at the Earth's surface—*e.g.*, the trade winds, or easterly winds, in equatorial regions and the westerlies in middle latitudes. Although the physical causes of these two circulatory patterns are complex (see WINDS AND STORMS), the reason that they lead to the existence of arid zones is easily understood.

Relation to atmospheric circulation

Basically, air that is heated at the Equator by incident solar radiation rises, cools, condenses, and releases its moisture content in the tropical zone; it then subsides toward the Earth's surface near latitudes 30° N and 30° S, thus producing two great belts of subtropical high pressure. This descending air is also the source of the easterly trade winds that blow toward the Equator. Some of the world's greatest deserts are located beneath or near these high-pressure belts—the hot, dry trade winds blow across the Sahara (Figure 4), the deserts of the Middle East and South Asia (Figure 5), and part of the North American Desert (Figure 6) in the Northern Hemisphere, and this effect of the general planetary circulation is also responsible for the occurrence of the Atacama–Peruvian desert of South America (Figure 7), the Namib and Kalahari in southern Africa, and the Australian Desert (Figure 8) in the Southern Hemisphere. It should be noted, however, that where the trade winds blow onshore, as along the east coasts of Africa, South America, and Australia, the moisture they bear precludes the existence of deserts.

These generalizations concerning the locations of the world's deserts and wind systems are borne out by several of the maps shown in Figures 4 to 8. Aside from the disruption in continuity of the arid zone along the east coasts of the continents mentioned above, it also may be noted that several of the world's deserts can be characterized as high-latitude types—*i.e.*, they lie north or south of the principal arid belts. Included in this group is much of the North American Desert and the Patagonian Desert of Argentina. These arid areas result principally from physiographic causes: a rain-shadow effect is involved. The latter term is applied when warm, moisture-laden winds must cross a mountain system or similar topographic barrier. When this occurs, as on the west sides of the Andes and the Sierra Nevada, the air masses cool as they rise, and condensation and precipitation occur over the mountains proper. When the air descends on the leeward sides of such barriers, it is thus devoid of moisture and deserts result. The deserts of Central Asia, principally the Takla Makan and Gobi, are also related to physiographic causes in the sense that their locations in the continental interior are quite remote from any sources of moisture. The distribution of land and sea can therefore be cited as a second general cause of aridity on the Earth's surface.

The definition of aridity. To any who have traversed one of the proverbial "burning wastelands" beneath a summer sun, the question "What is a desert?" would appear to be the ultimate inanity. But the definition is elusive and has from time to time been couched in terms of scarcity of inhabitants, lack of cultivated crops, and certain soil and vegetation characteristics. The common thread that binds all desert areas, however, is aridity; deserts are clearly drier than other environments, and thus the question "What is a desert?" really turns upon the question "What is an arid climate, or aridity in general?"

There are several possible approaches to this question, one of which is to attempt climatic classification based upon some set of attributes that is presumably suitable to delineate the arid areas. The word climate is derived from the Greek *klimatos*, meaning "inclination," which reflects the ancients' realization that climate varies with

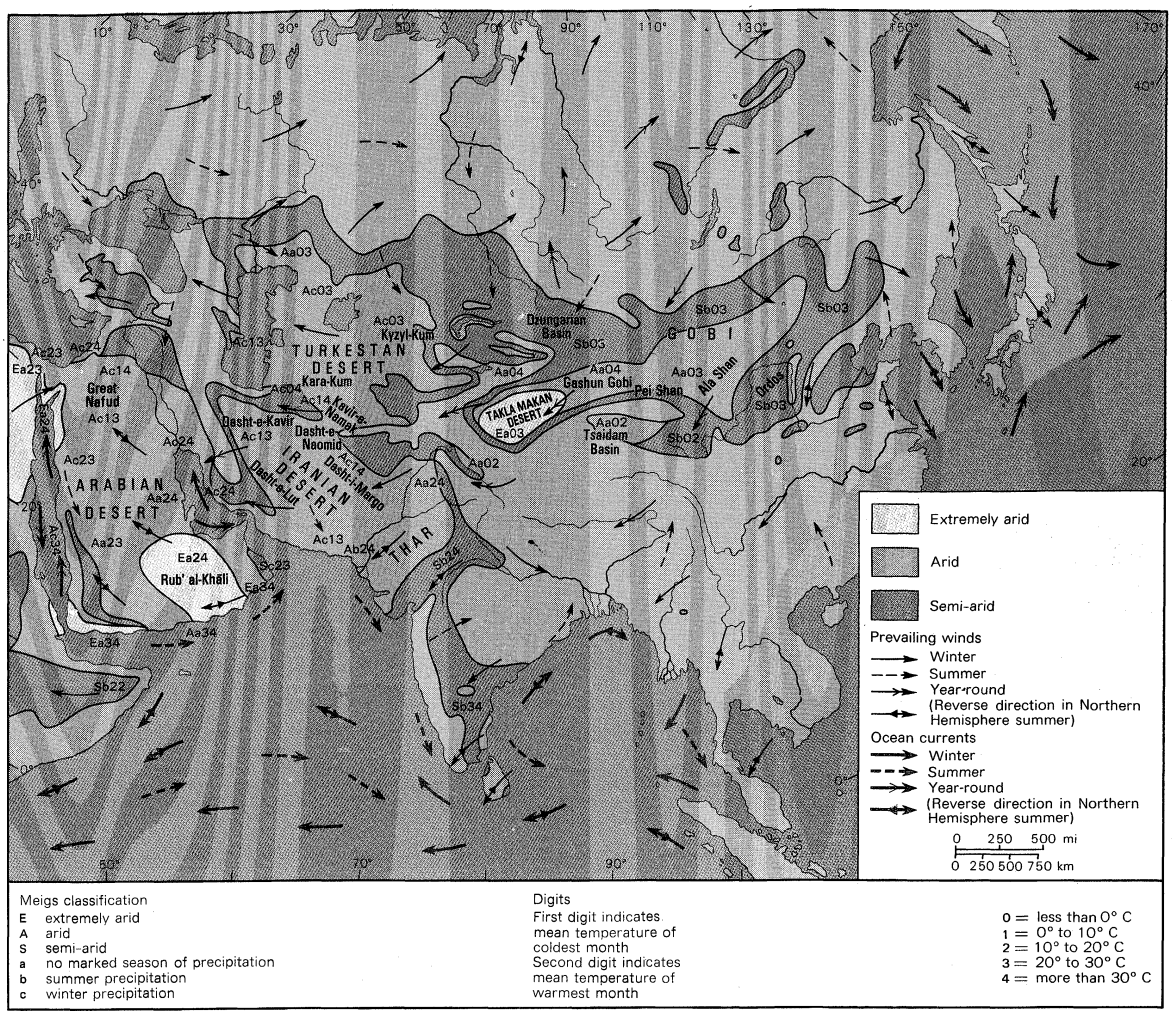


Figure 5: Deserts of Asia.
Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

latitude because of differential inclination of the Earth's surface to the Sun's rays. One of the first climatic classifications, in fact, was provided by Ptolemy, and its basis was geographic latitude and length of day. The most widely used classification of climates today is the Köppen system (see CLIMATE), or some modification thereof, which was based upon the world distribution of vegetation. Vegetation is sensitive, of course, to temperature and precipitation and tends to reflect these variables in a general way.

Indices of aridity

To delineate the boundaries of the arid zones more precisely, however, a large number of indices of aridity have been provided by various scientists, including Köppen. He believed that boundaries could be established in terms of the mean annual temperature (T) and mean annual precipitation (P). The limit between subtropical and semi-arid climatic types was expressed by the relation

$$P = 2\,(T + 7),$$

in which P is in centimetres and T is in degrees Celsius. Köppen's arid index was defined by setting $P = T + 7$, which for a typical temperature of 18° C (34° F) would yield an arid–semi-arid boundary at a mean annual precipitation value of 500 millimetres (20 inches). The index was later modified to reflect the differences that exist between areas with summer and winter rainfall.

A more widely used index of aridity was provided by the French scientist Emmanuel de Martonne, who related monthly precipitation and monthly temperature to the degree of aridity; ultimately he refined his formulation to the expression

$$I = \frac{n\bar{p}}{t + 10}$$

in which the index of aridity, I, equals the mean daily precipitation, $\bar{p}$, for any particular period times the number of days, n, in that period, divided by temperature, t, in degrees Celsius plus 10.

Many other indices of aridity have been set forth over the years, but, like those of Köppen and Martonne, they suffer from several failings. It is questionable, for example, whether mean annual values of precipitation indicate similarities between two regions sufficient to permit their placement in the same climatic category. Clearly, it is not inconceivable that two areas might receive 300 millimetres (12 inches) of precipitation a year, for example, which falls in one area at the rate of 25 millimetres per month but in the second in the form of two 150-millimetre storms. In such an instance the climate, vegetation, soils, drainage characteristics, and other physical factors might well be disparate, despite the equivalence of mean annual precipitation. And with regard to values of temperature, these are really used in lieu of evaporation data, which are more meaningful but more difficult to obtain over widespread desert areas. Moreover, it should be noted that nearly all such indices of aridity involve relationships between millimetres or centimetres of rainfall, on the one hand, and degrees Celsius, on the other. Such unitary chaos always indicates that the underlying physical relationships are not clear.

Definition in terms of water balance

Indeed, because aridity is a measure of dryness, the most superior index would be one based on the water balance of an area. This is, basically, the relationship between the moisture received by an area and the moisture that is lost. The former consists chiefly of precipitation, whereas the losses are produced by all forms of evaporation, surface runoff that transports water out of the area,

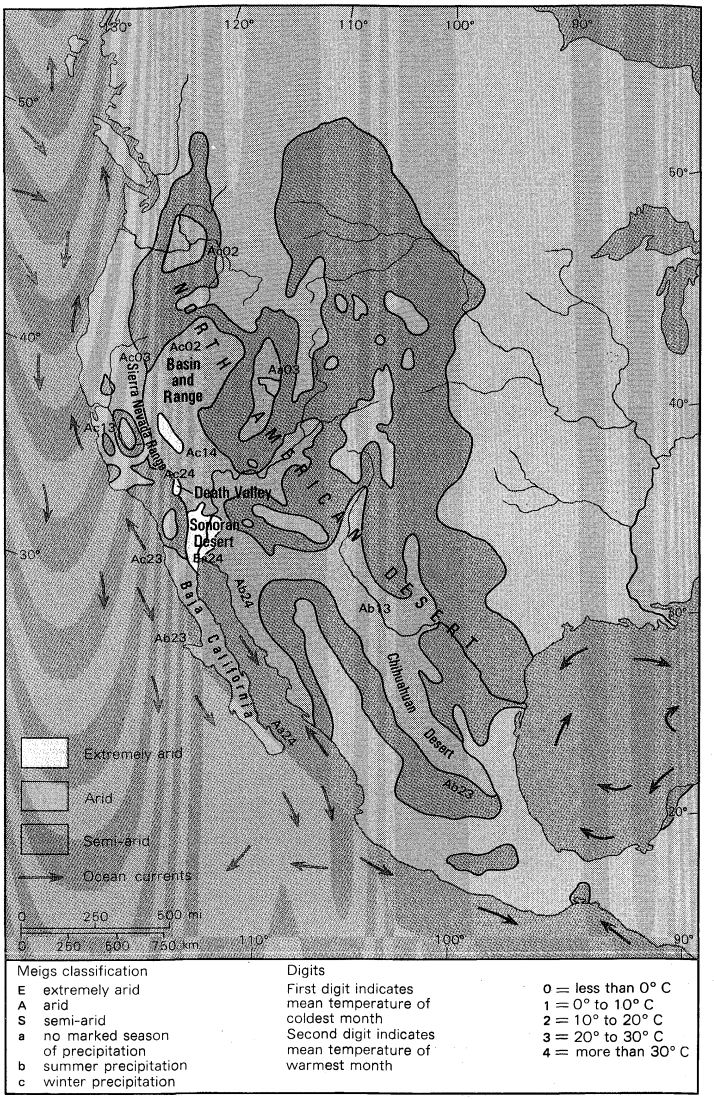


Figure 6: Deserts of North America.
Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

and seepage to the subsurface—*i.e.*, to the groundwater reservoir present. There are several difficulties involved in any water-balance determination (see HYDROLOGIC CYCLE), but this is particularly true in the arid regions because of their vast area and the paucity of necessary data in most instances.

The merits of considering aridity in terms of water input versus water output were so great, however, that the American scientist Charles Warren Thornthwaite devised several indices that are more closely reflective of water-balance considerations. His widely cited precipitation effectiveness ratio, for example, is of this nature. It relates precipitation, evaporation, and temperature in a relatively realistic way. Thornthwaite's aridity index is rather complex; it relates potential evapotranspiration—the total evaporation that would occur from a surface completely covered with vegetation when moisture is continually available—to the length of day and temperature at a particular location.

This aridity index can also be criticized on a number of meteorological grounds; among other things, it has been argued that the index lacks universal applicability and is biassed toward the climatic conditions prevailing at the particular stations (locations) employed by Thornthwaite in its derivation. The index is practical, however, and also is simple to apply because nomograms (graphs that permit immediate determination of the index for any set of climatic values) are available. In fact, the Thornthwaite index of aridity was used in the preparation of the maps shown in Figures 4 to 8. Certain values were selected to define varying degrees of aridity, and further distinctions among desert areas were obtained by considering seasonal distributions of temperature and rainfall. As explained by the legends on these maps, a combination of two letters and two numbers is used to indicate whether an area is classified as semi-arid, arid, or extremely arid, whether precipitation occurs seasonally, and the mean temperatures of the warmest and coldest months. It should be added that the areas classified as extremely arid on these maps not only have a high aridity index but also are areas in which (1) rainfall is not seasonally distributed and (2) periods as long as one year have occurred without recorded rainfall.

Although these small-scale global maps constitute one of the best available sets of maps for purposes of depiction of the arid regions of the world and their subdivision, it must be recognized that the boundaries shown remain somewhat subjective and may err considerably if examined in detail. Considering only the requirement that extremely arid regions are those in which rain has not fallen for periods as long as one year, the difficulties of delineating extremely arid boundaries become apparent. There are, of course, meteorological, or weather, stations in some of these desert areas (such as that at Iquique, in the Atacama) where no rainfall has been recorded for five or more years. But such stations are few and far between, and it cannot be stated with any certainty that rain has not fallen for at least as long as one year

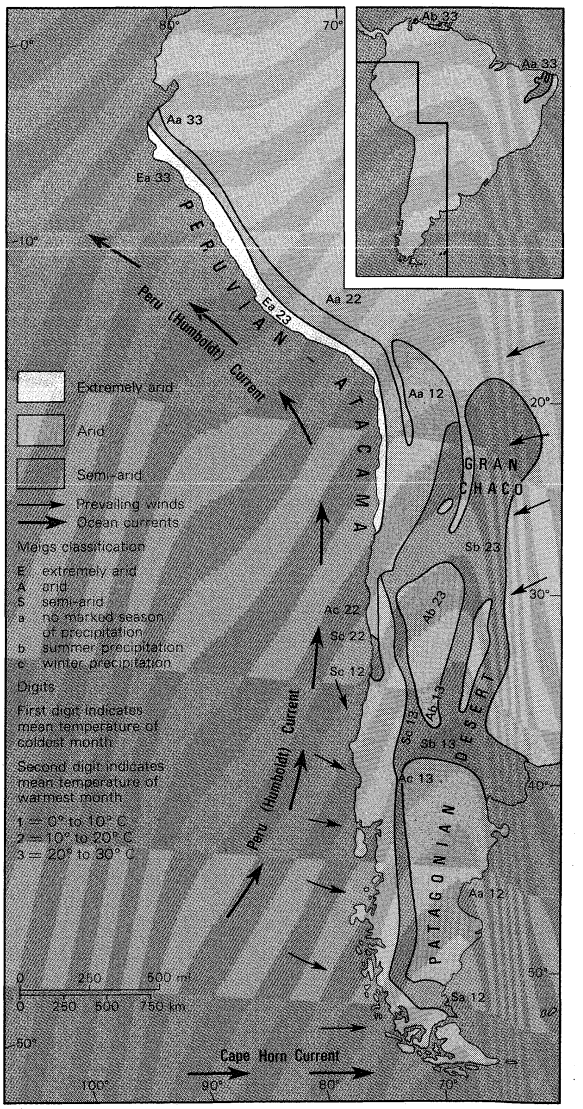


Figure 7: Deserts of South America.
Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

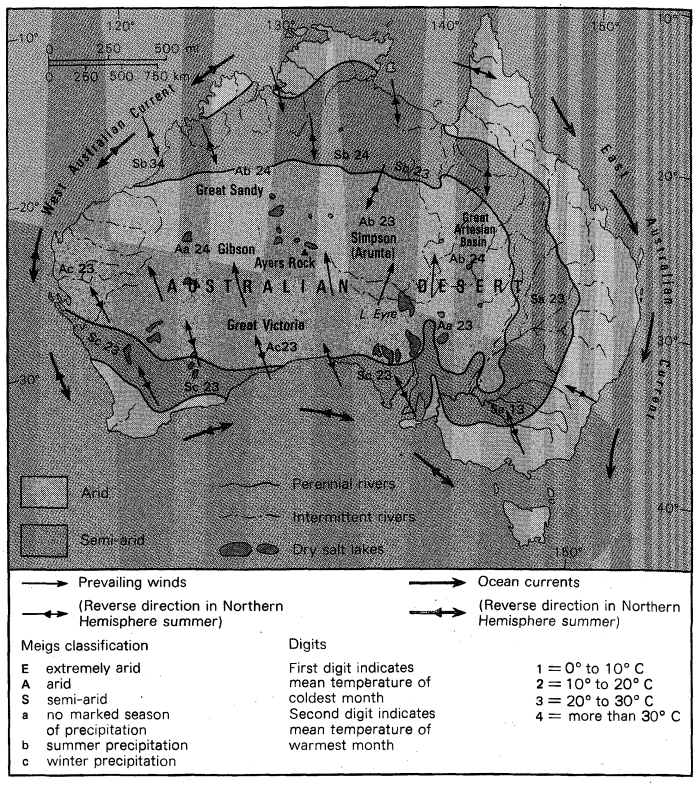


Figure 8: Deserts of Australia.
Adapted from UNESCO, Reviews of Research on Arid Zone Hydrology, "World Distribution of Arid and Semi-arid Homoclimates" (1953)

throughout the entire coastal Atacama, the Rub'al-Khali, the Takla Makan, or any other desert area that is indicated as extremely arid on Figures 4 to 8. It can be said, however, that these areas do generally tend to be hotter and drier than those designated as arid and that rainfall tends to be more spasmodic with respect to both frequency of occurrence and areal distribution. Some climatic data for selected stations in the deserts of the world are given in the Table.

HYDROLOGIC ASPECTS

Principal hydrologic factors. The factors that govern the hydrology of arid regions are climatic, topographic, and geologic. Considering climatic factors first, precipitation and evaporation are perhaps most important because, as mentioned in connection with aridity indices, these two variables tend to control the water balance of an area. Precipitation is highly variable in arid areas, and as a general rule the variability is inversely related to the mean annual precipitation—*i.e.*, as annual precipitation decreases, the variability increases. This variability, in turn, is directly related to the nature of surface runoff in desert regions; it is very spasmodic, and streamflow on a large scale will occur only after intense rainfall.

Another moisture source in deserts is dew, which is worthy of mention because its contribution to the water input and, hence, to the general hydrology of an area has been overemphasized upon occasion. Dew can be derived from the condensation of atmospheric water vapour or from water vapour that is emitted from moisture that is present in the soil. In the latter instance, there clearly can be no net water gain, so the dew derived from soil moisture can be disregarded. The maximum total dewfalls reported do not exceed 0.2 millimetres per day in any case, whereas total daily evaporation of about five to six millimetres is not uncommon. On these grounds it can be argued that the contribution of dew to the overall water balance in arid regions is of negligible import. Its greatest significance is (1) in west-coast deserts such as the Namib in South West Africa, where the influx of humid air from adjacent oceans permits optimal dewfall that benefits vegetation and (2) as a necessary moisture source for the chemical weathering of rocks.

Evaporation in all desert regions is a factor of great significance. It is here designated a climatic factor, be-

Desert Climatic Data for Representative Stations of the World

station	no. of years on record	mean annual precipitation (mm)	maximum recorded precipitation for a 24-hr period (mm)	relative humidity (percent)	mean annual temp. (°C)	maximum recorded temp. (°C)	minimum recorded temp. (°C)	average no. days of Sun per year	average wind velocity (km/hr)	maximum recorded wind velocity (km/hr)
North America										
Phoenix, Arizona (Sonoran Desert)	40	178	108	33	20	47.7	−8.3	209	8	96
Las Vegas, Nevada (Mojave Desert)	22	102	66	18	16	47	−13.6	220	14.4	86*
South America										
Los Canchones, Chile (Atacama Desert)	40	0	...	...	17	36	0.6	264	...	...
Lima, Peru (Peruvian Desert)	10	25.4	...	72	19	34	9	...	...	...
Central America										
Monterrey, Mexico (Chihuahuan Desert)	10	711	...	57	22	42	−9	...	...	...
Africa										
Algiers, Algeria (Sahara)	10	762	...	61	18	42	0	...	...	...
Cairo, Egypt (Sahara)	10	25.4	...	29	20	47	1.1	...	...	...
Tripoli, Libya (Libyan Desert, Sahara)	25	381	...	62	19	45	0.6	...	...	...
Asia										
Baghdad, Iraq (Syrian Desert)	25	152	...	29	...	49	−7.7	...	...	...
Lu-k'o-ch'in (Takla Makan)	2	...	...	...	13	48	−21	...	...	...
Ulaanbaatar, Mongolia	1	208	...	54†	−3	16	−26	...	...	...
Australia										
Alice Springs, Australia (Australian Desert)	90	252	147	...	21	47	−7.2	...	6.6	...
Marble Bar, Australia (Australian Desert)	70	340	305	...	27.7	49.2	1.1	...	...	...
Antarctica										
McMurdo	1	101	...	...	−17	−3	−29	...	...	...

*Six-year record. †Fifty-four year record.

Evaporation in desert regions

cause it is related to incident solar radiation and air temperature and to atmospheric humidity. The total annual evaporation everywhere exceeds total precipitation greatly, and, indeed, this relationship can be considered one of the climatic hallmarks of the arid regions. Wherever evaporation exceeds precipitation, the presence of free-standing natural water bodies such as lakes is generally precluded, unless some permanent source of water is involved. The Nile, for example, can cross the eastern part of the Sahara despite very high evaporation rates (approximately 120 inches per year at ʿAṭbarah, The Sudan) because its source is in the equatorial highlands and because groundwater contributes to its flow over part of its course. This is flowing water, however; an equal volume of standing water whose only input was derived from precipitation would eventually dry up.

Evaporation is greater over water surfaces than over bare soil surfaces, but it occurs over the latter as well and has much to do with depleting soil moisture and inhibiting plant growth. The occurrence of hot dry winds will tend to increase evaporation rates, and for these and other reasons—including difficulty of accurate measurement—existing evaporation data are in some instances debatable. Some representative values are listed in the Table, however.

With regard to topographic and geologic factors that influence desert hydrology, these can best be considered together. Topography may on occasion promote increased aridity by the rain-shadow effect previously mentioned, but it also may promote rainfall where highlands occur within arid regions. The Tibesti massif (Figure 4), within the central Sahara, for example, attains a maximum elevation of 3,415 metres (11,204 feet); this is sufficient to render it a semi-arid "island" in an extremely arid area by reason of the increase of precipitation with increasing elevation.

Geologic factors such as rock type and structure, the character of the soil, and the porosity and permeability of surface materials in general also can affect the hydrology in two ways. If water will readily seep to the subsurface because of the presence of rock fissures and fractures or of highly permeable materials, then the runoff from a given amount of precipitation will be less than normal because of water losses. On the other hand, this same set of conditions may promote recharge (the addition of water) of the groundwater reservoir in the area. Where geological factors inhibit the downward percolation of water, runoff following a storm may be enhanced, but, ultimately, greater evaporation may result.

Streamflow and drainage characteristics

Surface water. Although hydraulic and hydrologic principles are unvarying in all environments, some substantial differences exist between the characteristics of streamflow in arid regions and in more humid areas. The drainage networks that exist in the latter, for example, generally reflect the nature of runoff conditions today, whereas in desert areas there are many drainage systems that reflect establishment during Pleistocene time (2,500,000 to 10,000 years ago), when moister conditions prevailed and streamflow was more abundant.

Streamflow in humid and semihumid areas has been studied in detail, and it has been established that the channels of all rivers are in adjustment to a number of hydraulic parameters—*i.e.*, the width, depth, and velocity of flow and the sediment discharge are related to the bankfull water discharge, and this water discharge governs the nature of the channel. The frequency of occurrence of the bankfull discharge is generally about once every 1.5 to 2.5 years, and overbank floods of great magnitude tend to occur every 50 to 100 years. Rivers in these regions exhibit far more days of low flow than of high flow per year; they usually transport the bulk of their suspended sediment load during high-flow stages, but the many low flows during a year serve to transport much of the total load (the suspended load, dissolved load, and bed load). Finally, flow is perennial, with rivers linking up to ultimately reach the sea.

In desert areas, by way of contrast, streamflow is extremely sporadic. An effective precipitation sufficient to overcome both evaporation and seepage losses into the channel bed and the surrounding dry terrain of the drainage basin involved is required for streamflow to occur. This usually involves a storm yielding relatively intense precipitation (*e.g.*, 36 millimetres [1.4 inches] of rain were recorded in 40 minutes during a storm at Tamanrasset, in the central Sahara). Low flows tend to be few by reason of the seepage and evaporation losses, and, when flow does occur in arid areas, it will likely be of high stage, violent, and highly charged with sediment that is derived from the abundance of loose sand and gravel that blankets most areas. Discharge of any magnitude, however, and particularly bankfull discharge, may have a frequency of occurrence that is as small as once in 100 or more years, and the recurrence interval of major overbank floods may be measured in thousands of years in extremely arid areas. Streamflow is thus characterized as ephemeral rather than perennial, and because drainage generally fails to reach the sea, save for rivers of the first rank such as the Nile or those with relatively short distances to traverse such as the Kuiseb in the Namib Desert, it is termed interior drainage. The Kuiseb is of some interest because it flows to the sea with sufficient frequency to prevent the migration of dunes from the southern Namib Desert despite northern prevailing winds. Its channel thus serves as a boundary between the dunes and the gravel plains to the north (Figure 9).

L.K. Lustig—EB Inc

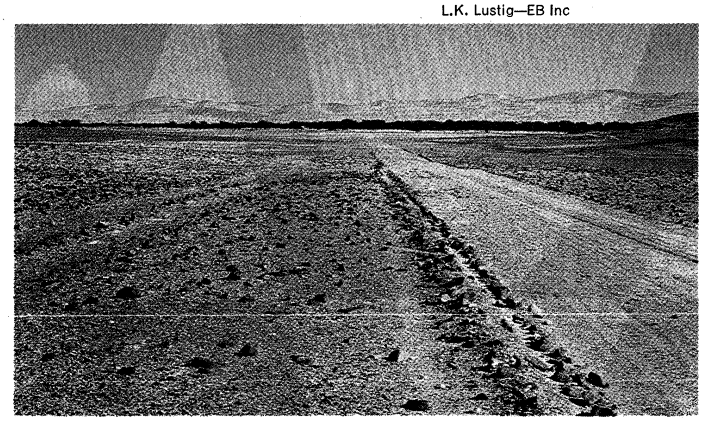


Figure 9: Red sand dunes of the southern Namib Desert separated from the northern gravel plain (foreground) by the Kuiseb River. The channel is marked by a row of vegetation, visible at the base of the dune area.

Desert lakes

Where surface flows are ponded or terminate in low depressions, marshy areas or lakes may result. If there is sufficient precipitation over the site of termination or if there is sustenance by groundwater flow, then a relatively freshwater lake may result. Most desert lakes are without outlet, however, and thus become highly saline, because each increment of inflow transports additional salts in solution and the constant evaporation of water from the lake causes salt concentrations. Salinity values as high as 200 to 300 parts per thousand are not uncommon (see further EVAPORITES).

One of the larger areas of interior drainage in a desert area is the Lake Eyre Basin, in Australia (Figure 8). It embraces a total of 1,300,000 square kilometres (500,000 square miles), and the mean annual rainfall is less than 250 millimetres over most of the headwater region. In 1967, however, an overnight rainfall of more than 150 millimetres occurred, causing ephemeral streams to flow far beyond any previously known distances; water depths as great as five metres were observed, and widths of flow were as great as several kilometres. Some of the streams linked up to provide a surface-water input to Lake Eyre proper, an exceedingly rare occurrence, but the point to be drawn here concerns the origin and significance of the many dry lakes in arid regions. The latter are variously termed playas, pans, and saline flats, all of which are depressional features that sometimes hold standing water but that are more often dry.

Where such features occur in isolation, as in the basin and range areas of the North American or Iranian des-

erts, this discussion is not applicable, but, in several areas of vast desert plains, a veritable legion of these dry lakes dot the landscape in relatively close association. Noteworthy examples are to be found in Western Australia and in western Botswana, in the Kalahari (Figure 10). These dry lakes show evidence of linkage by ephemeral streams, and, rather than relicts of Pleistocene time, they may well be parts of overall drainage systems that, like the central Australian case, flow today infrequently in response to precipitation events of great magnitude and intensity.

Subsurface water. Subsurface water is present in each of the desert areas, but lack of exploratory drilling precludes definitive statements on the total amount of water involved in most instances. Suffice to say that groundwater is easily the principal water resource of the arid regions. Water seeps into the ground and flows downward under the influence of gravity—always assuming, of course, that the materials through which it migrates possess the requisite porosity and permeability. Some of this water migration occurs in the deserts proper; surface-water losses caused by downward percolation into the sands and gravels of stream beds have been previously mentioned. Indeed, there is always some close relation between groundwater and surface water wherever stream channels occur, but much of the groundwater also is related to the vast areas over which rain may fall and ultimately reach the saturated zone in the subsurface.

In the near-surface zone, the water present, principally derived from rainfall, is termed soil moisture. Some of the moisture is lost because capillary action brings it back to the surface, where it may form dew, as previously noted, and is in general subject to evaporation. There is a second zone extending to perhaps 100 metres beneath the zone of soil moisture in many areas. This is a relatively dry zone, but it may contain lenses of water-bearing sediments surrounded by more impermeable beds or layers. Such groundwater is termed perched in reference to its mode of occurrence. Beneath this zone the groundwater proper occurs; the surface of this saturated zone is called the water table (see further GROUNDWATER).

Subsurface structure and the age of groundwater

In many instances the general subsurface structure, or configuration of strata, in desert areas is that of a basin. When the principal water-bearing strata, or aquifers, are upturned and are exposed at the surface in regions of high rainfall, much of the water input to the groundwater system will be so derived—rather than from seepage due to rainfall over the entire basin region. In the Great Artesian Basin, in Australia, for example, the total area is about 1,760,000 square kilometres (680,000 square miles), but the principal water intake is on the eastern edge of the basin, over an area of about 100,000 square kilometres, in which precipitation is 625 millimetres (25 inches) per year. The water flows downward through upturned sandstone aquifers and thence into the centre of the basin, which attains depths of as much as 2,100 metres (6,900 feet). This circumstance of distant intake for groundwater, wherever it occurs, is responsible for the age of the water that has been determined in North America, in the Libyan Desert of the eastern Sahara, and elsewhere. All absolute age measurements on groundwaters in desert areas indicate a range in age of between 20,000 and 35,000 years. This means that waters tapped by wells in Egypt, Libya, and adjacent regions of the Sahara entered the groundwater system from Pleistocene rainfall in highland areas to the south. The groundwater is often called fossil water for this reason.

If the principal water intake is at the upturned edge of a basin structure and if the water-bearing strata crop out at the surface (or occur near the ground surface) farther out in the basin, then there will be a pressure difference caused by the difference in elevation of the water within the aquifer. When this condition exists, the flow of water is termed artesian, and artesian water is the source for most springs, oases, and near-surface water wells in desert regions. The other principal mode of occurrence of these lifesaving caravan stops is also related to groundwater flow. Instead of artesian conditions, the termination of a subsurface aquifer within permeable sediments or the existence of fractures that penetrate the aquifer may be responsible.

DESERT LANDFORMS

Geomorphic processes. Geomorphic processes are those natural physical, chemical, and biological processes that affect landscape in general and the configurations of individual landforms in particular. Considering the arid regions, there are essentially four processes of importance in this regard: weathering, gravitational, fluvial, and eolian.

Weathering refers to the breakdown of solid rocks into smaller fragments or into their component parts by mechanical or chemical means. It is the first essential step in the molding of landforms and their ultimate reduction, because the weathered debris or loose sediment produced is then available for subsequent transportation by wind or water. The decay and reduction of bedrock is an extremely slow process under arid conditions. Many monuments that are thousands of years old remain relatively well preserved today and bear witness to this fact (Figure 11). The abundance of unconsolidated sediment that blankets desert areas, however, is mute testimony to the effects of time; given sufficient time even the slowest processes will be effective.

Desert varnish and duricrusts

Two pronounced effects of weathering processes in desert regions are desert varnish, a dark patina of iron and manganese oxides that coats individual particles and large rock surfaces alike, and duricrusts, indurated (hardened) soil layers that may consist principally of calcium carbonate or iron-, silica-, or alumina-rich compounds. Desert varnish is thought to require at least 2,000 years to form, largely because of its occurrence on datable

L.K. Lustig—EB Inc.

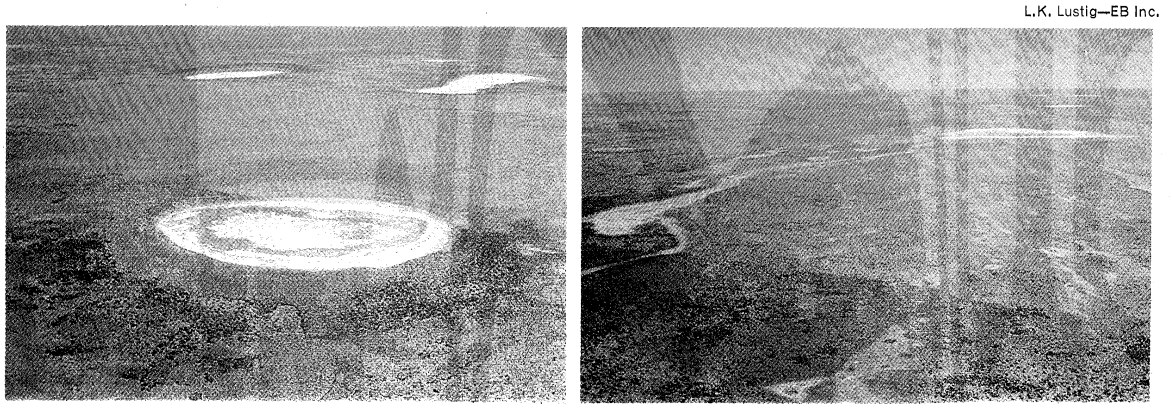


Figure 10: *Pans in the central Kalahari, a desert in Botswana:*
(Left) Circular pan surrounded by ridge of alluvium and windblown sediments appears to be isolated from the two irregular pans in the distance. (Right) An adjacent region after moderate precipitation, showing the tendency for pan linkage and the formation of a drainage system.

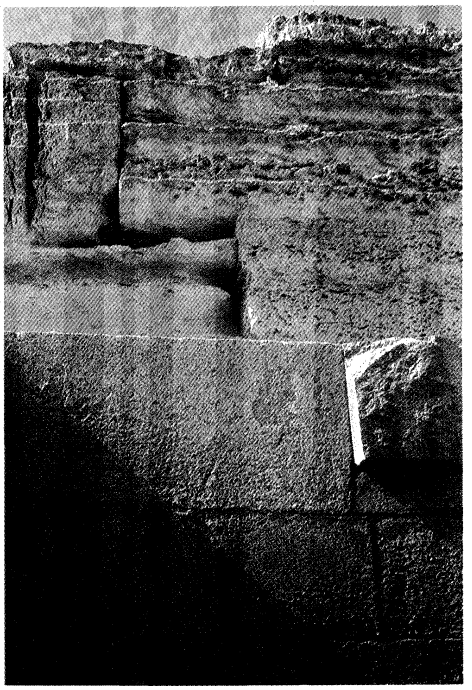


Figure 11: Limestone blocks in the north face of the Great Pyramid of Khufu, in Egypt. Weathering of this stone over several thousands of years has failed to mar the excellence of fit of the blocks; only minor surficial pitting has occurred under the arid conditions involved.
L.K. Lustig—EB Inc.

monuments, artifacts, and similar objects that are known to be of this age or older. There is one reported instance of possible varnish formation within 50 years (in the Mojave Desert), but, in the absence of additional evidence, the 2,000-year or greater range is accepted by most authorities. The iron and manganese involved in varnish formation are thought to be derived from the solid rock beneath the patina, with dew playing an important role as a moisture source in this chemical process.

Duricrusts result from soil-forming processes involving, in arid regions as in other environments, the leaching of chemical constituents from some soil layers and their concentration in others, where they cement and replace existing rock particles. The most common type of crust is perhaps the one rich in calcium carbonate, which is variously termed caliche and hardpan, among other names. No form of crust is rare, however, and in some areas, such as the Sudan, the ubiquitous reddish-brown colour of crusts and surface sediments (Figure 12) suggests the abundance of iron that exists. The crusts in this instance form a protective caprock and tend to preserve the djebels by their resistance to erosion. Duricrusts and varnish both result from chemical processes in arid regions, but the formation of crusts reflects a time requirement that is far in excess of that associated with varnish. The crusts are most abundant in areas that have been stable for the last few millions of years, such as the Australian Desert, where duricrusts were first described (see further DURICRUSTS).

L.K. Lustig—EB Inc.

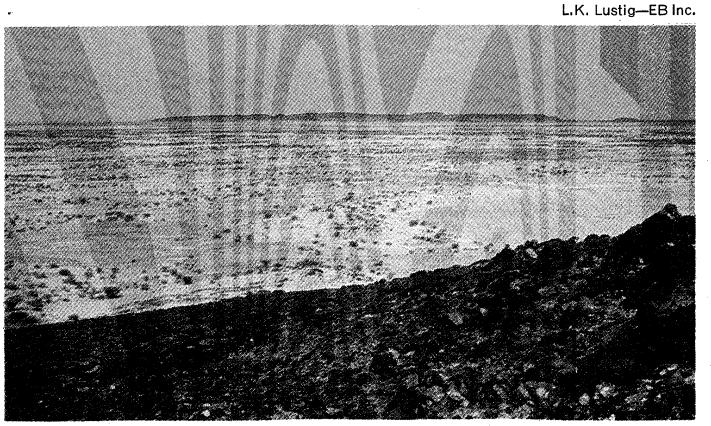


Figure 12: Crest of an isolated djebel near Omdurman, the Sudan, capped by an iron-rich duricrust. Djebels visible in the distance are similarly capped.

Mechanical weathering in deserts is thought to be reflected by the angularity of most of the weathered debris and by the spalling off of great rock sheets concentric about boulders and rock domes that are usually of granitic composition. For many years it was believed that these phenomena resulted from alternate heating and cooling due to large diurnal temperature changes in deserts, but this concept has been demonstrated to be wrong by laboratory experiments. It now seems clear that chemical processes are involved, with dew again serving as the principal moisture source. Although dew in the desert is small in quantity, the abundance of time for repeated increments of hydration and solution in arid areas makes possible these large-scale weathering results.

Gravitational and fluvial processes

Gravitational processes refer principally to the downslope movement of weathered materials under the influence of gravity. This will occur when instability develops on a hillslope or when a mass of material is sufficiently lubricated by water or a water–sediment mixture to overcome frictional resistance. The range of movements possible extends from the fall of a single particle or fragment from a cliff to great rockslides and landslides. The latter do occur in arid regions as elsewhere but are not of great quantitative importance save locally. (For information on the role of gravitational processes in hillslope development and the nature of these processes, see EARTH MOVEMENTS ON SLOPES; and HILLSLOPES.)

L.K. Lustig—EB Inc.

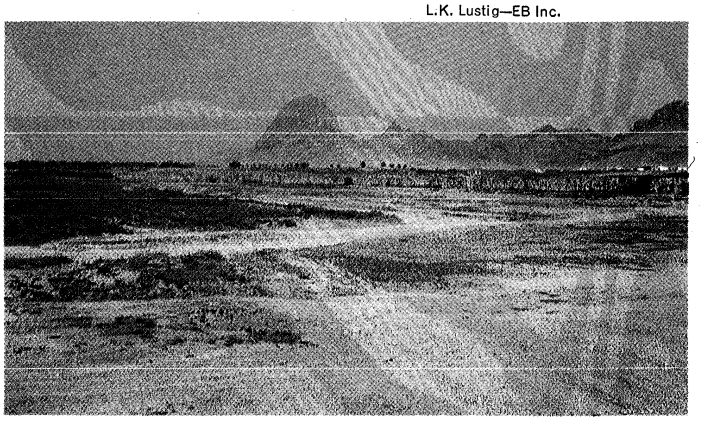


Figure 13: Meander in the Helmand River, Dasht-e Mārgow, southern Afghanistan, showing vertical bluff cut in valley alluvium. Rounded hill (centre) may be a relict feature.

Fluvial processes are those associated with flowing water. As mentioned above in the section *Surface water*, there is no difference between the laws governing fluid flow and channel characteristics in arid regions or other environments. The characteristics of desert streamflow derive from the lack of vegetation and abundance of weathered sediments and from the infrequent distribution of precipitation both in space and time. An intense storm in a given area leads very rapidly to the merger of runoff from slopes in small confined channels or in sheets. This water then pours into the larger arroyos, wadis, or desert washes present, and large quantities of sediment are transported. Some of this sediment is picked up on the initial slopes involved, but much sand and gravel that may have accumulated within the channels since the previous runoff event also is transported. Sediment is also scoured from channel banks, which cave in when wetted. This commonly occurs on the outside banks of meander bends, as in the case of the Helmand River, Afghanistan (Figure 13).

In mountainous regions the concentration of rainfall within a particular drainage basin often leads to the assimilation of so much sediment by surface runoff that mudflows—dense, viscous, water–sediment mixtures—

are produced. These flows can do much erosional work, particularly with respect to their capacity to transport very large blocks and boulders. Without continued rainfall in the drainage area or with a diminution of slope, mudflows will come to rest and dry out. The resulting sedimentary deposit will be virtually indistinguishable from other channel deposits.

Another fluvial process that is characteristic of arid regions is piping—a subsurface sapping of unconsolidated sediments that may effectively produce natural pipes, or tunnels, beneath large areas. Some parts of southern Arizona, for example, are honeycombed by pipes, and continual enlargement of them leads to the formation of natural bridges in these deposits (Figure 14). Piping is intimately related to the formation of desert washes, and in many cases the head of an arroyo receives water from a pipe network during storms, thus causing upstream migration and enlargement of the wash. It is a fluvial process that is underrated because the extensive development of pipes eventually obliterates evidence of their existence.

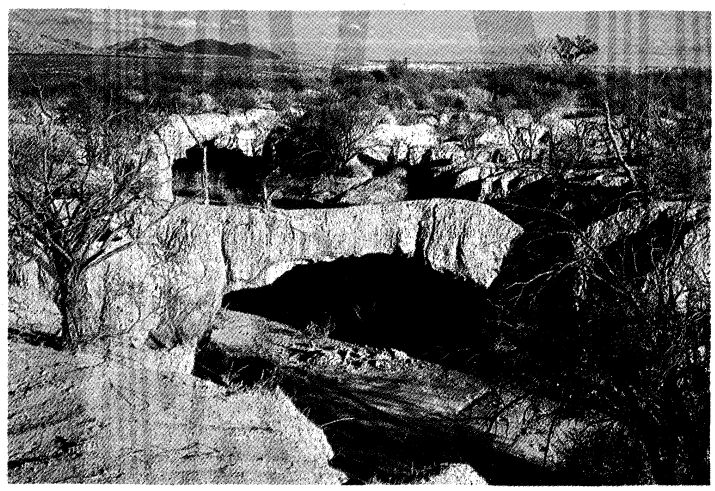


Figure 14: Natural bridges of alluvium in southern Arizona. The bridges result from a process of subsurface erosion and channel development called piping. When the bridges ultimately collapse, the channel beneath them will be indistinguishable from any other desert wash, or arroyo.
L.K. Lustig—EB Inc.

Effects of wind action

Eolian processes result from the work of the wind, and because wind action has always been more visible in deserts than surface runoff, early workers thought that desert features were largely attributable to the wind. Although this view is held to be in error today, much evidence of wind action is indeed present. Prolonged wind erosion can lead to lag-gravel deposits, in which the wind winnows finer particles, thus leaving coarse fragments behind; to blowouts, or deflation basins, which are circular to oval areas of depression thought to have been scoured out by the wind; and to ridges, flutes, and other scouring effects in bedrock surfaces when an abundance of fine particles suspended by the wind produce a sandblasting effect. The latter is also responsible for the production of ventifacts in desert regions; these are gravel particles the sides of which have been worn smooth by sandblasting, leading to a distinctively shaped stone formed by intersecting facets.

Desert winds also produce deposition of sediment whenever their velocities are checked by opposing winds, a mountain mass, or some other cause of wind shadows. The finest material in transport is usually called dust, and windblown dust can travel great distances. Dust storms in the western Sahara, for example, have led to deposition as far away as the eastern United States and Europe. Sedimentary deposits that consist largely of windblown dust are called loess; such deposits are exemplified by bluffs as thick as 100 metres in Mongolia and by widespread blankets of loess in eastern Europe and elsewhere (see further LOESS).

The deposition of sand by the wind is perhaps more widely known. The many desert sand sheets and the variety of existing sand dunes all result from eolian processes. Where dune migration is involved, the cause is wind transport of sands up and over the windward face of a dune; deposition of these same sand grains occurs on the lee face, and in this way an entire dune or group of dunes will migrate downwind. Wind directions are frequently variable rather than unidirectional in the course of a year, and dunes may well migrate in different directions in response to such shifts in stress. Nevertheless, there must always be some net wind component over a long period of time, and this will be the determining factor in considerations of net dune movement or migration. In many instances the net component is zero, in which case dunes may migrate to and fro but will tend to remain in a fixed average location through time.

Finally, the phenomenon of frosted sand grains is worthy of mention here. Because the sandblasting effect of windblown sand on rocks and other objects has long been known, the presence of quartz grains with milky surface coatings was assumed to be caused by the striking of one grain upon another during wind transport. Surface frosting can indeed develop in this way, and the mere presence of frosted grains was accepted by many as absolute proof of arid conditions. Philip Kuenen, the Dutch geologist, made a remarkably simple but astute observation of such frosted grains, however. Microscopic examination revealed that many grains possessed pits of small dimension that resembled miniature tunnels extending from the grain surfaces toward their centres. These pits were obviously far too small for sand grains to gain entrance, but frosting was nonetheless present. The only possible conclusion to be drawn is that frosting of sand grains can be produced by means other than grain to grain impact and, therefore, that the presence of frosting cannot be cited as unequivocable evidence of an arid environment.

Surficial features. A considerable number of landforms occur in desert areas, and though some are considered characteristic of arid regions, nearly all have counterparts in the more humid environments. Mountain ranges and plateaus, volcanic peaks and isolated bedrock knobs, peripheral landforms that are termed alluvial fans if composed of sediment and termed pediments if consisting of a sloping rock surface, sand sheets and sand dunes, playas and other dry lakes, and desert plains and depressions—these are the principal elements of the desert landscape, which is often described as stark, harsh, and angular in profile. This is to some extent more closely related to the relative absence of vegetation and soil cover in arid regions than to some fundamental physiographic distinction that is climatically based (see further LANDFORM EVOLUTION).

Mountains are present in some deserts, such as the Atlas and Anti-Atlas ranges of the northern Sahara or the many mountain ranges of the North American and Iranian deserts, but they are basically absent in others—*e.g.*, the Australian Desert. Wherever they do occur, these highlands tend to be deeply embayed by canyons and are blocky or steeply sloping in general aspect. The specific appearance of any particular mountain range or isolated mountain mass, however, is always governed by the rock type and structure involved and the response of these two elements to the fluvial, eolian, and weathering processes that are operative upon the landform. Hence, it is possible to encounter such disparate forms as the rounded rock mass that occurs in the Dasht-e Mārgow, Afghanistan (Figure 13), the mesa-like features of the Sudanese djebels that result from resistant iron-rich cappings (Figure 12), or the granitic inselberg (derived from German: "island mountain") in the Namib Desert (Figure 15). The last named is of some interest because it indicates the effects of fluvial erosion, albeit accompanied by some cavernous weathering, even in so arid an area as the central Namib, where the mean annual rainfall is approximately 50 millimetres (two inches) today.

Inselbergs, fans, and pediments

An inselberg is one kind of isolated highland landform that rises abruptly from desert plains; pinnacles of several types also occur. Erosional remnants near western Tibesti, in the Sahara, exist as spectacular rock towers that seem to thrust upward from the desert floor. These

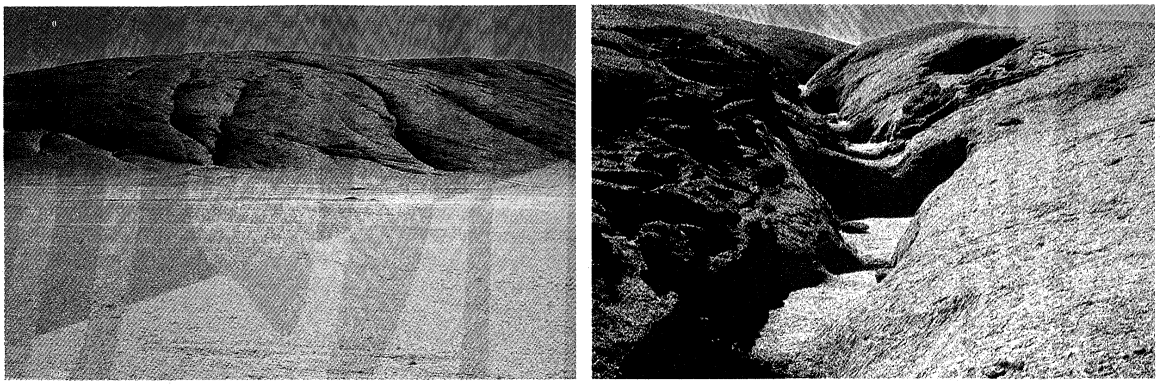


Figure 15: *Fluvial erosion in the central Nam ib Desert, South West Africa.*
(Left) Granitic inselberg rising abruptly from gravel plain. (Right) Close view of one of the channels shown at left.
L.K. Lustig—EB Inc.

pinnacles represent the remains of a formerly continuous sandstone layer, most of which has been long since eroded away. Other desert pinnacles, by way of contrast, may be the remains of algal deposits in a now-dry lake basin (Figure 16).

Many of the desert highlands, but not necessarily the smaller inselbergs and pinnacles, are surrounded by deposits of sediment that have been transported by floods from the highlands outward, toward desert plains. These deposits are commonly fan-shaped, with their apices located at the mouths of mountain canyons; they are called alluvial fans (*q.v.*), and they occur in most desert regions. If a relatively large open area exists beyond the mountain front, then the fan-shaped form may be nearly perfect, the size being governed only by the distance to which successive floods will debouch before fluvial transport of coarser sediment ceases. The alluvial fan at the mouth of Copper Canyon, in Death Valley, California, is a fine example of this circumstance (Figure 17, left). If, on the other hand, the sediment transported to the mountain front is intersected by a through-flowing transverse channel, then fan development may be nearly or entirely inhibited. This is the case along part of the mountain front in western Saudi Arabia (Figure 17, right). The similarity of colour and hue in this pair of photographs despite their very different geographic locales is noteworthy; the darker tones reflect the occurrence of desert varnish.

L.K. Lustig—EB Inc.

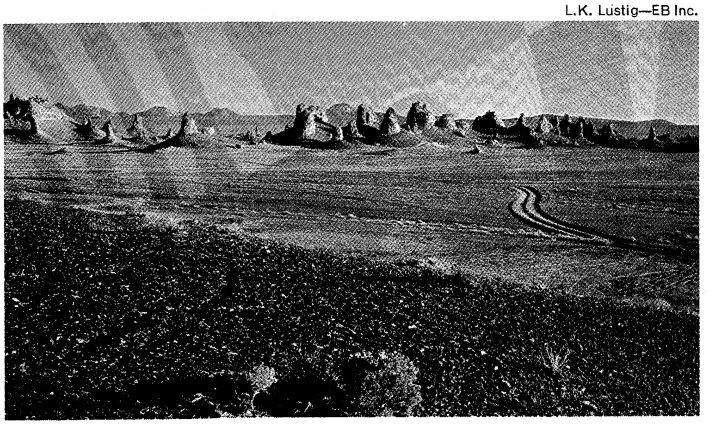


Figure 16: Pinnacles of calcareous tufa in the southern end of Searles Lake Basin, California. The tufa was deposited by algae in a former lake; the depth of water was at least as great as the height of the tallest pinnacle.

The same desert highlands that are fringed by alluvial fans, again excepting the smaller inselbergs and pinnacles, may also be bordered by sloping bedrock surfaces called pediments. These slope abruptly from the mountain front to the surrounding desert plain and may or may not be veneered with alluvium of varying thickness. The principal distinction between alluvial fans and pediments, aside from the fact that fans consist of unconsolidated sediments and pediments are solid rock surfaces, is that fans are depositional forms, whereas pediments represent the result of erosional processes. (For a thorough discussion of the several theories of pediment origin, see PEDIMENTS.)

Plains and basins

The lower boundaries of fans and pediments merge with desert plains, which by any estimate must be designated the predominant desert landform. In some desert regions the plains may be partially covered by shifting sands; elsewhere, the winds have winnowed the finer sediments, leaving a gravel surface in their wake (Figure 9); and in still other cases, the surface consists principally of exposed bedrock. Desert plains are generally characterized by their vast dimensions and extreme smoothness; there is much dispute among authorities with regard to their origin, however. Although it is generally held that the plains were formed by fluvial erosion in the distant past, the gradients involved are so small that it is difficult to conceive of streams traversing their entire length and breadth.

A number of rather large depressions in desert plains have been ascribed chiefly to deflation; that is, to erosion by wind action. Several basins in the Gobi and in the Atacama, together with the better known Qaṭṭārah Depression near al-ʿAlamayn, are in this category. Desert basins in general, however, are essentially structural features. In the basin and range terrain of the western United States, Iran, Pakistan, and elsewhere, the basins that exist were formed principally by the downward movement of large blocks of the Earth's crust between bordering, uplifted blocks that are now mountain ranges. Thus, desert basins alternate with mountain masses in these regions.

Smaller depressions in desert surfaces that are subject to periodic flooding and subsequent evaporation are termed dry lakes, playas, pans, and saline or alkali flats, among other names. They commonly occur at the lowest elevation within the larger basins discussed above; indeed, dry lakes are characteristic of basin and range regions. The playas in such areas may contain several thousands of feet of sedimentary fill beneath their saline crusts. The vertical sequence commonly includes silts, clays, and numerous salt beds, reflecting the occurrence of substantial hydrologic change during the alternating warm and cold episodes of the Pleistocene Epoch. Elsewhere, as in southern Africa and Western Australia, however, the climatic pulses were less pronounced, and only a few feet or tens of feet of sedimentary fill overlies bedrock in many instances. The pans of the Kalahari Desert (Figure 10) are of this type. Common to all dry lakes is the flatness of their upper surface; they are the flattest of all varieties of landforms, save for instances of disruption by pressure ridges which occasionally form along the margins of salt polygons. (For a detailed description of such surface features as well as of other characteristics of dry lakes in desert regions, see PLAYAS, PANS, AND SALINE FLATS; and for information on the nature of the associated salt deposits in arid regions, see EVAPORITES.)

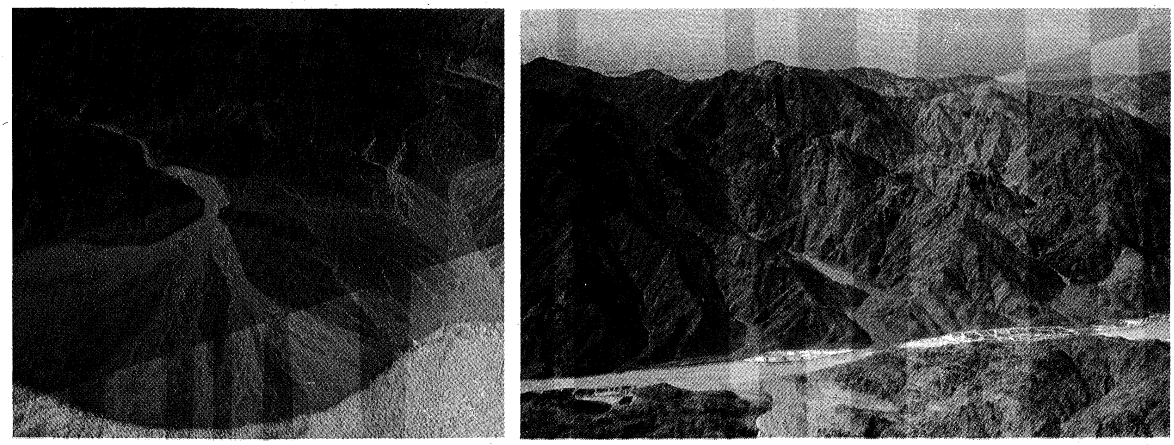


Figure 17: *Distribution of alluvium in mountainous regions.*
(Left) Alluvial fan at the mouth of Copper Canyon, Death Valley, California, an area of internal drainage. (Right) Mountains of western Saudi Arabia, north of Yeman (Ṣan'ā'), where occasional flow in the channel along the mountain front prevents the deposition of fans.
L.K. Lustig—EB Inc.

Desert sands

Desert sands and their configurations constitute the last category of landforms to be discussed here. In many ways they are the most interesting of desert landforms because of the complexities associated with the origin of the sands proper and the interpretation of the sheets and dunes that they comprise. It should be reiterated, however, that desert sands occupy only a minor fraction of the deserts as a whole. Even in the Sahara, which contains some of the largest sand seas in the world (*e.g.*, the Grand Erg Oriental in Algeria), the area occupied by sand has been estimated to be no more than about 10 percent. This figure would be somewhat greater for the Arabian Desert, in which the Rubʿ al-Khali and an-Nafūd are principally sandy areas, and for the Namib Desert of South West Africa which can be roughly divided between a southern sand-covered area and a northern area of gravel plains.

The sands involved in all instances are probably ancient, derived by the erosion of rocks in these areas millions of years ago. Many authorities believe that the great sand volumes of the Sahara and the Arabian Desert reflect accumulation in shallow seas during Tertiary time (65,000,000 to 2,500,000 years ago), followed by uplift of the land areas to their present elevations. In the Saharan case, some have argued that sand accumulation also may have occurred in ancient lakes, but the point to be made is that, contrary to popular belief, wind action is not responsible for the initial assemblage of these vast quantities of sand. The work of the wind has been to move, modify, add to, and subtract from these desert sands after their gathering by marine, fluvial, or lacustrine processes.

Many different kinds of sand dunes have been observed in desert areas, and a nearly equal number of classification systems have been set forth. Basically, dunes may be either fixed or mobile, the former designation being applied to those dunes sufficiently covered by vegetation to prevent their migration under prevailing winds. Some of the more common types of dunes are the following: longitudinal and transverse, referring to the orientation of linear dunes with respect to wind direction; sigmoidal, or S-shaped, dunes; pyramidal, domed, and star-shaped dunes; and the common crescent-shaped dunes termed barchans. Some longitudinal dunes in the western Kalahari are shown in Figure 18; they are fixed by vegetation on the dune crests. This type of dune also is widespread in the Australian Desert, where, again, the dunes are well spaced, symmetrical, with ridges ranging to about 30 metres in height. Sand dunes in general, however, may range in size from a small pile of sand about one metre in height to the so-called giant sand mountains that attain heights of 300 metres, as in the Rubʿ al-Khali, in the Arabian Desert as shown in Figure 19. These sand mountains commonly sustain other, smaller dune features on their surfaces.

L.K. Lustig—EB Inc.

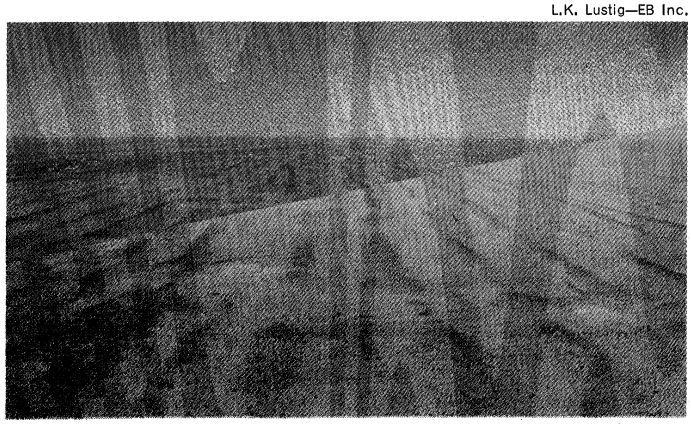


Figure 18: Longitudinal sand dunes in the western Kalahari, Botswana.

By courtesy of Arabian American Oil Company

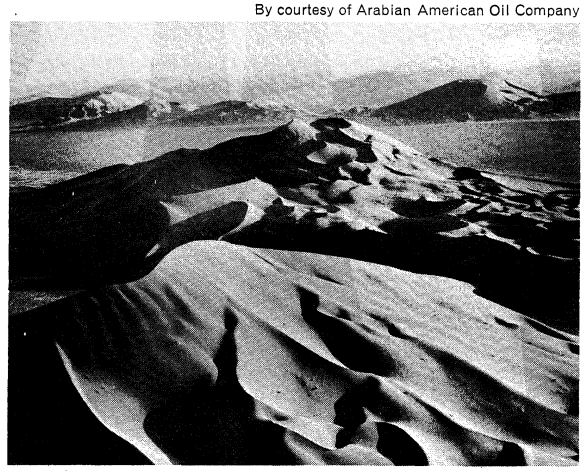


Figure 19: Giant sand mountains in the Rubʿ al-Khali, southern Saudi Arabia. These dunes attain heights of 300 metres; smaller dunes of several types may form on their surfaces.

Although a great amount of work has been done on the dynamics of windblown sand, it remains difficult to relate specific dune types to specific wind velocities, frequencies, and directions. This is partially because of a lack of satisfactory data on desert winds and partially because a time lag is always present—*i.e.*, a particular dune type or dune system may exist under the present prevailing wind regime but may still reflect the effects of an earlier and somewhat different wind regime (see further SAND SHEETS AND SAND DUNES).

THE PERMANENCE OF DESERTS

It was once believed that the present desert areas of the world had always been arid and that the phrase "once a desert always a desert" aptly summarized the subject of permanence. This view is patently indefensible, however, because there is abundant evidence that climatic changes have occurred throughout geological time and will doubtless continue to occur in the future. Changes of two kinds will be considered here, namely, those of the distant geological past and those of more recent vintage—within the past 3,000,000 years, approximately.

Climatic change. Meteorologists have argued on theoretical grounds that, because the Earth's rate of rotation and its heating by incident solar radiation have been essentially everpresent and invariant, some form of latitudinal climatic zonation like that existing today must have persisted throughout geological time. As indicated above in the section *General cause of the arid zones*, the existing climatic zonation and general atmospheric circulation are directly responsible for the Earth's two great belts of aridity at latitudes 30° N and 30° S, approximately. This being true, it is not surprising that the idea of the permanence of deserts once held sway.

Geological evidence

Considering but a few examples of relatively remote vintage, however, there are many ancient evaporite deposits in the Arctic areas today and sandstone beds with marked eolian characteristics in the mid-latitudes (including western Europe) that are as old as 200,000,000 years. Both evaporites and sandstones of this nature form today only under arid conditions. In today's deserts, on the other hand, many of the ancient rocks were formed under marine conditions, and even coals, reflecting a paludal (swampy) environment, are known. Moreover, in recent years an international group of respected geologists visited a number of locations in Algeria and concluded that widespread striations and other grooves in bedrock, heterogeneous sedimentary deposits, and other features definitely reflected a glacial origin. In fact, the group generally agreed that the present Saharan area was subjected to widespread glaciation during Ordovician time, around 450,000,000 years ago.

These examples indicate that some high- and middle-latitude areas have undergone arid conditions in the distant past, whereas the world's largest desert area today was at one time host to ice sheets. These data seemingly conflict with the meteorological view that the Earth's climatic zones have not changed through time. Reconciliation is possible, however, by recourse to the theories of continental drift and plate tectonics. The former, in its simplest aspect, holds that the present continents have "drifted" about on the Earth's surface since the breakup of a parent landmass about 200,000,000 years ago. Plate tectonics is essentially a refinement of this view, based on abundant evidence that ocean basins have been created along midoceanic ridges through time and that many of the Earth's major mountain systems and earthquake zones can be interpreted in terms of the intersection and abutment of six major "plates" (and a number of lesser ones) that comprise the Earth's crust and upper mantle. These plates—or plates of an earlier system—are thought to have been in motion since the beginning of geological time. The main point involved here is that the present climatic zones of aridity may well be relatively permanent global features but different land areas may have resided in these arid zones during different intervals of geological time. This then is one form of climatic change that has occurred, considering long-term changes.

Humid conditions during Pleistocene time

Evidence of climatic change in arid areas during the more recent past is rather abundant. Physical features that indicate the former existence of more humid conditions include (1) vast drainage systems that are today desiccated and partly filled by sand dunes, (2) the presence of large numbers of dry lakes in many parts of the world, some of which (*e.g.*, Great Salt Lake and Lake Chad) covered areas 10 times as great during intervals of Pleistocene time, (3) the presence of travertine and other spring deposits in currently arid areas, (4) the red soils and other soil varieties thought to form under tropical conditions that occur in the Sahara, Australia, and elsewhere, (5) rock weathering of many kinds that seems to require the presence of standing pools of water, (6) vast dune areas that are today fossilized because they have become fixed by vegetation, and (7) a variety of alluvial deposits and terrace sequences that have been interpreted in terms of climatic change; the persistent occurrence of gravel beds within a sequence of lake clays may reflect a change in hydrologic regime, for example (Figure 20).

L.K. Lustig—EB Inc.



Figure 20: Part of an irrigation system near Quetta, Pakistan, exposing alluvium in a basin floor. The gravel layer within the fine-grained sediments has been interpreted in terms of climatic change.

In addition, there are relict floras and faunas in arid areas that reflect more humid times; cave paintings, artifacts, and fossils that indicate the former presence of large mammalian faunas that were hunted by prehistoric man; and fluctuating widths of tree rings, as well as historic records of lake level and of streamflow, that suggest measurable climatic changes within the last few thousand years.

It is not claimed that arid areas were everywhere tropical during the several pluvial, or humid, episodes of the Pleistocene Epoch, when the ice ages descended on the middle- and high-latitude regions. Synchroneity of the ice ages and humid episodes in deserts is similarly doubtful, save in a few areas of basin and range topography. It is clear, however, that deserts are sensitive to the changes of precipitation and temperature that have occurred periodically throughout geological time and that these changes have had pronounced effects on the range, abundance, and kinds of plants and animals present and on the principal physical processes operative in the arid regions of the world. If this were not so, then the evidence cited would not exist.

Desert boundaries. Knowledge of the shifts in desert boundaries in response to climatic change, such as the estimated 200–300-kilometre shift toward the centre of the Sahara during the last (and mildest) moist episode of the Pleistocene, has occasionally led to cries of alarm that the deserts have been expanding in modern times. Such fears have been expressed with respect to the southern boundary of the Sahara and some parts of the Iranian and Thar deserts. In the latter case, evidence was assembled from all national scientific agencies responsible for the collection of relevant data, and a symposium on the question was organized and held. The conclusion in this case, as indeed in every instance of determined investigation, was that no evidence existed of natural expansion of desert boundaries. Overgrazing by livestock and the de-

nudation of hillslopes for firewood or fuel appear to be principally responsible for any changes from semi-arid to arid conditions in the vicinity of any desert boundary. It can be argued, however, that the precise location of any desert boundary is actually indeterminate except on small-scale maps and that constant pulsations, or expansions and contractions, about some mean position are the rule rather than the exception. This would accord with the climatic definition of these boundaries, the elements of which are known to change in value through time.

(L.K.L.)

LIFE IN THE DESERT

The desert as a biological environment. *Climatic patterns and the biota.* Biological activity is inevitably limited in the absence of water. Above all, it is the green plant that is so limited. Few plants can continue photosynthesis for long in the absence of water; and in the absence of plant growth other living forms have no opportunity to develop. Animals and micro-organisms can exist only if energy sources from green plants are available. They are further limited, though, in the absence of water, for most animals and all micro-organisms can be active only while moisture is adequate.

Plants as the basis of living communities

In the cold deserts in the interior of continental masses, drought is severe for at least a part of the year, and the daily and annual range of temperature may be great. Moreover, very low winter temperatures have a more severe impact on the biota than they might in a region with greater precipitation, because the protection by snow cover is less.

In the tropical deserts cold is no problem, but the clear skies result in a high radiation load on organisms and in high daytime temperatures, often with very large fluctuations from night to day. These factors, too, create more intense stresses on organisms than occur in most environments.

Other effects on the biota of the desert environment arise from the abrasive effects of windblown sand and from the high surface salinity in areas where temporary waters accumulate and evaporate or where evaporation leads to upward movement of salts through the soil profile.

Limited leaching and the small accumulation of organic matter in deserts result in desert soils remaining much closer to the types found in pioneer habitats than do soils in more moist environments. Desert soils seem particularly subject to the formation of impervious surface layers—either by physical processes or through the surface development of algae and lichens—so that the effective drought is accentuated by excessive runoff.

The pioneering conditions that organisms encounter in the desert are permanent, unlike those of other bare and inhospitable habitats, such as coastal sand dunes, rocks, and denuded areas left by human activity. In these climatically more favoured situations, the pioneering organisms initiate changes in the parent soil material that render it more hospitable to newcomers; and a succession of different living communities, each more complex than the last, can take place, leading to a climax community. In the deserts, harsh conditions are not ameliorated by biological activity, and the pioneers themselves constitute the climax.

There are limited desert areas—notably along the coast of South America but also in Baja California, Mauritania, and South West Africa—where the extremely low rainfall is supplemented by dew and the aridity is ameliorated by frequent mist and cloud. The relatively low evaporation rates make conditions in such areas much more moist than their precipitation records would suggest.

Spatial dispersion of biota. All of the organisms in a desert do not live under the same environmental conditions. Although the variety of habitats is less than in many wetter environments, it is still great enough to provide effective living conditions for a wide diversity of organisms. Moreover, many organisms show structural and behavioral responses to the environment that tend to neutralize at least some of its extreme features.

It is well known, for instance, that perennial plants in the desert tend to be widely spaced. But, although there are large unoccupied areas between the aerial parts of the plants, excavation may reveal that the root systems are in contact and they exploit the soil so effectively that establishment of new individuals becomes difficult. When rain penetrates the soil, it is rapidly absorbed by the continuous net of roots, and little penetrates to the subsoil. If situations differing in rainfall but otherwise similar are compared, it is found that the plants are more widely spaced as the rainfall decreases. This implies that the amount of plant material remains roughly proportional to the rainfall and that the individual plant is receiving about the same amount of rainfall, thanks to its inhibitory effects on the establishment of competitors.

A similar response leads to the establishment of denser vegetation around the watercourses, even when they run very infrequently, for there the longer periods of availability of subsoil water permit more plants to have similar productivity without more water per individual.

Different plant species in the desert under the same climatic conditions may differ markedly in their local environment, thanks to differences in soil conditions and topography. Even in the tropics there may be a marked difference in growth conditions between the steeper slopes that face north and those that face south. Desert soils, though with little organic matter and less well developed profiles than are usual in more humid conditions, differ sufficiently in their mechanical and chemical properties from place to place to support different types of vegetation. And where stones and boulders are scattered over the soil surface, the microclimates they generate—and the local increase in rainfall penetrating the soil around their edges—permit even some moisture-loving plants such as mosses to develop. Oases similarly are places where a locally dense vegetation corresponds with a locally available water supply.

The effects of soil types and local elevation differences on desert environment

For animals, the variety of habitats is still greater, for they are generated by the vegetation itself, and since animals are not rooted and dependent on light as green plants are—facts that oblige plants to be more or less exposed to the prevailing atmospheric conditions, however harsh they may be—animals are free to move to more comfortable locations if they desire, even underground. Each species of plant modifies the conditions for the animals around it and may generate living conditions that permit particular animal species to exist, whether providing food, shelter (shade or protection from predators), or layers of litter in which they can live. Movement between different habitats further varies the living range of animals, as when bats roost in caves during the day and emerge in the evening to feed or when the subterranean larvae of darkling beetles, feeding on living and dead roots, emerge as adults into the harsh world of the desert soil surfaces to begin their reproductive activities.

Like plants, some animals may show a wider dispersion in deserts than elsewhere, thus giving each individual a broader area of ground as "his own." Kangaroo rats, for instance, are reported to manifest more aggressive behaviour in the deserts than in other habitats, resulting in larger and more effectively defended territories for each individual.

Types of desert ecosystems. *Cold deserts and shrub steppes.* These occur mainly in the interior of Asia and in the intermountain zone of North America. Winter temperatures are well below freezing, and the ground may be snow covered for considerable periods. Summer temperatures are moderately high (up to 40° C), though the temperature extremes of regions further south near the tropics are avoided since the sun is not so high in the sky. Precipitation often is between 200 and 300 millimetres (8 and 12 inches), and a good part of it commonly falls as snow.

The dominant vegetation is of shrubs, with a discontinuous canopy, but not as widely spaced as in the drier deserts nearer the Equator. In the spring, as the snow melts, there is a flush of growth, and many short-lived annual plants occupy the soil surface. Among the shrubs, species of the genus *Artemisia* are prominent, as are a number

of species of the family Chenopodiaceae, particularly in lower lying areas, where there is some accumulation of salt.

Shrub deserts and thorn scrub. These ecosystem types occupy a large part of the area of the tropical deserts. The peripheral zone of the Sahara (the Sahel) is dominated by species of *Acacia*, many of them thorny. The Mojave and Chihuahuan deserts in North America are dominated by a range of shrubs, of which the creosote bush (*Larrea tridentata*) is probably the commonest. The same is true in similar deserts of South America. Much of the Australian dry country is dominated by sparse small trees of the very variable genus *Acacia*, while other species of this genus, *Eremophila, Cassia,* and members of the family Chenopodiaceae form a denser understory. In other peripheral parts of the Australian arid zone, the low-growing mallees (*Eucalyptus* species) take the place of the treelike *Acacia* species.

Under the shrubs there is often a partial ground cover of tufted perennial grasses, while the interspaces are filled with short-lived, small flowering plants following rain.

Larger plant-eating animals are not uncommon in this type of country, entering from the surrounding areas of somewhat higher rainfall and more grass during periods when forage is more abundant and retreating when times are harder. Some, such as the desert white-tailed deer and collar peccary in America and the kangaroo in Australia, make their permanent home in such conditions.

Stem-succulent deserts. These striking formations are of rather limited extent and consist of only the Sonoran Desert (Gran Desierto) of Arizona and northwestern Mexico, the area between Peru and Argentina, and the Namib Desert in South West Africa. Their special appearance rests on the geographical availability of certain plant groups absent from many of the deserts—the cactus family (Cactaceae) of the Americas and the arborescent members of the family Euphorbiaceae mainly found in Africa. These plants have adopted a special growth habit that provides one alternative solution to the problem of survival in drought. Elsewhere, their role is taken over by shrubs.

The deserts of cactus and cactus-like plants

Herbaceous deserts. There are desert areas where shrubs or stem succulents are few or absent and where the bulk of vegetative matter is in the form of perennial herbs. These are nonwoody plants whose tops die back each year but whose roots survive to produce new tops year after year. The Nullarbor Plain in southern Australia is of this type; the soil is thin over a continuous limestone substrate, permitting water to penetrate freely, and these harsh conditions are doubtless responsible for the lack of shrubs. Even the perennial species of the family Chenopodiaceae form only a sparse ground cover. In southern Africa, there are areas where the dominant vegetation is formed of low-growing succulents.

Salt deserts. Within the deserts, there are many areas where the soil contains a high concentration of salts; most commonly sodium chloride is predominant, but in some places sulfates or carbonates may be more important, and other substances may also, in part, take the place of sodium. These areas are usually in basins of internal drainage, and they may surround highly saline lakes, such as the Dead Sea in Jordan and Israel and the Great Salt Lake in the western United States, or large salt pans flooded only after exceptional rains, such as Lake Eyre in South Australia. Very few organisms can grow in a concentrated salt solution, and in consequence the area of a salt pan has a very simple community and may appear quite bare. As the salinity decreases, either around margins of such an area or through seasons of high rainfall, a limited range of higher plants can establish themselves—mainly shrubs or sub-shrubs of the family Chenopodiaceae. It is noteworthy that these are closely related to—sometimes, indeed, the same as—the species found in the coastal salt marshes. There, too, a high and variable salinity obtains, and near the spring high-tide mark, in some seasons, high salinity and low moisture content persist for a week or more. In the salt deserts, on the other hand, these conditions may persist for months. Whereas in salt marshes the increase in soil level with buildup of organic matter may eventually take the area above tide level, so that leaching can take place and succession proceed, the vegetation in salt deserts is a soil-dependent climax and will not develop further.

Sand dunes. Where sand dunes occur in deserts, their consolidation or stabilization by vegetation is much more difficult than along seacoasts. In both situations consolidation depends on plant species with wide-spreading root or rhizome (underground rootlike stems) systems. At the coast, this normally permits invasion and establishment of other species, and thus begins a succession. In the desert, the pioneers are the climax, and the barely consolidated dunes are likely to become mobile again if any disturbance breaks the fragile cover and permits wind blowouts. The free drainage of the sands implies that moisture conditions, as well as substrate mobility, are particularly harmful to plant life; and sand dunes in an arid landscape are likely to be among the most persistently bare areas on the Earth's surface.

Vegetation on dunes and along ephemeral streams

Wadis and watercourses. Along the ephemeral wadis and watercourses of the desert, the occasional violence of the waters discourages perennial plants; but along the sides, where flow is less frequent and less violent, a special vegetation of deep-rooted moisture-requiring shrubs and trees develops. These benefit not only from water that actually flows down the channel but also from its much deeper penetration into the soil—often quite permeable in these situations—and consequent longer availability. Some of these plants, called phreatophytes, may be able to tap water many metres below the surface and thus survive a drought during which most of the vegetation around dies. A special population of birds and other animals is associated with the vegetation around the watercourses.

Aquatic ecosystems in the deserts. A torrent that flows for two or three days and is then dry clearly does not permit the development of an aquatic ecosystem. But some waters in the deserts are longer lived, and they have their own special biota. Even in the temporary watercourses, there may be pools that remain filled for weeks, and in these an ecosystem with algae, insect larvae, and amphibians will develop. The same applies to the temporary lakes known as playa lakes lying in hollows and filled for a month or so following rain. If these temporary lakes are saline, the ecosystem may be much simpler, for the salt concentration may considerably exceed that of seawater and the species that can endure it are few.

Virtually the only permanent waters of low salinity originating in the deserts are springs, renewed by flow from underground. These may develop complete ecosystems of similar structure to aqueous systems elsewhere, though generally simpler—due, doubtless, to isolation. These include rooted aquatic and riparian higher plants, algae, crustaceans, insects, amphibians, and fish.

The desert biota and their adaptations. *Protists.* Soil micro-organisms occur in desert soils in about the same degree of diversity as in other soils and can survive frequent desiccation and remoistening. In some deserts, there is a noteworthy development on the soil surface of blue-green algae, which are thought to be active in nitrogen fixation (converting atmospheric nitrogen to forms that plants can utilize) and may also modify the infiltration properties of the soil.

Plants. A great deal has been written of desert plants and their adaptations to the rigours of desert life. Apart from the blue-green algae mentioned above, other algae are practically confined to the limited water areas of the desert. Parasitic fungi are as abundant as in other environments, though mainly confined to the higher groups not requiring free water for swimming reproductive cells. Saprophytic fungi (*i.e.*, fungi obtaining food from dead organic matter) are not abundant, the limited organic matter in and on the soil providing little in the way of nourishment, and active growth being possible only during the short periods when the soil is moist.

Algae, fungi, lichens, and mosses

Lichens are numerous and often prominent—particularly the crustose types. They often form an almost continu-

ous covering to the soil surface and are then able to take advantage of any wetting of the soil for short-lived periods of photosynthesis.

Mosses and liverworts are few and are often confined to special habitats more moist than the deserts as a whole. Ferns are not infrequently found in rock crevices and similar habitats. Club mosses also occur, some showing the remarkable adaptation of being able to dry out and withstand desiccation and lack of active growth in the form of a ball, which may be blown about freely.

Some gymnosperms (nonflowering seed plants) are noteworthy inhabitants of the deserts. Conifers and cycads are few in desert areas, but several species of *Ephedra* are important desert shrubs—a role to which their near-leaflessness suits them; on the other hand, the remarkable *Welwitschia*, with its large, split leaves, is also a desert plant, confined to the Namib in South West Africa.

Among the angiosperms (flowering seed plants), there are some families that are especially characteristic of the deserts. The family Chenopodiaceae is such a characteristic inhabitant of salt deserts (and other saline habitats) throughout the world. Some members of the family Zygophyllaceae are also normally desert plants. In the Americas, species of the family Cactaceae are characteristic of the deserts, and the small family of Fouquieriaceae are found there only. Another group playing an important part in many of the world's deserts, though not confined to them, is the family Mimosaceae, and particularly the genus *Acacia*. Species of this genus characterize the peripheral zones of the Sahara, are dominant in many parts of the Australian desert, and are far from negligible in the American and Asian deserts.

Plants in the deserts have to surmount a difficulty that the photosynthetic mode of nutrition imposes on them. In order to produce food by means of photosynthesis, carbon dioxide from the atmosphere must be taken into the plant. This unavoidably leads to the loss of water from the plant at the same time, since the same small openings in the plant surface (called stomates) serve for both purposes. Various solutions to this problem are to be seen in desert plants.

One way to avoid the adverse effects of limited water availability is to live through dry periods in a completely inactive form—as a seed. This is the solution to the problem adopted by the ephemerals, or short-lived annual plants. These plants shorten the duration of their life cycle so that seed germination follows very promptly after an adequate storm and seed formation is already in progress by the time soil moisture is reduced to the wilting point. This means, of course, that the total period of photosynthesis is limited to a few weeks every year—perhaps in some years to none at all. Nevertheless, this group of plants is a very successful element in desert vegetation, as is evident in the flowery display often seen over the desert floor a month or so after rain.

Another solution, rather similar, is for the plant to restrict itself during the dry season to inactive tissues that have very limited gas exchange and to form tissues capable of free gaseous diffusion only during periods when soil moisture is available. This mode of growth is called the deciduous habit, and it is expressed by plants everywhere that drop their leaves during winter or, in the case of deserts, during dry periods. A good many species of desert plants have short-lived leaves and are bare for much of the year. The spectacular ocotillo (*Fouquieria splendens*) of the Sonoran Desert of southwestern North America is one example.

The problem of water loss during photosynthesis

Another solution is to separate temporally (*i.e.*, in time) the gaseous-diffusion process from that of energy utilization. Normally, the photosynthetic surfaces must be open to carbon dioxide diffusion during daylight periods, in order that radiant energy may be converted to chemical energy in the photosynthetic process, and this incurs the penalty of water-vapour loss at a time when the diffusion gradient for it is particularly high because of high daytime temperatures and dry air and that for carbon dioxide is no more favourable than at any other time. If carbon dioxide can be taken up at night instead, the concomitant loss of water can be much reduced. Many desert plants have this ability, carbon dioxide being stored in the form of carboxylic acids at night, which are converted to carbohydrates when radiant energy becomes available in the day. The stomates in these plants open by night to allow entry of carbon dioxide into the intercellular spaces and close by day to restrict loss of water vapour while photosynthetic reduction of the "stored" carbon dioxide proceeds. This so-called crassulacean metabolism is not confined to the family Crassulaceae but is found in a good many other plants of the deserts.

Another mechanism, like the deciduous habit, restricts gas exchange when drought conditions prevail not by loss of leaves (a wasteful process) but by direct restriction of gaseous diffusion. The stomates, through which most gas exchange occurs, can generally open or close according to prevailing environmental conditions; adaptations may increase their responsiveness to water stress and at the same time reduce the amount of gaseous diffusion in other ways. The means used include increased thickness of the cuticle (outer layer of stem and leaves), development of wax layers, rolling of leaves in drought, and so forth—a series of adaptations often described under the heads of xeromorphy or xerophily.

Another mechanism, often combined with restriction of gaseous diffusion by stomates during periods of drought and with crassulacean metabolism, is water storage in succulent tissues. This has developed in rather few taxonomic groups—the families Aizoaceae, Crassulaceae, Cactaceae, some Asclepiadaceae (especially the subfamily Stapeliae), some Chenopodiaceae, some Liliaceae, the arborescent Euphorbiaceae, and a few others.

In some cases, plants can grow in deserts without suffering water shortage through having an extremely deep root system penetrating to subsoil water. A high ratio of root to shoot tissue is indeed a common feature of desert plants, whether the root system is deep or spreads widely.

Adaptations of plants to high temperatures

A hazard of the deserts, when water loss from aerial parts is restricted, is the high temperature in these organs that may result from intense sunlight when cooling by evaporation is limited. Resistance to high tissue temperatures may be part of the adaptational equipment of these plants.

Temperatures on the soil surface in deserts may be very high (70° C, or 158° F), and plants growing there must be prepared to endure them. Lichens and blue-green algae, which compose the bulk of the soil-surface flora, are usually dry at times when the soil-surface temperature is highest, in which state they are highly resistant to extreme heat. Perennial plants usually provide shade to the soil immediately around their bases, so that very high temperatures are eliminated there. For small annuals, on the other hand, high soil temperatures around their stem bases when soil moisture has been depleted may cause some death of tissues and accelerate flower and seed maturation.

Another factor to which adaptation is required in some desert situations is the high salinity of the soil solution. Usually plants growing in such situations have difficulty taking water into their roots from such solutions. Many plants have adapted to these conditions with a highly concentrated sap, enabling the root to take up water even from saline solutions.

Invertebrate animals. Many groups of invertebrates require a marine environment and are consequently absent from the deserts; other groups occur in desert waters but not on land. The flatworms (phylum Platyhelminthes) occur as parasites inside the bodies of animals, as elsewhere. The various groups of nematodes (phylum Aschelminthes)—animal parasites, plant parasites, and those living free in the soil—are found in the deserts as in other habitats, though activity in the soil is restricted by their need for free water. The segmented worms (phylum Annelida) are represented by a few earthworms, occurring only in favoured situations. There are a fair number of snails found in deserts, some species of which are remarkably resistant to arid conditions.

It is, however, of the insects and spiders and their relatives (phylum Arthropoda) that one thinks when desert invertebrates are mentioned. The crustaceans of that phy-

lum are represented by the brine shrimps that live in concentrated salt solutions and by the terrestrial isopods (wood lice, or sow bugs) that occur fairly widely. Myriopods are reasonably numerous in the desert, both the vegetarian millipedes and the carnivorous centipedes. All the groups of arachnids (spiders, scorpions, etc.) except the primitive marine *Limulus* occur in the deserts, and some are more at home than in many other environments. Scorpions are numerous, as are ground-dwelling spiders (web-building spiders are not particularly common), and there are plenty of mites, including those inhabiting the soil.

Insects are abundant, though forms with aquatic larvae such as the dragonflies and mayflies are found only around the waters; termites, beetles, butterflies, moths, flies, ants, bees, and wasps are all universally present in deserts, as are the grasshoppers and true bugs.

Adaptations to desert conditions by insects and their relatives

For many insects and arachnids, the problems of life in the deserts are easier than for most vertebrates. Like many desert plants, they have a waterproof cuticle, or skin, to retain water; and the frequently short life cycle, often combined with metamorphosis (changes during the life cycle from egg to larva, etc.), enables vulnerable stages to be timed for the less adverse seasons. By persisting through unfavourable periods in an inactive stage (egg or pupa), the animal can avoid many of the difficulties of desert life, much as does the ephemeral plant that persists as a seed or the deciduous plant when reduced to inactive woody tissue during dry periods. Another means of avoiding environmental difficulties is to burrow beneath the soil. Many arthropods burrow for at least a part of their life cycle; and there is an abundant fauna that spends its whole life beneath the soil. This is true of nematodes and the few desert earthworms, of course, as well as large numbers of mites and almost all insects of the order Collembola (springtails). These soil arthropods form part of a trophic cycle (food chain) within the soil mass, in which material originally derived from plant roots is eaten by herbivores and they, in turn, by several carnivore levels. This material eventually returns to detritus, or the decaying organic state, to be acted upon by micro-organisms, which are again taken into the animal sequence.

Some of the subterranean arthropods are larval stages of adults that spend their life on the soil surface. There are important groups, however, whose adult life is divided between the upper and lower worlds. The most important are the ants and the termites. Not all ants have subterranean nests, but, in deserts, by far the majority do so; and all termites avoid the light, building covered runways when they come above ground level or developing channels through woody stems. Thus the ants, and even more the termites, avoid the risk of exposing the more vulnerable immature forms to the high radiation, temperature, and evaporation rate of the open desert; and they limit the time of exposure of the more resistant adults to what is required for food gathering (in the case of ants) and defense.

Vertebrate animals. The desert waters, when permanent and not too saline, contain fish, and in some cases the species are endemic (peculiar to the locality), because of the isolation of some of these ecosystems. They are generally closely related to species in nearby non-desert areas. Little is known of their special adaptations, though they must often tolerate a higher and more variable salinity than do their relatives elsewhere.

Frogs and toads are far from uncommon in the deserts, despite their need of an aqueous environment for reproduction. Consequently, desert amphibians need to be opportunistic, responding very quickly to rain that leaves standing water and then passing through the stages of egg laying and larval life before this water dries up. The selection pressure for rapid larval development must be very intense, for those larvae still dependent on gill breathing when the pools dry will die. The adults, with a moist and unprotected skin, are also very vulnerable to high evaporation rate and spend most of the time between rains inactive in a protected spot, often burrowing deep into the soil.

Reptiles are perhaps the most characteristic group of the desert animals. Lizards and snakes are numerous, tortoises and even turtles occur, and crocodiles are common in some tropical desert waters. The majority of the lizards are insectivorous, though some are herbivores. The snakes mostly feed on other vertebrates, in the main small mammals and birds' eggs and nestlings. Some of them have refuges in holes in the soil—perhaps more from predators than from the climate. But the majority spend most of their life exposed to high temperatures and evaporation on the surface of the soil. Their scaly skin is indeed resistant to water loss, but the radiation load they suffer is often considerable. They have no means of temperature regulation except seeking shade, and their tissues must be resistant to temperatures well above the normal limits for warm-blooded animals. Though some show bright warning coloration, many are camouflaged in the colours of the soil surfaces on which they live.

Birds are as numerous in deserts as food availability permits but often spend only part of their lives in this environment. Apart from seasonal migration, they may move between roosting or nesting sites and feeding areas. The lack of trees or tall shrubs to provide nesting positions excludes many species (apart from foraging excursions), and many desert birds nest on the ground. The eggs are in some cases covered with mounds of vegetable material, which give partial protection and provide a more uniform temperature for incubation. Running birds are numerous in the deserts and include some of the largest—the ostrich in South West Africa (and previously in North Africa and the eastern Mediterranean countries); and the emu in Australia, both of which are herbivores. Some birds are present that are prominent predators and carrion eaters, such as the roadrunner and magpie of the North American deserts. Raptors (hawks, eagles, etc.) are important and have in the deserts excellent visibility for attacking prey from the air. They may nest in rocky parts of the deserts themselves, but their range of movement is such that they can often operate from areas in adjacent but different vegetation.

Desert birds

Some bird species show notable adaptation to the desert environment in their reproductive habits. Gambel quail, for instance, do not reproduce in years with below-normal rainfall, when food for the young would be scanty; ducks in central Australia have lost the annual reproductive rhythm and will breed at any time when rain falls.

There are many mammals in the deserts, though the modifications required in behaviour or physiology (or in both of these) are considerably greater than for the reptiles. This is largely because they are warm-blooded and because they excrete urea rather than uric acid, which imposes extra demands on body water.

Most mammalian orders, other than those that are purely aquatic, are found in the deserts. Primates (apes, monkeys, etc.), mainly tree dwellers, are naturally not prominent, although several species of baboon are found in the African and Arabian deserts. Edentates (anteaters) and insectivores are few, but bats are often numerous, roosting in the caves and rock crevices that are commonly found in desert regions. The rodents (rats, mice, etc.) and lagomorphs (rabbits, hares, etc.) form the largest group of desert-inhabiting mammals. Some of them are remarkably successful there—the European rabbit, for instance, which, following its introduction from Europe into Australia in the 19th century, not only spread through the temperate parts of the continent but also colonized a large part of the southern desert region. The native rodents of the deserts often show marked physiological adaptation to the desert, water requirements being reduced to a minimum. Some, in fact, can subsist on the metabolic water produced by respiration of their foodstuffs. Many desert mammals avoid high radiation loads and decrease their water losses by spending the midday hours in the shade of bushes and feeding in the early morning, late evening, or at night. Many seek the shelter of burrows in the earth, with their much more moist and consistent conditions, at least during the daytime hours, and their life above ground may be limited to nocturnal foraging.

Herbivores in the desert can benefit from the ability to move rapidly, from one feeding place to another or between feeding and drinking places, so it is not surprising that a fair number of ungulates (hoofed animals) make the desert their home. Many are nomadic and range over large areas, thus being able to benefit from vegetation growth following local storms, without the risk of starvation imposed by eating out the forage within a limited territory. There are numerous antelopes, goats, and sheep in the deserts of all continents but Australia. Some of the wild asses are at home in the deserts of western Asia and North Africa, and escaped domestic equines (burros in North America, brumbies in Australia) have taken successfully to the deserts and desert margins. Outstandingly successful among ungulates in the deserts, however, are the camels, whose adaptation to life there is well-known. Their broad feet suit them to movement over sandy areas (though such areas are extensive only in limited parts of the desert regions); they have large reserves of fat, to enable them to live through long periods of scanty forage; and they can live for many days without drinking, making up their water deficit by enormous drafts when opportunity presents itself.

Among the carnivores, lions inhabit the desert margins of Africa and Asia, and some smaller cats also hunt in the deserts, both in the Old World and the New. Several of the hyenas are desert animals, and some of the dogs are also prominent desert carnivores, notably the coyote and kit fox in North America and the dingo in Australia.

The marsupials (pouched animals) include a number of species well adapted to life in the Australian desert—many of the kangaroos and wallabies and a variety of smaller herbivorous and omnivorous species.

Man. A wide variety of human races have adapted successfully to life in the deserts. Caucasoid peoples inhabit the deserts of North Africa and western Asia, and the Australian Aborigines, thought to be more closely related to them than to other major racial groups, live in the deserts of Australia. Negroid peoples live on the southern margins of the North African desert and in South West Africa. Mongoloid peoples occupy the Central Asian deserts and those of North and South America.

As would be expected in an animal of such behavioral diversity as man, his adaptations have been behavioral and cultural rather than physiological. Irrigation civilizations have been developed in many arid regions of the world—in fact, the Nile Valley and the valley of the Tigris and Euphrates, which were two of the cradles of civilization, run through deserts. Where irrigation was not possible, low vegetative productivity imposed a nomadic culture on desert-dwelling societies.

Human uses of deserts

Apart from nomadism and irrigation works, man's behavioral adaptations include his use of thick-walled dwellings with few windows; water storage in deep, covered cisterns; clothing, often white, protecting against excessive radiation; and domestication of desert animals, notably the camel.

Productivity of deserts. The natural productivity of the deserts is low compared with that of most other ecosystems. The dependence of all plants on water and of other organisms on plants means that the arid conditions limit productivity where radiation, temperature, and other growth factors would otherwise permit the highest levels of productivity on the Earth. The proportion of solar energy incorporated in organic matter, rarely exceeding 2 percent in natural communities, is in the deserts a small fraction of 1 percent. The bulk, or biomass, of vegetation in the Syrian Desert has been reported as ranging from 470 to 4,800 kilograms per hectare (kg/ha; or 420 to 4,300 pounds per acre) and, in the cold deserts of Central Asia, from 55 to 7,000 kg/ha.

In cold deserts in North America dominated by sagebrush, figures between 1,000 and 3,000 kg/ha of dry matter per year have been recorded, as compared with 50,000 to 100,000 kg/ha for mountain forests in the same latitude. For the Sonoran Desert a figure of 1,400 kg/ha per year has been quoted.

While the standing plant biomass may be well over 1,000 kg/ha, the animal biomass is far less. Figures on which to base even a rough estimate are very scanty, but 100 kg/ha would probably be exceptionally high, and 10 kg/ha the more usual order of magnitude.

The deserts have been used by food-gathering human societies for many millennia. Provided the population is sparse and nomadic, a reasonably reliable living may be wrung even from these unpromising habitats. Good examples are the Australian Aborigines, who made use of a wide range of fruits and roots and insects, reptiles, and mammals. Other peoples use the deserts by grazing on them flocks of domesticated livestock and constantly moving the flocks to fresh pastures as the limited growth is grazed off. (D.W.G.)

BIBLIOGRAPHY. The literature on arid regions is both vast and specialized and there are few general works that embrace it in its entirety. The interested reader should consult WILLIAM G. MCGINNIES, BRAM J. GOLDMAN, and PATRICIA PAYLORE (eds.), *Deserts of the World: An Appraisal of Research into Their Physical and Biological Environments* (1968), for reviews of work on weather and climate, geomorphology and surface hydrology, surface materials, and plants and animals in the arid regions. Each of these reviews is accompanied by an extensive bibliography. Another excellent source of information has been provided by the UNESCO series of publications categorized as reviews of research or proceedings of symposia. Important volumes in the series include: *Reviews of Research on Arid Zone Hydrology* (1953); *Proceedings of the Ankara Symposium on Arid Zone Hydrology* (1953); *A History of Land Use in Arid Regions* (1961); *The Problems of the Arid Zone* (1962); *Changes of Climate* (1963); and *Geography of Coastal Deserts* (1966). The latter provides a complete descriptive summary of every coastal desert region in the world and contains much information that was touched on only briefly in this article.

Some useful articles on specific characteristics of deserts and on the potentialities of the arid regions are presented in WILLIAM G. MCGINNIES and BRAM J. GOLDMAN (eds.), *Arid Lands in Perspective* (1969). Other books and monographs covering the physical characteristics of the several world deserts are listed in the bibliographies accompanying articles on particular deserts. A few papers from the more specialized journals might be mentioned here, however, to serve as a guide to some of the topics not normally mentioned in popular accounts. Two summary papers by R.F. PEEL provide excellent insight to knowledge of the arid regions and the limitations thereof. These are: "Some Aspects of Desert Geomorphology," *Geography*, 45:241–262 (1960); and "The Landscape in Aridity," *Trans. Inst. Br. Geogr.*, no. 38 (1966). Interpretations of alluvial sequences, dune patterns, and lake expansion are set forth in three papers treating Saharan regions, namely: C. VITA-FINZI, "Late Quaternary Alluvial Chronology of Northern Algeria," *Man*, 2:205–215 (1967); A.T. GROVE, "The Ancient Erg of Hausaland, and Similar Formations on the South Side of the Sahara," *Geogrl. J.*, 124: 528–533 (1958); and R.A. PULLAN, "The Recent Geomorphological Evolution of the South Central Part of the Chad Basin," *J. W. Afr. Sci. Ass.*, 9:115–139 (1964). In this vein, the recent work of KARL W. BUTZER and CARL L. HANSEN, *Desert and River in Nubia* (1968), is an outstanding example of such interpretations of alluvial chronology and history in the Nile Basin.

The phenomenon of piping in arid regions is also seldom mentioned in general works and the interested reader should consult the basic paper by G.G. PARKER, "Piping: A Geomorphic Agent in Landform Development of the Drylands," *International Association of Scientific Hydrology*, no. 65, pp. 103–113 (1963). Finally, the subject of possible glaciation in Africa and Australia is well covered in two modern papers, namely those by L.A. FRAKES and J.C. CROWELL, "Late Paleozoic Glaciation. II, Africa Exclusive of the Karroo Basin," *Bull. Geol. Soc. Am.*, 81:2261–86 (1970); and R.W. GALLOWAY, "Late Quaternary Climates in Australia," *J. Geol.*, 73: 603–618 (1965).

None of the works cited here should be considered as the best or only papers available; they are merely examples of the thousands of specialized articles to be found in the literature on physical aspects of the arid regions. The biological aspects of deserts are also treated in such articles, but the following books and symposium volumes provide good summary coverage: GEORGE W. BROWN (ed.), *Desert Biology* (1968), a symposium volume in which numerous specialist authors discuss their own fields of desert biology in varying depth—attention is concentrated on the deserts of North America; J.L. CLOUDSLEY-THOMPSON (ed.), *Biology of Deserts: The Proceedings of a Symposium on the Biology of Hot and Cold Deserts Organized by the Institute of Biology*

(1954), includes 28 papers on a wide range of biological topics relevant to desert life, some in French, and mainly relating to the deserts of Africa and the Near East; *Desert Life* (1965), a general semipopular account; and with M.J. CHADWICK, *Life in Deserts* (1964), a general account intended for use in colleges; and KNUT SCHMIDT-NIELSEN, *Desert Animals: Physiological Problems of Heat and Water* (1964), a discussion of how vertebrates (particularly mammals) adapt themselves to the shortage of water and often its salinity, and to the high temperatures in the deserts.

(L.K.L./D.W.G.)

Detroit

Its name an international byword for the American automotive industry, Detroit and its metropolitan area in southeastern Michigan comprise one of the major industrial–commercial complexes in the United States. In spite of an increasingly diversified manufacturing and shipping base, its economy remains unusually sensitive to national or international events that affect the prosperity of its major industry, automobile production, and economic booms and depressions tend to be felt more keenly in Detroit than in most areas of the country.

General character of the city

Fronting the Detroit River, which connects Lake Erie with Lake St. Clair, the impressive skyline of downtown Detroit looks across the international boundary to Windsor, Ontario. By the early 1970s, however, this facade disguised a city in which the characteristic urban problems of the U.S. city existed to an extreme degree. Of its 1970 population of more than 1,500,000 (metropolitan area nearly 4,200,000), some 44 percent were black, and racial tensions that had exploded in 1967 threatened to do so again at any time. The city was losing industry and much of the more affluent white community to the suburbs, and its ability to meet the constantly spiralling costs of welfare and other social services was diminishing. Crumbling ghettoes covered much of its landscape, and problems of air and water pollution added to the civic malaise. Other observers saw, on the other hand, more than a few signs of a city making a dynamic comeback: especially in the many areas of physical rebuilding and in the many persons, particularly from the black community, moving into positions of leadership and providing a new sense of buoyancy. (For information on related topics, see the articles MICHIGAN; GREAT LAKES.)

THE GROWTH OF THE METROPOLIS

Detroit was founded in 1701 by a French trader, Antoine de la Mothe Cadillac, who was seeking a strategic location for a new post to protect and encourage the French fur trade and to advance his own economic interests. On July 24 he came ashore near the present site of the Veterans Memorial Building with 50 soldiers, an equal number of voyageurs and settlers, and about 100 friendly Indians. A palisaded enclosure, about 200 feet square, was erected and named Fort-Pontchartrain du Détroit (of the strait) in honour of Cadillac's patron, comte de Pontchartrain, Louis XIV's minister of state. Later, the British called it simply Detroit. Despite his success, Cadillac was plagued by enemies in Quebec and Paris. To placate them, the King removed Cadillac from Detroit in 1710 and appointed him governor of Louisiana.

In the century after Cadillac's departure, Detroit slowly laid the foundation for its future growth despite the effects of three changes in control.

First, on November 29, 1760, as part of the treaty ending the French and Indian War, Detroit was surrendered without resistance to a British force under the command of a famous ranger, Maj. Robert Rogers. But a threat to British control developed shortly after, when some of France's Indian allies, notably the Ottawas under Chief Pontiac, organized a plot that was to culminate in the seizure of Detroit through a ruse. In May 1763 Chief Pontiac made his move, but Maj. Henry Gladwin, commandant, learned of the plan and foiled the conspiracy. A bitter siege followed, and, in the subsequent treaty of peace, France ceded its claims to the territory east of the Mississippi River. The British then came into undisputed possession of the area.

At the close of the American Revolution, Great Britain ceded the lands west of the Alleghenies to the United States but, in violation of the treaty, refused to withdraw troops from Detroit and various other posts. Since the British then had the friendship of the Indians, they kept stirring them up against Americans migrating to the West. After a series of campaigns against the Indians culminating in the Battle of Fallen Timbers, the Jay Treaty provided for the evacuation of the British posts. On July 11, 1796, Capt. Moses Porter arrived at Detroit with 65 soldiers to occupy the fort.

Capture by the British

There was to be one more transfer of control, one that distinguishes Detroit as the only major American city ever occupied for an extended period of time by a foreign foe. Soon after the outbreak of the War of 1812, Gen. William Hull surrendered the town to the British forces without a defense. The British held Detroit until late September, 1813, when Lieut. (later Commodore) Oliver H. Perry's victory in the Battle of Lake Erie (September 10) had made further occupation of the city impossible. Thus, the way was cleared to the continued development of the settlement that had been incorporated as a town in January 1802, with a board of trustees as the governing body.

In 1805 Michigan Territory (later to be Michigan, Wisconsin, and a part of Minnesota) was created by Congress and placed under a governor and judges. Detroit was designated the capital, and it remained the capital for ten years after Michigan became a state in 1837.

On June 11, 1805, the whole town was swept by a devastating fire. Every building except one was burned to the ground, but, miraculously, no lives were lost, and only two persons were injured. Each citizen whose home was burned was given a larger piece of land than he had previously held, and the city became less compact but, nevertheless, carefully planned. The governor, Augustus B. Woodward, produced a layout based on the design of a regular hexagon, with a round park in the centre and broad streets radiating from it. To accommodate later growth, other hexagons could be laid out adjoining the original one. As the city spread beyond its original boundaries, however, the plan was not extended, and years later only in the downtown area could one see vestiges of Woodward's dream.

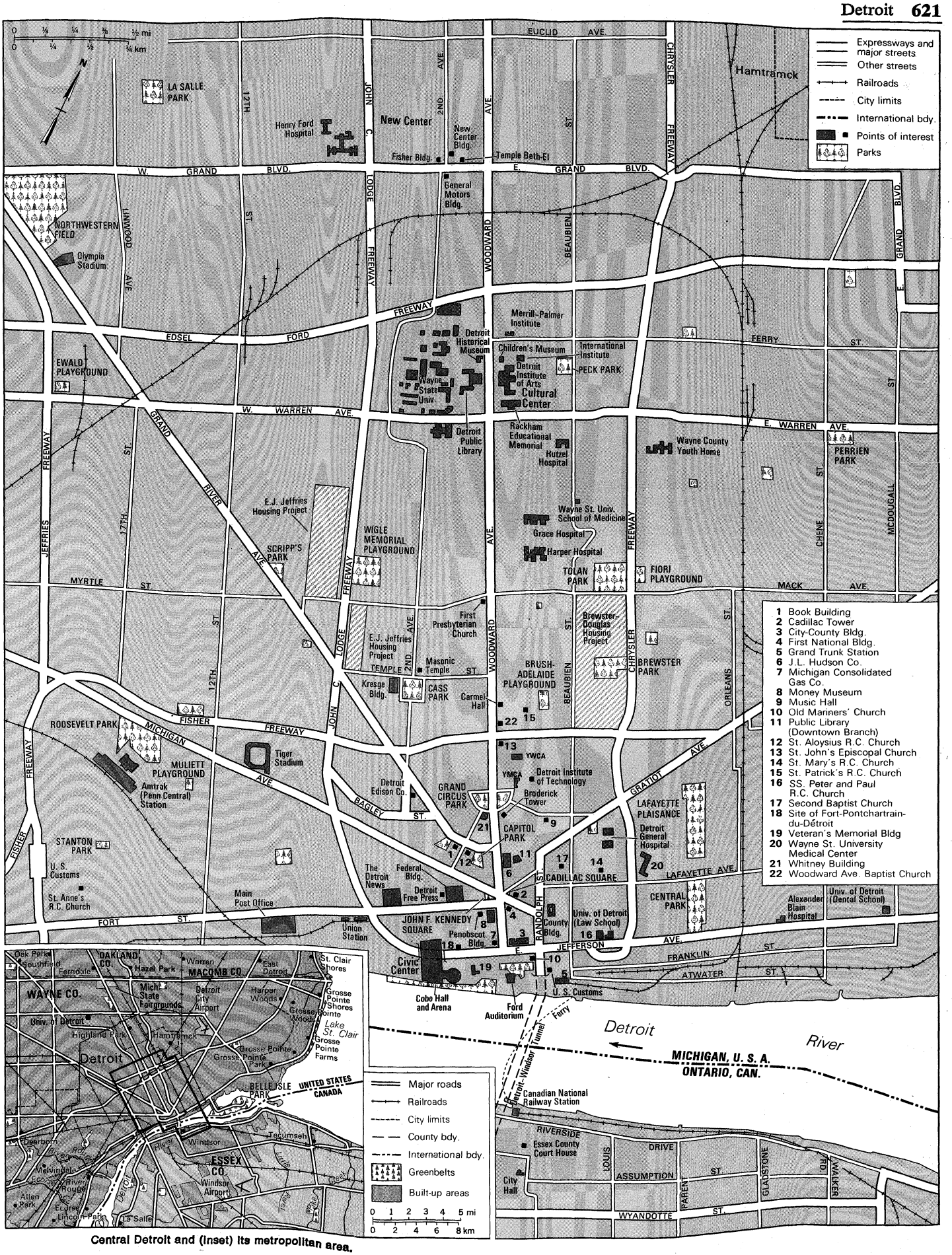
In the meantime, thought also was given to government. Michigan Territory officials revoked Detroit's 1802 charter for a new one in 1806, providing for a mayor appointed by the governor and an elected council of two houses with three members each.

By 1810, Detroit had a population of 1,650 and in 1815 was reincorporated as a city. A board of trustees, who chose their own president, became the governing body, and in 1824 a new charter provided for a mayor and a council of aldermen elected by wards. With slight variations, this pattern was to persist until a new charter in 1918 provided for a nonpartisan mayor and council of nine, the latter elected at large. This nonpartisan "strong mayor" form remained intact despite many challenges in the mid-1900s.

Early economic development

Those early years also helped set the pattern for economic and social development. In 1818 the "Walk-in-the-Water," the first steamboat on the upper Great Lakes, began regular runs between Buffalo and Detroit—and thus was dramatized the strategic location of the city. Soon thereafter other lines appeared, and shipbuilding and chandlering became profitable enterprises. Forwarding and commission houses flourished. By mid-century, although not the queen of the Great Lakes, Detroit was at least a prominent member of the queen's court. But, more importantly, Detroit became a centre for capital, provided by fortunes amassed in the lumber industry, and a centre for skilled workmen, due to the shipbuilding and manufacturing of machinery that developed there in the mid-19th century. This combination of money and trained manpower was the base on which Detroit depended when it assumed the leadership of the automobile industry.

The peak of water transportation was reached in the 1880s. Then, gradual decline set in until the St. Lawrence



Central Detroit and (Inset) its metropolitan area.

Seaway Project of the 1950s revived interest in the Port of Detroit. During the late 1960s overseas commerce exceeded 1,000,000 tons annually.

In the 1830s several railroads were incorporated, and construction started from Detroit. A decade later, a considerable number of miles of track stretched into the hinterland. In 1852 through-rail connection was made with Chicago when the Michigan Central completed its line. By railroad and water, grain and other produce poured into Detroit for reprocessing and forwarding to other parts of the nation or to Europe. By mid-century Detroit was one of the recognized flour-milling centres of the United States, with lively transactions on the Board of Trade.

In the 35 years following the U.S. Civil War, Detroit changed from its early role as a country merchant to that of industrial magnate. Iron and steel, foundry, railroad car, boot and shoe, stove, wheel and axle, chemical, and pharmaceutical industries all flourished by the end of the century—fortunately for those who were to move toward putting America on wheels.

Advent of the horseless carriage and emergence of Henry Ford

Charles B. King, a local businessman, drove a horseless carriage through the city streets in 1896, soon followed by a local engineer and would-be entrepreneur named Henry Ford; a new era dawned for Detroit. Its place as the automobile capital of the world was to be firmly established with the coming of the assembly line and the $5-a-day minimum wage, introduced by Ford (1914). What was a boom became an industrial explosion. Although Detroit did not become a one-industry city, the automobile industry inevitably put its stamp on every aspect of life in the city. It helped to spur the city's growth in population; it affected its economy—and that of the nation—as the average wage of auto workers moved from 76 cents per hour in 1936 to more than $5.23 per hour by 1972; and it was the target for organization by what became one of the largest industrial unions (the United Automobile Workers) in America.

In the mid-20th century, the city that achieved industrial renown in peace with its automobiles and in war with its production of armaments was also to become the city that was the stage for major developments in the civil rights revolution. This was not an entirely new role for the city, however, since before the U.S. Civil War Detroit was an important station on the "Underground Railroad," through which hundreds of runaway slaves passed to freedom into Canada.

Its industrial growth served as a magnet for many blacks, particularly as the demand for manpower grew at the time of World War II. Few realized it at the time, but events in Detroit were to foreshadow the urban problems of the future, when, in June 1943, a riot left 35 dead (29 of them black) and 1,000 injured before federal troops were called to help restore order.

Detroit was quick to establish a community relations commission, and it earned the reputation of being progressive in its handling of human relations problems. But in July 1967 rioting once again required the calling of federal troops. The death toll was 43, with $200,000,000 in property damage and 3,304 arrests. Incalculable were costs in bitterness and heartbreak that were to be long in being eradicated and repaired. A start was made with the establishment of a New Detroit Committee (New Detroit Inc.) on community relations.

THE CONTEMPORARY CITY

Population. In 1810 the population of Detroit was 1,650. By 1820 it had declined slightly to 1,442, partly because of the War of 1812. The opening of the Erie Canal, growing commerce on the Great Lakes, and development of the railroads helped increase the population to 21,019 by 1850. The importance of the city continued to draw increasing numbers of people, so that in 1860 the population had jumped to 45,516. In 1870, with a population of 79,577, it was the 18th largest U.S. city, and by 1900 the population had reached 285,704.

The development of the automobile and other industries after 1900 greatly stimulated Detroit's growth. By 1920, with a population of 993,678, it passed St. Louis to become the nation's fourth largest city, a rank it held until 1950, when it was passed by Los Angeles.

By 1970, while it maintained its rank, Detroit's population followed the trend of major American industrial cities with a dramatic drop. In 1970 the population was 1,511,482 (a decrease of 9.5 percent from 1960). Meanwhile, the standard metropolitan statistical area (Macomb, Oakland, and Wayne counties) increased to 4,199,931 in 1970 (an increase of 11.6 percent over 1960).

Detroit's ethnic background

From the early part of the 19th century, Detroit was a cosmopolitan community. French, Canadians, and Americans were prominent. In 1850, 47 percent of the population was of foreign birth. The Irish and the Germans predominated, comprising more than 50 percent of the foreign born. Although an increasing number of foreigners moved to Detroit before the more restrictive immigration laws of 1921 and 1924, the percentage of foreign born gradually declined, and the origin of the foreign born also changed. In 1880 those of German birth alone constituted 20.4 percent of the total population in the city, while Irish made up only 5.8 percent. By 1900, foreign born constituted about 34 percent of the population. Russians, Austrians, and Hungarians began arriving in considerable numbers between 1900 and 1910 and in 1910 made up 8 percent of the total population, ranking behind the Germans (9.6 percent) and the Canadians (9 percent). The greatest Polish immigration came in the next decade; by 1920 they constituted 6 percent of the population. In 1920 the foreign born were 29 percent of the whole population. The percentage dropped to about 26 percent in 1930, about 20 percent in 1940, about 17 percent in 1950, and 12 percent in 1960. By 1950 most of the foreign born were older people.

After the immigration laws of the 1920s, the great migration from the South set in and continued into the second half of the 20th century. In 1940 there were almost 200,000 blacks in Detroit, in 1960, 482,000, and by 1970 the city had become 43.7 percent black.

Administration and finance. Detroit has a mayor-council nonpartisan type of government. The charter adopted in 1918 provided for the nomination and election of the mayor, city clerk, city treasurer, council, and board of education, without party designations. The mayor has broad powers to appoint most of the administrative officials and commissions: the board of assessors, the board of health, the city planning, public lighting, water, street railway, fire, public welfare, and other commissions. A single commissioner appointed by the mayor, runs the police department.

The Common Council is composed of nine members elected at large. Eighteen are nominated in the primary, and the councilman with the most votes in the final vote automatically becomes president of the council and acting mayor in the absence of the mayor. The first nonpartisan government took office in January 1919. Originally, the officials were chosen for two-year terms, but in 1953 an amendment to the charter lengthened the terms to four years. The charter also provides for recall, and in 1930, at a special election, Mayor Charles Bowles was recalled because of alleged relations with the underworld and the Ku Klux Klan; it was the first time such action was taken against a mayor of a major U.S. city.

The chief source of revenue for Detroit is the real estate tax. To increase revenue, the council imposed an income tax, which in 1970 was 2 percent. Nonresidents working in the city pay ½ of 1 percent.

Housing and planning. In 1934, when the federal housing program began, Detroit sponsored slum clearance and low-cost housing projects, which continued into the second half of the 20th century.

After World War II, municipal planning included a new civic centre on a 58.3-acre riverfront site. Many old and unsightly buildings were demolished, and in 1950 the first of the new civic buildings was completed—the Veterans Memorial. Other buildings include a combined city–county building, replacing the former city hall and old courthouse; the Henry and Edsel Ford Auditorium, where the Detroit Symphony Orchestra plays; Cobo Hall, a convention and exhibit building; and Cobo Con-



Detroit's business centre, located along the Detroit River. In the background is Windsor, Ontario.
Charles E. Rotkin—P.F.I.

vention Arena, which can seat up to 12,000 persons for concerts and sports events.

Major buildings

Most of the city's largest office buildings, hotels, department stores, and banks are in the downtown area. The tallest is the 47-story Penobscot Building, and the largest office building outside downtown is the General Motors Building in the New Center area at Grand Boulevard and Second Avenue. Another large office building in the same neighbourhood is the Fisher Building. The General Motors, Fisher, and New Center buildings (a third but smaller office building) are connected by underground pedestrian tunnels. To accommodate the growing metropolitan area, many large, out-city shopping centres were built by private capital.

Transportation. By the early 1970s, the city had virtually completed an impressive network of expressways that provided relatively easy access to any area.

Dissatisfied with privately owned streetcar lines as early as the 1890s, Detroit in 1913 voted for a change in the city charter to allow for municipal ownership, but it was not until 1921 that the first such line was operated. A bitter feud between the city and the private lines ended when the city purchased the lines in 1922. In the late 1920s the city began operation of a few motor buses but again met with private competition, the Detroit Motor Bus Company, which it took over in 1932.

After World War II, patronage on the transit system began to fall, streetcars gradually were eliminated, and by 1956 the Detroit Street Railway had turned entirely to buses. The system operated as if it were privately owned, but by the late 1960s patronage had fallen sharply and deficits soared. In 1969 the citizens voted to allow a subsidy of the system from tax income.

Seven major railroad systems serve the city. Three airports—City, Metropolitan, and Willow Run—serve more than a dozen private, passenger and freight-carrying airlines.

The economy. Detroit is more than an automobile city; for about 1920 it became a steel centre, and the chemical business also prospered after the discovery of a layer of rock salt 34 feet thick, 1,000 feet below the surface. The Detroit Rock Salt Mine has approximately 60 miles of tunnels with fully mechanized equipment and transportation. In the second half of the 20th century Detroit ranked at or near the top in United States production of pharmaceuticals, adding machines and calculators, foundry products, and electrical household appliances. The Chrysler Corporation is the city's largest single employer.

Because of Detroit's location on the Great Lakes, the Michigan Customs District ranked third of U.S. ports in foreign trade at mid-20th century. Exports totalled about $700,000,000 a year. Imports included about 300 items used in automobile production. Detroit is a port of call for cargo boats and oceangoing freighters operating on the Great Lakes.

Education and culture. The public school system, formally organized in 1842, had more than 290,000 students in the second half of 20th century. Its facilities include trade schools as well as elementary, junior high, and high schools. Parochial and private schools had more than 60,000 students. Higher education is provided by Wayne State University and numerous sectarian, professional, technical, and junior colleges. The University of Michigan has a branch campus at Fairlane, the old Ford estate in Dearborn. Wayne State University, Eastern Michigan University, and the University of Michigan cooperate in an adult, noncredit educational program; and Wayne State and the University of Michigan operate the Institute of Labor and Industrial Relations, located on the Wayne campus.

Cultural societies

Detroit always has had a rich cultural life. Numerous cultural societies, such as the Lyceum, Mechanics' Society, Young Men's Society, Athenaeum, fraternal groups, musical, charitable, and religious societies, reading clubs, and scientific organizations, have contributed to the intellectual life of the citizens. There are two daily newspapers: the *Detroit Free Press* and *The Detroit News*. With the great influx of immigrants in the 70 years after 1850, many cultural and social societies of the various nationality groups were started, some of which still existed in the second half of the 20th century.

Particularly notable among Detroit's cultural institutions is the Cranbrook Academy of Art, established in 1927 by two wealthy philanthropists, George and Ellen Booth. Situated on the Cranbrook estate in the suburb of Bloomfield Hills, it houses an extensive collection of paintings and sculptures. Also on the Cranbrook estate is the Cranbrook Institute of Science, founded in 1930.

Bordering Wayne State University, the Detroit Cultural Center includes the main public library, the Detroit Institute of Arts, the Detroit Historical Museum, the Horace H. Rackham Educational Memorial, and the Detroit Public Schools Children's Museum. The municipally owned Institute of Arts houses paintings, sculptures, and decorative arts from prehistoric to modern times. The Detroit Historical Museum preserves the city's history in physicial survivals. In cooperation with the Detroit Historical Society, it also sponsors a large number of historical projects. The Children's Museum contains mobile and permanent exhibits to supplement the curriculum of the public schools. The Rackham Educational Memorial is the Detroit headquarters for the University of Michigan's extension services and is also the headquarters of the Engineering Society of Detroit. In suburban Dearborn, the Henry Ford Museum exhibits all forms of transportation equipment, while Greenfield Village contains reconstructions of many famous early American buildings and exhibits of early crafts.

Since 1914, Detroiters have maintained a symphony orchestra, which offers winter and summer seasons of concerts each year, in addition to a series of children's concerts. Evening band concerts are presented outdoors at Belle Isle Park during the summer. In 1928 the Detroit Civic Theatre was begun as a community venture. Its playhouse was later taken over by Wayne State University for drama projects. Because of the large number of foreign born, an International Institute has been organized; in addition to other activities, it maintains a program of folk music and dancing.

Detroit's churches

Detroit is also a city of churches, there being more than 1,000 of all denominations. St. Anne's Catholic Church is the oldest, dating from the settlement in 1701. The first Protestant society, interdenominational, was organized in 1816 and in 1825 became the First Presbyterian Church. All other main Protestant faiths are represented in Detroit. Temple Beth-El, the first Jewish society, was organized in 1850.

Parks and recreation. Detroit and the metropolitan area are well supplied with parks and recreational facilities. Among the larger ones that offer a variety of enter-

tainment are Belle Isle and Palmer parks. Belle Isle, a wooded, 1,000-acre island in the Detroit River, was acquired by the city in 1879. It has riding stables and bridle paths, a bathing beach, a children's zoo, a botanical conservatory, an aquarium, a museum devoted to the Great Lakes, facilities for boating, and picnic areas. Palmer Park, on the north side of the city, has no waterfront but offers ice-skating in season, golf, riding, tennis, archery, and picnic grounds. The Detroit Zoological Park, in Royal Oak, is among the nation's leading zoos, with cageless exhibits on a site of over 120 acres.

Spectator sports are popular, and Detroit is the home of several professional teams: the Tigers (baseball), Lions (football), Red Wings (ice hockey), and Pistons (basketball). Throughout the city and metropolitan area there are a large number of public and private golf courses. Bois Blanc (Bob-Lo) Island, on the Canadian side of the Detroit River, offers typical amusement park entertainment. Great Lakes cruises and various excursion trips are available during the summer months.

BIBLIOGRAPHY. Historical accounts of the city of Detroit include SILAS FARMER, *The History of Detroit and Michigan* (1884); CLARENCE M. BURTON, *The City of Detroit, Michigan, 1701–1922*, 5 vol. (1928); GEORGE W. STARK, *City of Destiny* (1943); ARTHUR POUND, *Detroit, Dynamic City* (1940); JOHN C. LODGE, *I Remember Detroit* (1949); GEORGE B. CATLIN, *The Story of Detroit* (1923); F. CLEVER BALD, *Detroit's First American Decade, 1796–1805* (1948); and FRANK B. and ARTHUR M. WOODFORD, *All Our Yesterdays: A Brief History of Detroit* (1969). Two studies of the race question in Detroit are JOHN C. LEGGETT, *Class, Race, and Labor: Working-Class Consciousness in Detroit* (1968); and LEONARD GORDON (comp.), *A City in Racial Crisis: The Case of Detroit Pre- and Post- the 1967 Riot* (1971). Also of interest are SIDNEY GLAZER, *Detroit: A Study in Urban Development* (1965); ALMON E. PARKINS, *The Historical Geography of Detroit* (1970); and J. DAVID GREENSTONE, *A Report on the Politics of Detroit* (1961).

(Fr.A.)

De Valera, Eamon

A leader in Ireland's fight for independence from England, Eamon De Valera headed his country's government with little interruption between 1932 and 1959, when he became president of Ireland, a post he held until 1973.

De Valera's command of public confidence owed little to popular appeal but rested upon his long record of austere integrity and patriotism and his sagacity as a political leader. His academic attainments and status inspired wide respect; he became chancellor of the National University of Ireland in 1921 and was founder of the Dublin Institute for Advanced Studies.

By courtesy of the Irish Embassy; photograph, Lensmen Ltd. Press Photo Agency, Dublin



De Valera, *c.* 1965.

Early life

De Valera was born in the United States, in New York City, on October 14, 1882. His father, a Spanish artist, having died, De Valera was sent as an infant to his mother's family in a labourer's cottage in County Limerick, Ireland. He was educated at the local national school and at Blackrock College, Dublin; he graduated from the Royal University and became a teacher of mathematics and an ardent supporter of the Irish-language revival. In 1913 he joined the Irish Volunteers, which had been organized to resist opposition to Home Rule for Ireland. In the Dublin anti-English uprising of Easter 1916, he commanded an occupied building and was the last commander to surrender. Because of his U.S. birth, he escaped execution but was sentenced to penal servitude.

Released in 1917 but arrested again and deported to England in May 1918, De Valera was acclaimed as the chief survivor of the uprising and elected president of the revolutionist Sinn Féin party, which won three-quarters of all the Irish constituencies in December 1918. After a dramatic escape from Lincoln Jail in February 1919, he went in disguise to the United States, where he collected funds. He returned to Ireland before military repression ended with the truce of 1921 and appointed plenipotentiaries to negotiate in London. He repudiated the treaty that they signed to form the Irish Free State, however, because it accepted the exclusion of Northern Ireland and imposed an oath of allegiance to the British crown.

When Dáil Éireann (the assembly of Ireland) ratified the treaty by a small majority, De Valera supported the republican resistance in the ensuing civil war. William Thomas Cosgrave's Irish Free State ministry imprisoned him; but he was released in 1924 and then organized a Republican opposition party that would not sit in the Dáil. In 1927, however, he persuaded his followers to sign the oath of allegiance as "an empty political formula," and his new Fianna Fáil party then entered the Dáil, demanding abolition of the oath of allegiance, of the governor general, and of the Seanad as then constituted and of land-purchase annuities payable to Great Britain. The Cosgrave ministry was defeated by Fianna Fáil in 1932, and De Valera, as head of the new ministry, embarked quickly on severing connections with Great Britain. He withheld payment of the land annuities, and an economic war resulted. Increasing retaliation by both sides enabled De Valera to develop his program of austere national self-sufficiency in an Irish-speaking Ireland, while building up industries behind protective tariffs. In 1937 the Free State declared itself a sovereign state, as Ireland or Éire, conceding voluntary allegiance to the British crown.

Head of government

De Valera's prestige was enhanced by his success as president of the Council of the League of Nations in 1932 and of its assembly in 1938. The menace of war in Europe induced Neville Chamberlain, in 1938, to conclude the "economic war" with mutual concessions. Great Britain relinquished the naval bases of Cobh, Berehaven, and Lough Swilly. In September 1939 De Valera proclaimed at once that Ireland would remain neutral and resist attack from any quarter. Besides avoiding the burdens and destruction of war, he had brought temporary prosperity, and he retained office after repeated elections.

In 1948 a reaction against the long monopoly of power and patronage held by De Valera's party enabled the opposition, with the help of smaller parties, to form an interparty government under John A. Costello. But this precarious coalition collapsed within three years, ironically, after declaring Ireland a republic by formal law, an act De Valera had avoided. De Valera resumed office until 1954, when he appealed unsuccessfully for a fresh mandate, and Costello formed his second interparty ministry. No clearly defined difference now existed between the opposing parties in face of rising prices, continued emigration, and a backward agriculture. But De Valera claimed that a strong single-party government was indispensable and that all coalitions must be weak and insecure. On this plea he obtained, in March 1957, the overall majority that he demanded. In 1959 De Valera agreed to stand as a candidate for the presidency. He resigned his position as *taoiseach* (head of government) and leader of the Fianna Fáil party. In June he

President of the republic of Ireland

was elected president and was re-elected in 1966. He retired to a nursing home near Dublin in 1973 and died there August 29, 1975.

De Valera's career spanned the dramatic period of Ireland's modern cultural and national resurgence. It was marred, however, by the deep political divisions spawned by the civil war of the 1920s, the persistence of the partition of the island that excluded six British-ruled counties from the republic, and the relative weakness of Ireland's economic position. As an anti-colonial leader, a skillful constitutionalist, and a symbol of national liberation, De Valera dominated Ireland in the half century following the country's independence. (D.R.G.)

BIBLIOGRAPHY. The biography by DENIS GWYNN, *De Valera* (1933), is concise, comprehensive, and impartial; the EARL OF LONGFORD and THOMAS P. O'NEILL, *Eamon De Valera* (1970), has direct interview material from De Valera; F.S.L. LYONS, *Ireland Since the Famine* (1971), puts the career of De Valera into context and includes the views of his critics; CONSTANTINE FITZGIBBON and GEORGE MORRISON, *The Life and Times of Eamon De Valera* (1974), has a brief and impartial text combining biography with a history of Ireland and is extensively and superbly illustrated. (Ed.)

Development, Animal

Development, in the context of this article, includes the processes that lead eventually to the formation of a new animal starting from cells derived from one or more parent individuals. Development thus occurs following the process by which a new generation of organisms is produced by the parent generation. This article is divided into the following major sections:

I. General features
  Reproduction and development
  Preparatory events
II. Early development
  Embryo formation
    Cleavage
    Gastrulation
  Embryonic adaptations
    Adaptations in animals other than mammals
    Adaptations in mammals
III. Organ formation
  Primary organ rudiments
  Organogenesis and histogenesis
IV. Ectodermal derivatives
  The nervous system
    The brain and spinal cord
    Major sense organs
  The epidermis and its outgrowths
V. Mesodermal derivatives
  The body muscles and axial skeleton
  The appendages: tail and limbs
  Excretory organs
  Circulatory organs
  Reproductive organs
VI. Endodermal derivatives
  The alimentary canal
  The pharynx and its outgrowths
  The liver, pancreas, and lungs
VII. Postembryonic development
  The larval phase and metamorphosis
  Direct development
VIII. Maturity and death

I. General features

REPRODUCTION AND DEVELOPMENT

In multicellular animals (Metazoa), reproduction takes one of two essentially different forms: sexual and asexual. In asexual reproduction the new individual is derived from a blastema, a group of cells from the parent body, sometimes, as in *Hydra* and other coelenterates, in the form of a "bud" on the body surface. In sponges and bryozoans, the cell groups from which new individuals develop are formed internally and may be surrounded by protective shells; these bodies, which may serve as resistant forms capable of withstanding unfavourable environmental conditions, are released after the death of the parent. In certain animals the parent may split in half, as in some worms, in which an individual worm breaks into two fairly equal parts (except that the anterior half receives the mouth, "brain," and sense organs if they are present).

Obviously, in such a case it is impossible to say which of the two resulting individuals is the parent and which the offspring. Some brittle stars (starfish relatives) may reproduce by breaking across the middle of the body disk, with each of the halves subsequently growing its missing half and the corresponding arms.

A common feature of all forms of asexual reproduction is that the cells—always a substantial number of cells, never only one cell—taking part in the formation of the new individual are not essentially different from other body, or somatic, cells. The number of chromosomes (bodies carrying the hereditary material) in the cells participating in the formation of a blastema is the same as in the other somatic cells of the parent, constituting a normal, double, or diploid (2n), set.

In sexual reproduction, a new individual is produced not by somatic cells of the parent but by sex cells, or gametes, which differ essentially from somatic cells in having undergone meiosis, a process in which the number of chromosomes is reduced to one-half of the diploid (2n) number found in somatic cells; cells containing one set of chromosomes are said to be haploid (n). The resulting sex cells thus receive only half the number of chromosomes present in the somatic cell. Furthermore, the sex cells are generally capable of developing into a new individual only after two have united in a process called fertilization (*q.v.*).

Advantages of sexual and asexual reproduction

Each type of reproduction—asexual and sexual—has advantages for the species. Asexual reproduction is, at least in some cases, the faster process, leading most rapidly to the development of large numbers of individuals. Males and females are independently capable of producing offspring. The large size of the original mass of living matter and its high degree of organization—the new individual inherits parts of the body of the parent: a part of the alimentary canal, for instance—make subsequent development more simple, and the attainment of a stage capable of self-support easier. New individuals produced by asexual reproduction have the same genetic constitution (genotype) as their parent and constitute what is called a clone. Though asexual reproduction is advantageous in that, if the parent animal is well adapted to its environment and the latter is stable, then all offspring will benefit, it is disadvantageous in that the fixed genotype not only makes any change in offspring impossible, should the environment change, but also prevents the acquisition of new characteristics, as part of an evolutionary process. Sexual reproduction, on the other hand, provides possibilities for variation among offspring and thus assists evolution by allowing new pairs of genes to combine in offspring. Since all body cells are derived from the fertilized egg cell, a mutation, or change, occurring in the sex cells of the parents immediately provides a new genotype in each cell of the offspring. In the course of evolution, sexual reproduction has been selected for, and established in, all main lines of organisms; asexual reproduction is found only in special cases and restricted groups of organisms.

PREPARATORY EVENTS

In the case of multicellular animals we find there are two kinds of sex cells: the female sex cell (ovum, or egg), derived from an oocyte (immature egg), and the male sex cell (spermatozoon or sperm), derived from a spermatocyte. Eggs are produced in ovaries; sperm, in testes. Both the egg and the sperm contribute to the development of the new individual; each providing one set of genes, thereby restoring the diploid number of chromosomes in the fertilized egg. The sperm possesses a whiplike tail (flagellum) that enables it to swim to the egg to fertilize it. In most cases the egg, a stationary, spherical cell, provides the potential offspring with a store of food materials, or yolk, for its early development (Figure 1). The term yolk does not refer to any particular substance but in fact includes proteins, phosphoproteins, lipids, cholesterol, and fats, all of which substances occur in various proportions in the eggs of different animals. In addition to yolk, eggs accumulate other components and acquire the structure necessary for the development of the new

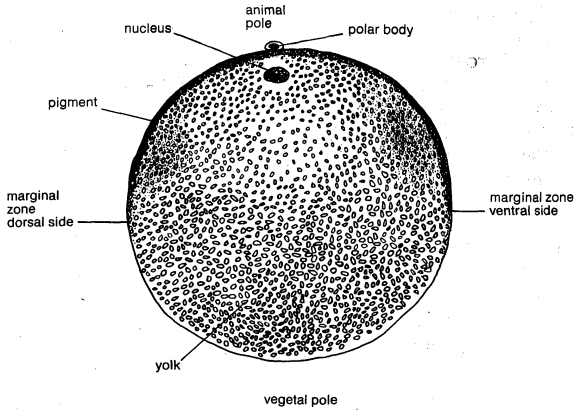


Figure 1: Amphibian egg, vertical section, showing distribution of yolk and pigment in interior.

individual. In particular the egg acquires polarity—that is, the two ends, or poles, of the egg become distinctive from each other. At one pole, known as the animal pole, the cytoplasm appears to be more active and contains the nucleus (meiotic divisions occur in this region); at the other, called the vegetal pole, the cytoplasm is less active and contains most of the yolk. The general organization of the future animal is closely related to the polarity of the egg.

When the amount of food reserve is comparatively small, as it is in many marine invertebrates and mammals (in the latter the embryo is nourished by materials in the mother's blood), the egg may be barely visible to the unaided eye. The egg of the sea urchin is about 75 microns (0.003 inch) in diameter; that of a human being is slightly more than 0.1 millimetre. Eggs are classified according to the amount of yolk present. An egg with a small quantity of evenly distributed yolk is called an oligolecithal egg. One with more yolk that is unevenly distributed (*i.e.*, concentrated towards the vegetal pole) is telolecithal; and one with still greater amounts of yolk in granules or in a compact mass is megalecithal.

Protective membranes

The egg is surrounded by protective membranes, which may be soft and jellylike or hard and calcified, like shells. Egg membranes are produced while the egg is either in the ovary or being carried away from the ovary in a tube called an oviduct. The eggs of many animals have both kinds of membranes. In insects, a hard shell (chorion) forms around the eggs in the ovaries. In frogs, a very thin vitelline membrane forms around the eggs in the ovary; subsequently a layer of jelly is deposited around the eggs while they pass through the oviducts. In birds, a very thin vitelline membrane is produced around the egg in the ovary; then several layers of secondary membranes are formed in the oviduct before the egg is laid. The outermost of these secondary membranes is the calcareous shell. In mammals the egg is surrounded by the so-called pellucid zone, which is equivalent to the vitelline membrane of other animals; follicle cells form an area called the corona radiata around this zone.

After fertilization the egg, now called a zygote, is endowed with genes from two parents and has begun actual development. (Activation of the egg may be brought about by an agent other than sperm in certain animals, but such cases of parthenogenesis are exceptional. See the articles on FERTILIZATION; and REPRODUCTION.)

After fertilization, the zygote undergoes a series of transformations that bring it closer to the essential organization of the parents. These transformations, initiated at a physiological, perhaps even at a molecular, level, eventually result in the appearance of certain structures. The whole process is called morphogenesis (Greek *morphē*, "shape" or "form"; *genesis*, "origin" or "production"). The process of development is more easily understood if, at every step, the changes necessary to bring the system nearer the goal are considered. Depending on the achievements necessary at any step, development can be subdivided into a number of discrete phases, the first of which, cleavage, immediately follows fertilization.

II. Early development

EMBRYO FORMATION

Cleavage. Since the goal of development is the production of a multicellular organism, many cells must be produced from the single-celled zygote. This task is accomplished by cleavage, a series of consecutive cell divisions. Cells produced during cleavage are called blastomeres. The divisions are mitotic—*i.e.*, each chromosome in the nucleus splits into two daughter chromosomes, so that the two daughter blastomeres retain the diploid number of chromosomes. During cleavage, almost no growth occurs between consecutive divisions, and the total volume of living matter does not change substantially; as a consequence, the size of the cells is reduced by almost half at each division. At the beginning of cleavage, cell divisions tend to occur at the same time in all blastomeres, and the number of cells is doubled at each division. As cleavage progresses, the cells no longer divide at the same time.

Cleavage in most animals follows an orderly pattern, with the first division being in the plane of the main axis of the egg (Figure 2). This cleavage plane is arbitrarily called vertical, on the assumption that the main axis of the egg is vertical. The second cleavage plane is again vertical but at right angles to the first, giving rise to four equal cells arranged around the main axis of the egg (Figure 2). The third cleavage plane is at right angles to both the first and second cleavage planes and is horizontal, or equatorial. Subsequent divisions may alternate between vertical and horizontal cleavage planes, but later cleavage divisions become randomly oriented. This pattern is typical of many animal groups; however, more complicated patterns of cleavage are found in such animals as annelids, mollusks, and nematodes.

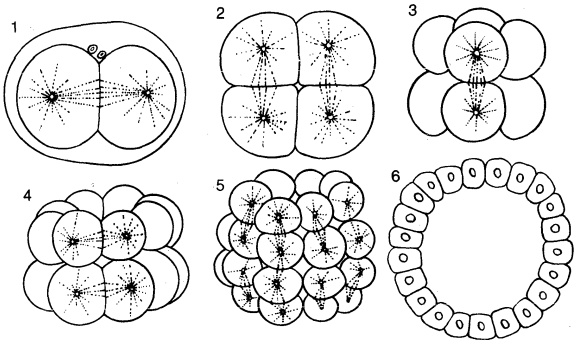


Figure 2: Cleavage of an oligolecithal egg (sea cucumber) in progressive stages 1 through 6 (see text).

Influence of yolk on cleavage

As the amount of yolk in the egg increases, it influences cleavage by hindering the cytoplasmic movements involved in mitosis. If there is only little yolk (oligolecithal eggs), the yolk granules follow the movements of the cytoplasm and are distributed in the resulting blastomeres. But if the amount of yolk is larger (megalecithal eggs), cleavages occur nearer the animal pole, where there is less yolk; as a result, the blastomeres nearer the animal pole are smaller than those nearer the vegetal pole. The presence of yolk masses may retard the onset of cleavage in a part of the egg or even suppress it altogether; in this case cleavage is partial, or meroblastic. Only a part of the egg material then is subdivided into cells, the rest remaining as a mass that serves as nourishment for the developing embryo.

Cleavage is complete, or holoblastic, in many invertebrates including coelenterates, annelids, echinoderms, tunicates, and cephalochordates. The blastomeres may be either about equal or only slightly different in size. Cleavage in amphibians is holoblastic, but the size of the blastomeres is very uneven. Blastomeres are smallest at the animal pole and largest (and yolky) at the vegetal pole. Somewhat similar conditions prevail in many mollusks. In most fishes, birds, reptiles, and egg-laying mammals (monotremes), cleavage is discoidal—*i.e.*, restricted

to a disk of cytoplasm at the animal pole of the egg, most of the yolky egg material remaining uncleaved (Figure 3). Cleavage in insects and many other arthropods is superficial—*i.e.*, the entire surface layer of egg cytoplasm subdivides into cells, and the egg contains a central mass of uncleaved yolk. The conditions of cleavage in placental mammals, including man, are peculiar.

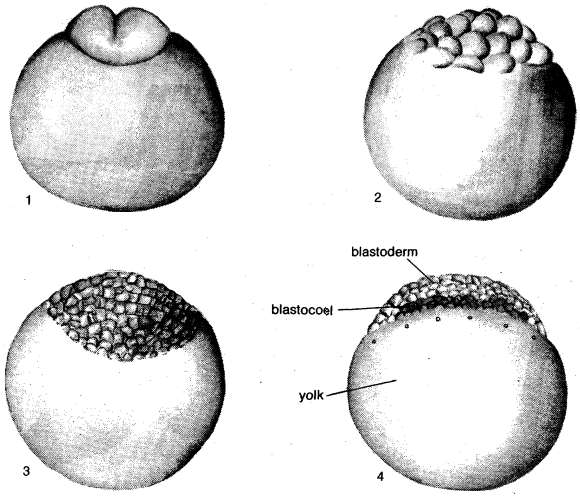


Figure 3: Discoidal cleavage of an egg of a bony fish.

During cleavage, development involves only an increase of cell numbers; the shape of the embryo does not change, and chemical transformations within the embryo are restricted to those necessary for cell division. Chemical and structural transformations are concerned with accumulating chromosomal material in the nuclei of the blastomeres. Before each division the chromosomes carrying the genes double in number; this means that the chromosomal material, deoxyribonucleic acid (DNA), has to be synthesized. This synthesis proceeds possibly at the expense of cytoplasmic ribonucleic acid (RNA) but certainly also from simpler organic compounds. A certain amount of protein synthesis is also necessary for cleavage to proceed: if developing eggs are treated with puromycin, a substance which is known to suppress protein synthesis, cleavage stops immediately. The proteins concerned have not yet been identified. No proteins are synthesized, however, that would foreshadow the future differentiation of parts of the embryo. It is believed that the genes in the chromosomes remain largely inactive during cleavage. The rhythm (speed) of cleavage is wholly dependent on the cytoplasm of the egg.

Although the shape and volume of the embryo do not change during cleavage, one important change in gross organization does take place. As the blastomeres are produced, they move outward, leaving a centrally located fluid-filled cavity. In cases of holoblastic cleavage, the blastomeres become arranged in a layer from one to several cells thick surrounding the cavity. The embryo at this stage may be likened to a hollow ball and is known as a blastula (Figure 4:1). The outer layer of cells is called the blastoderm, and the fluid-filled cavity the blastocoel. In discoidal cleavage the cells, which do not surround the whole embryo, lie only on the animal pole; nevertheless, a blastocoel may be formed by a crevice appearing between the blastomeres and the mass of yolk (Figure 3). The blastomeres then may be arranged as a saucer-shaped blastodisk covering the blastocoel.

The formation of the blastula signifies the end of the period of cleavage. The next stage of development is concerned not with an increase in cell number, though cell divisions continue at a slower pace, but with rearrangement of the available cell masses to conform with the gross features of the future animal.

Gastrulation. The embryo in the blastula stage must go through profound transformations before it can approach adult organization. An adult multicellular animal typically possesses a concentric arrangement of tissues of the body; this feature is common to all animal groups above the level of the sponges. Adult tissues are derived from three embryonic cell layers called germinal layers: the outer layer is the ectoderm, the middle layer is the mesoderm, and the innermost layer is the endoderm (entoderm). The ectoderm gives rise to the skin covering, to the nervous system, and to the sense organs. The mesoderm produces the muscles, excretory organs, circulatory organs, sex organs (gonads), and internal skeleton. The endoderm lines the alimentary canal and gives rise to the organs associated with digestion and, in chordates, with breathing.

Germinal layers

The blastula, which consists of only one cell layer, undergoes a dramatic reshuffling of blastomeres preparatory to the development of the various organ systems of the animal's body. This is achieved by the process of gastrulation, which is essentially a shifting or moving of the cell material of the embryo so that the three germinal layers are aligned in their correct positions.

The rearrangement of the blastula to form the germinal layers is seen clearly in certain marine animals with oligolecithal eggs. The hollow blastula consists of a simple epithelial layer (the blastoderm), the transformation of which can be likened to the pushing in of one side of a rubber ball (Figure 4). As a result of such inpushing (or invagination), the spherical embryo is converted into a double-walled cup, the opening of which represents the position of the former vegetal pole. The involuted part of the blastoderm, lining the inside of the double-walled cup, gives rise to the endoderm and mesoderm, and the blastomeres remaining on the exterior become the ectoderm. As a consequence of the infolding at the vegetal pole, the blastocoel is reduced or obliterated, and a new cavity is created, the primitive gut cavity, or archenteron, which eventually gives rise to the hollow core (lumen) of the alimentary canal. At this stage the embryo has a primitive gut with an opening to the exterior and is known as a gastrula. The opening of the gastrula is the blastopore, or primitive mouth; both terms are somewhat misleading. It would seem that the term blastopore should be applied more appropriately to an opening in a blastula, in which, of course, no opening exists. As to the term primitive mouth, it must be pointed out that the blastopore does not always give rise to the adult mouth. In certain animal groups it becomes the anus, and a mouth forms as a completely new opening.

In some coelenterates, cells at the vegetal pole do not form an invaginating pocket, but individual cells slide inward, losing connection with other cells of the blastoderm. Eventually these cells fill the blastocoel and form a compact mass of endoderm. The cavity (archenteron) within this mass and the opening (blastopore) to the exterior are then produced secondarily by the separation of these cells.

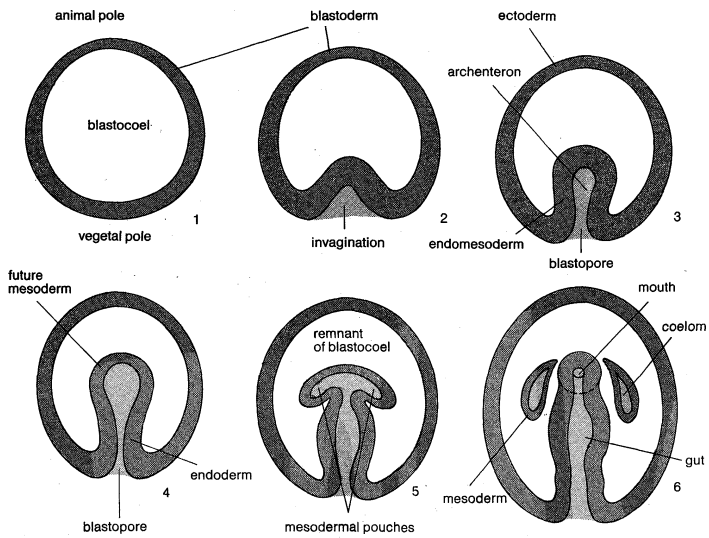


Figure 4: Gastrulation and mesoderm formation in the embryo of a starfish.

Amphioxus, echinoderms, and amphibians. Gastrulation does not always proceed exactly as described above. In the course of evolution, certain animal groups have modified this critical stage of embryonic development, and these modifications have undoubtedly contributed to the successful continuation of species. In the primitive fishlike chordate amphioxus, for example, the invaginating blastoderm eventually comes into close contact with the inner surface of the ectoderm, thus practically squeezing the blastocoel out of existence or at least reducing it to a narrow crevice between the ectoderm and the endomesoderm. In echinoderms, on the other hand, a smaller portion of the blastoderm invaginates, and the blastocoel remains as a spacious internal cavity between the ectoderm and the endomesoderm. It persists as the primary body cavity and is the only body cavity (apart from the cavity of the alimentary canal) in such invertebrates as nematodes and rotifers.

In the double-walled-cup stage, the two internal germinal layers—endoderm and mesoderm—may not yet be distinct. Their separation may occur later, in the second phase of gastrulation, by one of two methods. One is the development of outpocketings from the wall of the archenteron. In starfishes and other echinoderms, the deep part of the endomesodermal invagination forms two thin-walled sacs, one on each side of the gastrula (Figure 4). These are the rudiments of the mesoderm; the remaining part of the archenteron becomes the endoderm and produces the lining of the gut. The cavities within the mesodermal sacs expand to become the coelom, the secondary body cavity of the animal. A somewhat similar process of mesoderm and coelom development occurs in amphioxus among the chordates, except that a series of mesodermal sacs forms on either side of the embryo, foreshadowing the segmented (metameric) structure common to chordates. Only the most anterior pairs of the mesodermal sacs actually contain a cavity at the time of their formation; the more posterior ones are solid masses of cells separating from the archenteric wall and from one another and developing coelomic cavities later.

Gastrulation in starfishes

A second method of mesoderm formation is by the splitting off of mesodermal cells from the original common mass of endomesoderm. This may take the form of single cells detaching themselves from the archenteron or of whole sheets of cells splitting off from the endoderm. An example of the latter type is seen in the gastrulation of amphibians. The development of specific regions of the early amphibian embryo—by the use of natural pigmentation or artificially introduced dyes—can be followed and their location in the adult recorded in diagrams called fate maps. The fate map of a frog blastula, just prior to gastrulation demonstrates that the materials for the various organs of the embryo are not yet in the position corresponding to that in which the organs will lie in a fully developed animal. The endodermal material for the foregut, for example, lies not far from the vegetal pole; the ectodermal component of the mouth region (stomodeum) is situated close to the animal pole. Extensive rearrangement of the embryo is necessary to bring all the parts into their correct relationships.

Because of the large amount of yolk and resulting uneven cleavage, gastrulation in amphibians cannot proceed by a simple infolding of the vegetal hemisphere. A certain amount of invagination does take place, assisted by an active spreading of the animal hemisphere of the embryo; as a result, the ectoderm covers the endodermal and mesodermal areas. The spreading is sometimes described as an "overgrowth"—an inappropriate term, since no growth or increase of mass is involved. The future ectoderm simply thins out, expands, and covers a greater surface of the embryo in a movement known as epiboly.

Gastrulation in amphibians, in lungfishes, and in the cyclostomes (hagfishes and lampreys) begins with the formation of a pit on what will become the back (dorsal) side of the embryo (Figure 5). The pit represents the active shifting inward of the cells of the blastoderm. As these cells undergo a change in shape, there occurs also a contraction at the external surface, with adjacent cells being drawn toward the centre of the contraction even before an actual depression is formed. The cells most concerned in this process will become part of the future foregut. Further movement of the cells inward results in the formation of a distinct pit, which rapidly develops into a pocket-like archenteron with its opening, the blastopore. Once the archenteron is formed, more and more of the exterior cells roll over the edge of the blastopore and disappear into the interior. In the course of gastrulation the shape of the blastopore changes from a simple pit to a transverse slit and finally into a groove encircling the yolky material at the vegetal pole. As a result of epiboly of the animal hemisphere, the upper edge of the groove is gradually pushed down until the yolky cells of the vegetal pole are covered completely. The edges of the blastopore then converge toward the vegetal pole, the slit between them being eventually reduced to a narrow canal, which lies at the posterior end of the embryo and, in some species, becomes the anal opening. (In other cases the canal closes, and a new anal opening breaks through nearby, slightly more ventrally.)

Formation of archenteron in amphibians

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

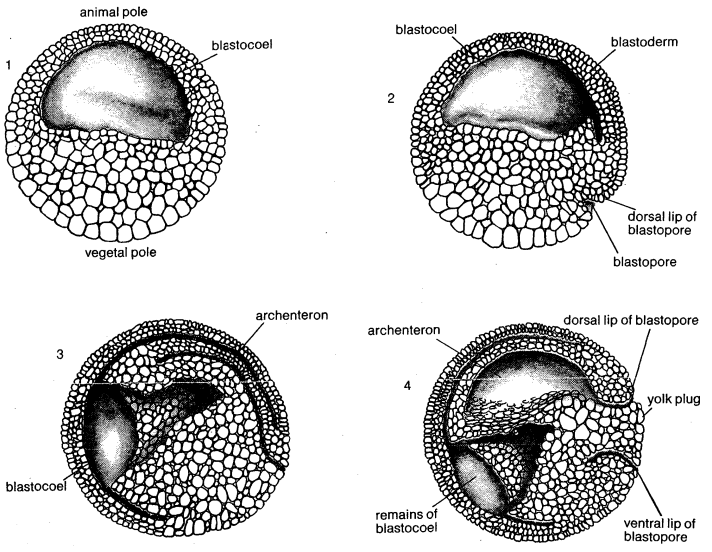


Figure 5: Gastrulation in the frog embryo. Embryos, in four progressive stages, are represented as cut in the median plane.

The cavity of the archenteron increases as more material from the outside is transferred inward, and the blastocoel becomes almost completely obliterated. Both mesoderm and endoderm are shifted into the interior, and only the ectoderm remains on the embryo surface. The mesoderm splits from the endoderm: the endoderm lines the archenteric cavity (and eventually becomes the lining of the alimentary canal), as the mesoderm surrounds the endoderm to form the chordamesodermal mantle. By the time the blastopore closes, the three germ layers are in their correct spatial relationship to each other.

Reptiles, birds, and mammals. Although amphibian gastrulation is considerably modified in comparison with that in animals with oligolecithal eggs (*e.g.*, amphioxus and starfishes), an archenteron forms by a process of invagination. Such is not the case, however, in the higher vertebrates that possess eggs with enormous amounts of yolk, as do the reptiles, birds, and egg-laying mammals. Cleavage in these animals is partial (meroblastic), and, at its conclusion, the embryo consists of a disk-shaped group of cells lying on top of a mass of yolk. This cell group often splits into an upper layer, the epiblast, and a lower layer, the hypoblast. These layers do not represent ectoderm and endoderm, respectively, since almost all the cells that form the embryo are contained in the epiblast. Future mesodermal and endodermal cells sink down into the interior, leaving only the ectodermal ma-

terial at the surface. In reptiles, egg-laying mammals, and some birds, a pocket-like depression occurs in the epiblast but encompasses only chordamesoderm or even only the notochord. Individual cells of the remainder of the mesoderm and endoderm migrate into the interior and there arrange themselves into a sheet of chordamesoderm and of endoderm, the latter of which mingles with cells of the hypoblast if such a layer is present. The migration of the cells destined to form mesoderm and endoderm does not take place over the whole surface of the disk-shaped embryo but is restricted to a specific area along the midline. This area is more or less oval in reptiles and lower mammals; distinctly elongated in higher mammals and birds, it is called the primitive streak (Figure 6), a thickened and slightly depressed part of the epiblast that is thickest at the anterior end, called the Hensen's node.

In animals having discoidal cleavage, the three germinal layers at the end of gastrulation are stacked flat; ectoderm on top, mesoderm in the middle, and endoderm at the bottom. The embryo is produced from the flattened layers by a process of folding to form a system of concentric tubes (Figure 7). The edges of the germ layers, which are not involved in the folding process, remain attached to the yolk and become the extra-embryonic parts; they are not directly involved in supplying cells for the embryo but break down yolk and transport it to the developing embryo.

Gastrulation in higher mammals

Higher mammals—apart from the egg-laying mammals—do not have yolk in their eggs but, having passed through an evolutionary stage of animals with yolky eggs, retain, particularly in gastrulation, features common to reptiles (and birds, which also had reptilian ancestors). As a result, at the end of cleavage the formative cells of the embryo—the cells that will actually build the body of the animal—are arranged in the form of a disk over a cavity that takes the place of the yolk of the reptilian ancestors of mammals. Within the disk of cells a primitive streak develops, and the three germinal layers are formed much as in many reptiles and birds.

Gastrulation and the formation of the three germinal layers is the beginning of the subdivision of the mass of embryonic cells produced by cleavage. The cells then begin to change and diversify under the direction of the genes. The genes brought in by the sperm exert control for the first time; during cleavage all processes seem to be under control of the maternal genes. In cases of hybridization, in which individuals from different species produce offspring, the influence of the sperm is first apparent at gastrulation: paternal characteristics may appear at this stage; or the embryo may stop developing and die if the paternal genes are incompatible with the egg (as is the case in hybridization between species distantly related).

The diversification of cells in the embryo progresses rapidly during and after gastrulation. The visible effect is that the germinal layers become further subdivided into aggregations of cells that assume the rudimentary form of various organs and organ systems of the embryo. Thus the period of gastrulation is followed by the period of organ formation, or organogenesis.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

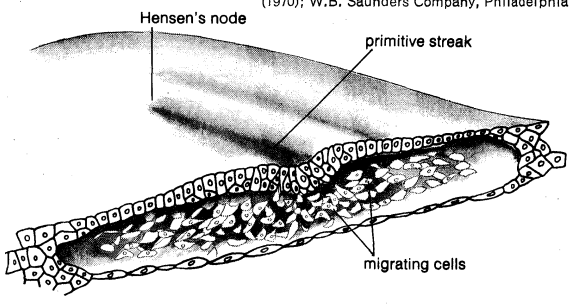


Figure 6: Anterior half of blastoderm of a bird embryo cut transversely to show migration of mesodermal and endodermal cells from the primitive streak.

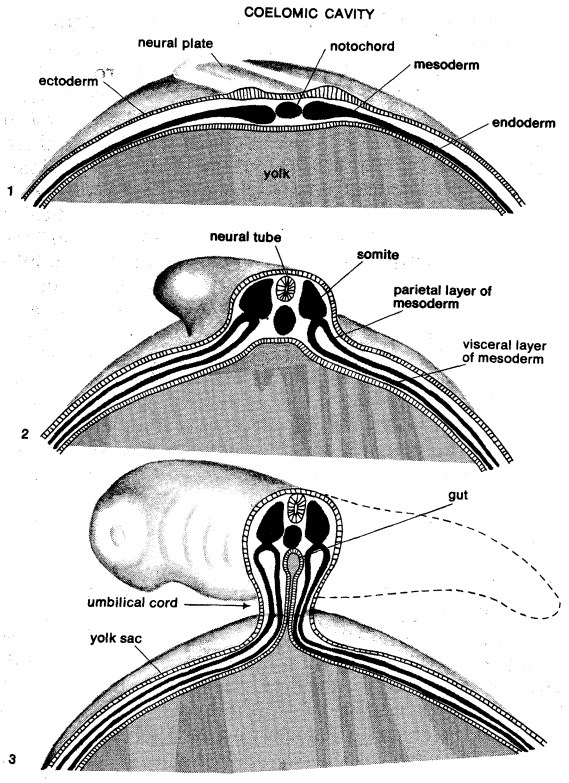


Figure 7: Vertebrate body formation from three germinal layers showing separation from yolk sac after discoidal cleavage.

EMBRYONIC ADAPTATIONS

Throughout its development the embryo requires a steady supply of nourishment and oxygen and a means for disposal of wastes. These needs are met in various ways, depending in particular on (1) whether the eggs develop externally (oviparity), are retained in the maternal body until ready to hatch (ovoviviparity), or are carried in the maternal body to a later stage (viviparity); and (2) the length of embryonic development.

Adaptations in animals other than mammals. Eggs of many marine invertebrates are discharged directly into water, and the period of development before the larva emerges is relatively brief. Oxygen diffuses easily into the small eggs, and nourishment is provided by a moderate amount of yolk. During cleavage the yolk is distributed to all the blastomeres. Much of the nourishment in the egg is stored as animal starch, or glycogen, which is almost completely used by the time the larva emerges from the egg. A small amount of water and inorganic salts are taken in by the embryo from surrounding seawater. Eggs developing in freshwater carry their own supply of necessary amounts of certain salts that are not present in sufficient quantities in the environment. Products of metabolism—especially carbon dioxide and nitrogenous wastes in the form of ammonia—diffuse out from small embryos developing in water.

Mechanisms to overcome drying

The eggs of terrestrial animals must overcome the hazard of drying. In certain species this danger is avoided because the animal returns to water to breed, as frogs and salamanders. Some groups of insects (*e.g.*, dragonflies, mayflies, and mosquitoes) also lay eggs in water, and the larvae are aquatic. Eggs of other animals (*e.g.*, snails, earthworms) are laid in moist earth and thus are protected from drying up. In terms of evolution, however, a decisive solution to the problem of development on land was arrived at by most insects and by reptiles and birds, which developed eggs with a shell impermeable to water or, at least, resistant to rapid evaporation. The shells of bird and insect eggs, while restricting evaporation of water, allow oxygen to diffuse into the egg and carbon dioxide to diffuse out. Apart from gas exchange, the eggs constitute closed systems, which give nothing to the outside and require nothing from it. Such

eggs are called cleidoic. Because the products of nitrogen metabolism in cleidoic eggs cannot pass through the eggshell, animals (birds and insects) have had to evolve a method of storing wastes in the form of uric acid, which, since it is insoluble, is nontoxic to the embryo.

After a short period of development in the egg, the emerging young animal has to fend for itself, unless there is some form of parental care. Exposure to the external environment at a tender age results frequently in loss of life, a hazard met by many animals through an increase in the supply of nourishment within the egg, thus allowing the young to attain a greater size and development. This tendency to produce large yolky eggs has been achieved independently in different evolutionary lines: in octopuses and squids among the mollusks, in sharks among the fishes, and in reptiles and birds among the terrestrial vertebrates.

As has been indicated, cleavage is incomplete in eggs with large amounts of yolk. Although some yolk platelets may be enclosed in the formative cells of the embryo, the bulk of the yolk remains an uncleaved mass, overgrown and surrounded by the cellular part of the embryo. In such cases a membranous bag, or yolk sac, is formed (Figure 8) and remains connected to the embryo by a narrow stalk (the evolutionary precursor of the umbilical cord of mammals). The cellular layers surrounding the yolk sac and forming its walls may consist of all three germinal layers (in reptiles and birds), so that the yolk virtually comes to lie inside an extension of the gut of the embryo; or (in bony fishes) the yolk sac may be enclosed in layers of ectoderm and mesoderm. In either case a network of blood vessels develops in the walls of the yolk sac and transports the yolk products to the embryo. As the yolk is broken down and utilized, the yolk sac shrinks and is eventually drawn into the body of the embryo. In addition to the yolk sac, extra-embryonic parts are also encountered in the form of embryonic membranes, which are found in higher vertebrates and in insects. Vertebrates have three embryonic membranes: the amnion, the chorion, and the allantois.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

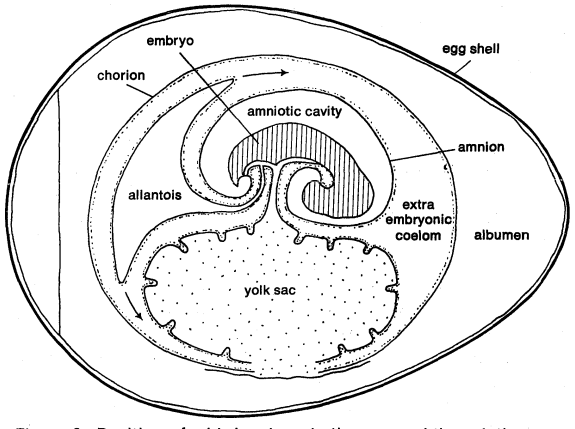


Figure 8: Position of a bird embryo in the egg and the relation of the allantois to the amniotic cavity and the yolk sac.

In reptiles, birds, and mammals, folds develop on the surface of the yolk sac just outside and around the body of the embryo proper. These folds, consisting of extra-embryonic ectoderm and extra-embryonic mesoderm, rise up and fuse dorsally, enclosing the embryo in a double-lined, fluid-filled chamber known as the amniotic cavity. The inner lining of the fold becomes the amnion, and the outer becomes the chorion, which ultimately surrounds the entire embryo. The amniotic fluid protects the embryo from drying, prevents the adhesion of the embryo to the inner surface of the shell, and provides the embryo with efficient shock absorption against possible damaging jolts. (The amnion and chorion develop in the same way in insect embryos.) The third membrane, or allantois, is originally nothing more than the urinary bladder of the embryo. It is a saclike growth of the floor of the gut, into which nitrogenous wastes of the embryo are voided. It enlarges greatly during the course of development, eventually expanding between the amnion and chorion and also between the chorion and the yolk sac, to become the third embryonic membrane. In addition to providing storage space for the nitrogenous wastes of the embryo, the allantois takes up oxygen, which penetrates into the egg from the exterior, and delivers it, by way of a network of blood vessels, to the embryo.

The allantois

Adaptations in mammals. At some early stage during the evolution of viviparous mammals, eggs came to be retained in the oviducts of the mother. The embryo then was provided with nourishment from fluids in the oviduct; the yolk, which became redundant, gradually ceased to be provided, and the eggs became oligolecithal. The eggshell, present in reptiles, was no longer needed and eventually disappeared, as did the white of the egg. The chorion, however, remained as the most external coat of the developing embryo through which nourishment reaches the embryo. It acquired the ability to adhere closely to the walls of the uterus (which was what that part of the oviduct holding the embryo had become) and became the so-called trophoblast. The blood-vessel network of the underlying allantois conveys nutrients that diffuse through the trophoblast to the body of the embryo proper. These modifications gave rise to a new organ, the placenta, formed from tissues of both the mother and the embryo: the uterine wall with its blood vessels provided by the mother; the trophoblast and allantois—and in some mammals also the yolk sac—with their blood vessels provided by the embryo.

The overall development of placental mammals as a result of these changes is profoundly different from that of their ancestors, the reptiles, and proceeds in the following way: the tiny yolkless egg is fertilized in the upper portion of the oviduct by sperm received from the male in the process of coupling (coitus); cleavage starts as the egg is propelled slowly down the oviduct by action of cilia in the oviduct lining. At the end of cleavage a solid ball of cells called a morula is produced. The surface cells of the morula become the trophoblast and the inner cell mass gives rise to the embryo (the formative cells) and also its yolk sac, amnion, and allantois. A cavity appears within the morula, converting it into a hollow embryo, called the blastocyst. This cavity resembles the blastocoel but, in fact, is analogous to the yolk sac of meroblastic eggs, except that there is no yolk and the cavity is filled with fluid. At the blastocyst stage, the embryo enters the uterus and attaches itself to the uterine wall. This attachment, or implantation, a crucial step in the development of a mammal, is attained through the action of the trophoblast, which forms extensions, known as villi, that penetrate the uterine wall (Figure 9). In higher placental mammals, the lining of the uterine wall and, in varying degrees, the underlying tissues as well are partially destroyed, resulting in a closer relationship between the blood supplies of the mother and the embryo. Indeed, in man and in some rodents, the blastocyst sinks completely into the uterine wall and becomes surrounded by uterine tissue.

Implantation

While implantation takes place, the formative cells arrange themselves in the form of a disk under the trophoblast. In the disk, the germinal layers develop much as in birds, with the formation of a primitive streak and migration of the chordamesoderm into a deeper layer. A layer of endoderm is formed adjoining the cavity of the blastocyst, and an amniotic cavity develops, enclosing the embryo; in lower placental mammals, the allantois also develops. The embryo proper, lying in the amniotic cavity, is connected to the extra-embryonic parts by the umbilical cord. The umbilical cord lengthens greatly during later development. In higher mammals, the cavity of the allantois is reduced, but the allantoic blood vessels become well developed and extend through the umbilical cord, connecting the embryo to the placenta. The blood that circulates in the placenta brings oxygen and nutrients from the maternal blood to the embryo and carries away carbon dioxide and other waste products from the em-

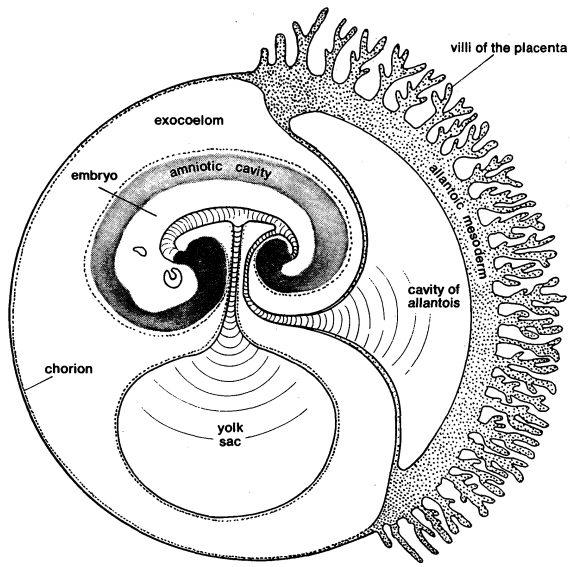


Figure 9: Embryo of placental mammal surrounded by embryonic membranes.

bryo to the maternal blood for disposal by the maternal body.

Although tissues of maternal and embryonic origin are closely apposed in the placenta, there is little actual mingling of the tissues. Despite an occasional penetration of an embryo cell into the mother and vice versa, there is a placental barrier between the two tissues. The blood circulation of the mother is at all times completely separated from that of the embryo and its extra-embryonic parts. The placental barrier, however, does allow molecules of various substances to pass through; such differential permeability is indeed necessary if the embryo is to obtain nourishment. The permeability of the placental barrier differs in different animals; thus antibodies, which are protein molecules, may penetrate the placental barrier in man but not in cattle.

The maintenance of the fetus—as the more advanced embryo of a mammal is called—in the uterus is under hormonal control. In the initial stages of pregnancy, the continued existence of the embryo in the uterus depends on the hormone progesterone, which is secreted by the corpora lutea, "yellow bodies," that develop in the ovary after an egg has been released.

At birth the fetal parts of the placenta separate from the maternal parts. Contraction of the uterine wall first releases the fetus from the uterus; the fetal parts of the placenta (the afterbirth) follow. In certain cases of intimate connection between fetal and maternal tissues, the maternal tissues are torn, and birth is accompanied by profuse bleeding.

III. Organ formation

PRIMARY ORGAN RUDIMENTS

Immediately after gastrulation—and sometimes even while gastrulation is underway—the germinal layers begin subdividing into regions that will give rise to various parts of the body. Subdivision proceeds in stages: initially a mass of cells is set aside for an organ system (for the alimentary canal, for instance) and subsequently further subdivided into the rudiments of various parts of the organ system, such as the liver, stomach, and intestines. The initially formed larger units are referred to as primary organ rudiments; those they later give rise to, as secondary organ rudiments.

The type of organ rudiment produced depends on the organization of the body in any particular group in the animal kingdom. In the vertebrates the earliest subdivision within a germinal layer is the segregation within the chordamesodermal mantle of the rudiment of the notochord from the rest of the mesoderm (Figure 7). During gastrulation the material of the notochord comes to lie middorsally in the roof of the archenteron. It separates by longitudinal crevices from the chordamesodermal mantle lying to the left and right. The material of the notochord then rounds off and becomes a rod-shaped strand of cells immediately under the dorsal ectoderm, stretching from the blastopore toward the anterior end of the embryo, to the midbrain level. In front of the tip of the notochord, there remains a thin sheet of prechordal mesoderm.

Development of notochord

The mesodermal layer adjoining the notochord becomes thickened and, by transverse crevices, subdivided into sections called somites (Figure 10). The somites, which later give rise to the segmented body muscles and the vertebral column, are the basis of the segmented organization typical of vertebrates (seen especially in the lower fishlike forms but also in the embryos of higher vertebrates). The lateral and ventral mesoderm, which remains unsegmented, is called the lateral plate. The somites remain connected to the lateral plate by stalks of somites that play a particular role in the development of the excretory (nephric) system in vertebrates; for this reason they are called nephrotomes. Rather early the mesodermal mantle splits into two layers, the outer parietal (somatic) layer and the inner visceral (splanchnic) layer, separated by a narrow cavity (Figure 11) that will expand later to form the coelomic, or secondary, body cavity. The coelomic cavity extends initially through the nephrotomes into the somites; in the somites it is eventually obliterated. Endoderm completely surrounds the lumen of the archenteron (when present) and produces the cavity of the alimentary canal. If no arch-

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

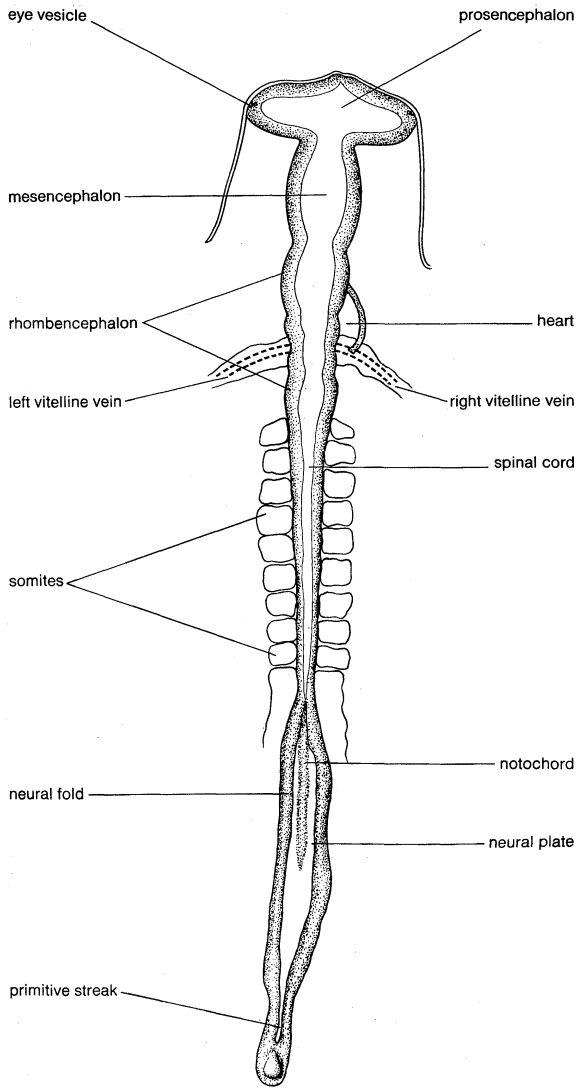


Figure 10: Early chick embryo, showing development of the brain vesicles and of some of the mesodermal structures.

enteric cavity is formed during gastrulation, the cavity of the alimentary canal is formed by the separation of cells in the middle of the mass of endoderm (as in bony fishes) or by folding of the sheet of endoderm (Figure 7).

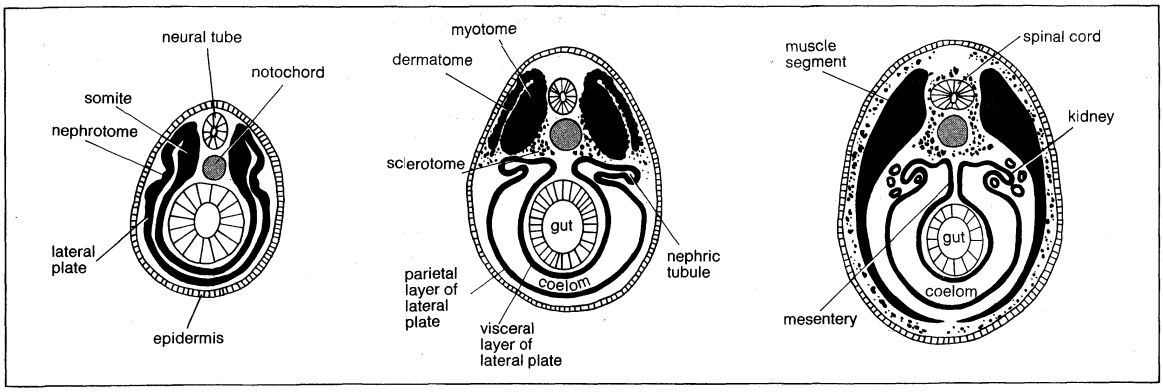


Figure 11: (Left to right) Subdivision of the mesoderm in a vertebrate embryo.

The endodermal gut sooner or later acquires an extended anterior part called the foregut and a narrower and more elongated posterior part, the hindgut. Characteristic of chordates is the development of the nervous system from a part of ectoderm lying originally on the dorsal side of the embryo, above the notochord and the somites. This part of the ectodermal layer thickens and becomes the neural plate (Figure 7:1), whose edges rise as neural folds that converge toward the midline, fuse together, and form the neural tube. In vertebrates the neural tube lies immediately above the notochord and extends beyond its anterior tip. The neural tube is the rudiment of the brain and spinal cord; its lumen gives rise to the cavities, or ventricles, of the brain and to the central canal of the spinal cord. The remainder of the ectoderm closes over the neural tube and becomes, in the main, the covering layer (epithelium) of the animal's skin (epidermis). As the neural tube detaches itself from the overlying ectoderm, groups of cells pinch off and form the neural crest, which plays an important role in the development of, among other things, the segmental nerves of the brain and spinal cord.

Formation of the neural crest

In developing the primary organ rudiments mentioned above, the embryo acquires a definite organization clearly recognizable as that of a chordate animal. Similar processes, which occur in the development of other animals, establish the basic organization of an annelid, a mollusk, or an arthropod.

The organization of the embryo as a whole appears to be determined to a large extent during gastrulation, by which process different regions of the blastoderm are displaced and brought into new spatial relationships to each other. Groups of cells that were distant from each other in the blastula come into close contact, which increases possibilities for interaction between materials of different origin. In the development of vertebrates in particular, the sliding of cells (presumptive mesoderm) into the interior and their placement on the dorsal side of the archenteron (in the archenteric "roof"), in immediate contact with the overlying ectoderm, is of major importance in development. Experiments have shown that, at the start of gastrulation, ectoderm is incapable of progressive development of any kind; that only after invagination, with chordamesoderm lying directly underneath it, does ectoderm acquire the ability for progressive development. The dorsal mesoderm, which later differentiates into notochord, prechordal mesoderm, and somites, causes the overlying ectoderm to differentiate as neural plate. Lateral mesoderm causes overlying ectoderm to differentiate as skin. The influence exercised by parts of the embryo, which causes groups of cells to proceed along a particular path of development, is called embryonic induction. Though induction requires that the interacting parts come into close proximity, actual contact is not necessary. The inducing influence—whatever it might be—is a diffusible substance emitted by the activating cells (the inductor). The inducing substance of the mesoderm is a large molecule, probably a protein or a nucleoprotein, which presumably penetrates reacting cells, though direct and unequivocal proof of such penetration is still unavailable. Inducing substances are active on vertebrates belonging to many different classes; *e.g.*, inductions of primary organs have been obtained by transplanting mammalian tissues into frog embryos or by transplanting tissues of a chick embryo into the embryo of a rabbit.

Induction is responsible not only for the subdivision of ectoderm into neural plate and epidermis but also for the development of a large number of organ rudiments in vertebrates. The notochord is a source of induction for the development of the adjoining somites and nephrotomes; the latter appear jointly to induce development of limb rudiments from the lateral plate mesoderm. Further examples are mentioned below in connection with development of the various organs.

Inducing substances

Since the results of induction are different for different organ rudiments, it must be presumed that there exist inducing substances with specific action, at least to a certain extent; thus, the lateral mesoderm induces differentiation of the skin but not neural plate from the very same kind of ectoderm. The number of inducing substances need not, however, be the same as the number of different kinds of tissues and organs, since certain differentiations could possibly be induced by a combination of two or more inducing substances, or the same inducing substance might have different effects on different tissues. It has been suggested that the regional organization of the entire vertebrate body could be controlled by the graded distribution of only two inducing substances—provisionally named the neuralizing substance and the mesodermalizing substance—along the length of the embryo. The neuralizing substance, concentrated at the anterior end, gradually decreases toward the posterior end; the mesodermalizing substance, on the other hand, is concentrated at the posterior end and decreases toward the anterior end. The differentiation of induced structures depends on the relative amounts of the two inducing substances at any given point in the embryo. Acting alone, the neuralizing substance induces only nervous tissue, which takes the form of the forebrain, and the mesodermalizing substance induces only mesodermal structures (*e.g.*, somites, notochord).

In the amphibian embryo, induction appears to have its primary source in the dorsal lip of the blastopore, which eventually gives rise to the notochord and adjoining somites. Induction by the notochord and somites is responsible for the development of the neural plate in the ectoderm, of lateral and ventral parts of the mesodermal mantle, and of the lumen of the alimentary canal in the endoderm. The dorsal lip of the blastopore for this reason has been called the primary organizer. In higher vertebrates, in which gastrulation occurs through the medium of a primitive streak, the anterior end of the streak and the Hensen's node have properties similar to those of a primary organizer. Organization centres have been found, or suspected, in embryos of animals belong-

ing to a few other groups, in particular the insects and sea urchins, but the interpretation of the experimental results in these animals is less satisfactory than in the case of vertebrates.

Concept of organization centre

The concept of an organization centre suggests that a part of the embryo differs from the rest of the embryonic tissues in being more active. The more active parts of the embryo (and also of animals in later stages of development) are particularly sensitive to certain noxious influences in their environment. If an embryo is deprived of oxygen or subjected to weak concentrations of poisons, the first parts to suffer are the most morphogenetically active ones. In vertebrate embryos the anterior end of the head is most sensitive. Early sea-urchin embryos have two centres of maximal sensitivity: one at the animal pole and the other at the vegetal pole. The damage done by noxious influences may result in actual breakdown of cells in a region of maximal sensitivity and may also lead to a depression of the developmental potential of the cells. Thus, the graded distribution of certain physiological properties appears to play a part in morphogenetic processes: physiological gradients are in fact also morphogenetic gradients.

Gradients in the embryo can be used to control development to a certain extent, by exposing the embryo to influences that, while reaching all parts, have a local effect as the result of differences in sensitivity. Disturbances of normal development often are the result of disruptions of gradients (see also MALFORMATION, BIOLOGICAL).

ORGANOGENESIS AND HISTOGENESIS

The primary organ rudiments continue to give rise to the rudiments of the various organs of the fully developed animal in a process called organogenesis. The formation of organs, even those of diverse function, shares some common features, which are considered in this section. As the organs form, so do their component tissues, in a process termed histogenesis.

A germinal layer, as the name implies, is a sheet of cells. An organ rudiment may be formed and separated from such a sheet in several ways. A groove, or fold, may appear within the layer, become closed into a tube, and then separated from the original layer. A tube once formed may be subdivided into sections by constrictions and dilations of the tube at certain points. This is the way the nervous system rudiment is formed in vertebrates as already described.

Alternatively, the germinal layer may produce a round depression, or pocket. The pocket may then separate from the layer as a vesicle, or it may elongate and branch at the tip while still connected with the layer. The latter method is common in the development of various glands and also the lungs in vertebrates.

Still another method of rudiment formation in a germinal layer is by the development of local thickenings, elongated or round, and detachment from the epithelial sheet. If a lumen appears later within such a body, the result may be the same as that achieved by folding—that is, a tube or vesicle may be formed. Indeed, the same sort of organ may develop even in related animals in either of these ways. The epithelial layer may further be cut up into segments, with the layer losing continuity, as in the formation of somites in vertebrates or similar mesodermal blocks in segmented invertebrates (*e.g.*, annelids and arthropods).

Embryonic mesenchyme

Lastly, the cells of a germinal layer may give up their connection to each other and become a mass of loose, freely moving cells called embryonic mesenchyme. This mass gives rise to various forms of connective tissue but may also condense into more solid structures, including parts of the skeleton and the muscles.

Many organs are comprised of all three germinal layers. It is very common for glands, for instance, to derive their lining from an ectodermal or endodermal epithelium and their connective tissue (sometimes in the form of a capsule) from mesenchyme of mesodermal origin. Parts of ectoderm and endoderm cooperate also in the development of the lining of the alimentary canal, and mesoderm provides the connective tissue and muscular sheath of the canal.

In this section the development of organs of the body are dealt with according to the germinal layer that contributes the most important part, and only the development of vertebrate organs is considered.

IV. Ectodermal derivatives

THE NERVOUS SYSTEM

The vertebrate nervous system develops from the neural plate—a thickened dorsal portion of the ectoderm—which forms a tube, as described earlier. From the very start the tube is wider anteriorly, the end that gives rise to the brain. The posterior part of the neural tube, which gives rise to the spinal cord, is narrower and stretches as the embryo lengthens. Stretching involves the head to only a very minor degree.

The brain and spinal cord. Constrictions soon appear in the brain region of the neural tube, subdividing it into three parts, or brain vesicles, which undergo further transformations in the course of development. The most anterior of the primary brain vesicles, called the prosencephalon, gives rise to parts of the brain and the eye rudiments. The latter appear in a very early stage of development as lateral protrusions from the wall of the neural tube (Figure 10), which are constricted off from the remainder of the brain rudiment as the optic vesicles. The rest of the prosencephalon constricts further into two portions, an anterior one, or telencephalon, and a posterior one, or diencephalon. The telencephalon gives rise, in lower vertebrates, to the smell, or olfactory, centre; in higher vertebrates and man, it becomes the centre of mental activities. The diencephalon, with which the eye vesicles are connected, was presumably originally an optic centre, but it has acquired, in the course of evolution, a function of hormonal regulation. The floor of the diencephalon forms a funnel-shaped depression, the infundibulum, which becomes connected with the pituitary, or hypophysis, the most important gland of internal secretion (*i.e.*, endocrine gland) in vertebrates. Indeed, the posterior lobe of the hypophysis is actually derived from the floor of the diencephalon. Tissues of the infundibulum and the posterior lobe of the hypophysis produce certain hormones (oxytocin and vasopressin) and stimulate the production and release of other hormones from the anterior lobe of the hypophysis.

The second primary brain vesicle, the mesencephalon, gives rise to the midbrain, which, in higher vertebrates, takes part in coordinating visual and auditory stimuli.

The third primary brain vesicle, the rhombencephalon, is more elongated than the first two; it produces the metencephalon, which gives rise to the cerebellum with its hemispheres, and the myelencephalon, which becomes the medulla oblongata. The cerebellum acts as a balance and coordinating centre, and the medulla controls functions such as respiratory movements.

Differentiation of nervous tissue

The cells constituting the wall of the neural tube and, later, of the brain and spinal cord become arranged in such a way that they point into the central cavity of the tube. The differentiation of nervous tissue involves many cells abandoning their connection to the inner surface of the neural tube and migrating outward, where they accumulate as a mantle. The first cells to migrate become the neurons, or nerve cells. They produce outgrowths called axons and dendrites, by which the cells of the nervous system establish communication with one another to form a functional network. Some of the outgrowths extend beyond the confines of the brain and spinal cord as components of nerves; they establish contact with peripheral organs, which thus fall under the control of the nervous system. Cells migrating from the inner surface of the neural tube later in development become astrocytes, which are the supporting elements of nerve tissue.

The fate of nerve cells is dependent largely on whether they succeed, directly or indirectly (through other neurons), in connecting with peripheral organs. Nerve cells that fail to establish connections die. Thus, if in early stages of embryonic development, some organ, a limb

rudiment for instance, is surgically removed, the nerve cells in the centres supplying nerves to such an organ are reduced in number, and the corresponding nerves also diminish or disappear. On the other hand, if an organ is introduced by transplantation into a developing embryo, the organ will be supplied by nerves from a nerve centre in which the number of cells apparently increases; no additional cells are provided, but cells that would otherwise have degenerated remain active and differentiate into functional neurons, thus satisfying the demand created by the additional organ.

Nerves do not consist entirely of outgrowths of neurons located in the brain and spinal cord. Many components of nerves are outgrowths of neurons, the cell bodies of which are located in masses called ganglia; there are three main types of ganglia: spinal ganglia, cranial ganglia, and ganglia of the autonomous nervous system. The spinal ganglia are derived from cells of the neural crest—the loose mesenchyme-like tissue that remains between the neural tube and skin after separation of the two. Part of the cells of the neural crest in the region of the trunk and tail accumulate in segmental groups (corresponding to the mesodermal somites) and provide fibres to peripheral organs and to the spinal cord. These fibres constitute the sensory pathways in the spinal nerves. The motor components of the spinal nerves—fibres that activate muscles—are outgrowths of neurons lying in the spinal cord. The ganglia of the cranial nerves are produced only in part from cells of the neural crest; an additional component comes from the epidermis on the side of the head. Cells of the epidermal thickenings called placodes detach themselves and contribute to the formation of the cranial ganglia and thus of the cranial nerves.

The ganglia of the autonomous (sympathetic) nervous system are derived, as are the spinal ganglia, from neural-crest cells, but, in this case, the cells migrate downward to form groups near the dorsal aorta, near the intestine, and even in the intestinal wall itself. The outgrowths of cells in these ganglia are the nerve fibres of the sympathetic nerves (see also NERVES AND NERVOUS SYSTEMS).

Major sense organs. *The eye.* As has been pointed out, the rudiments of the eyes develop from optic vesicles, each of which remains connected to the brain by an eye stalk, which later serves as the pathway for the optic nerve. The optic vesicles extend laterally until they reach the skin, whereupon the outer surface caves in so that the vesicle becomes a double-walled optic cup. The thick inner layer of the optic cup gives rise to the sensory retina of the eye; the thinner outer layer becomes the pigment coat of the retina. The opening of the optic cup, wide at first, gradually becomes constricted to form the pupil, and the edges of the cup surrounding the pupil differentiate as the iris. The refractive system of the eye and, in particular, the lens of the eye are derived not from the cup but from the epidermis overlying the eye rudiment. When the optic vesicle touches the epidermis and caves in to produce the optic cup, the epidermis opposite the opening thickens and produces a spherical lens rudiment. The lens develops by an induction by the optic vesicle on the epidermis with which it comes in contact. A further influence emanating from the eye changes the epidermis remaining in place over the lens into a transparent area, the cornea. Influence of the optic cup on the surrounding mesenchyme causes the latter to produce a vascular layer around the retina and, outside of that, a tough fibrous or (in some animals) even a partly bony capsule called the sclera. Thus a complex interdependence of different materials produces the fully developed and functional vertebrate eye (see also PHOTORECEPTION).

Formation of the optic cup

The ear. The main part of the ear rudiment is derived from thickened epidermis adjoining the medulla. This area of the epidermis invaginates to produce the ear vesicle, which separates from the epidermis but remains closely apposed to the medulla. The ear vesicle becomes complexly folded to produce the labyrinth of the ear. Subsequently, a group of cells of the ear vesicle becomes detached and gives rise to the acoustic ganglion. Neurons of this ganglion become connected by their nerve fibres to the sensory cells in the labyrinth, on the one hand, and with the brain (the medulla), on the other. The ear vesicle, acting on the surrounding mesenchyme, induces the latter to aggregate around the labyrinth and form the ear capsule. Further parts with various origins are added to the ear: the middle ear, from a pharyngeal pouch and the associated skeleton, and the external ear (where present), from epidermis and dermis.

The olfactory organ. The olfactory organ develops from a thickening of the epidermis adjacent to the neural fold at the anterior end of the neural plate. This thickening is converted into a pocket or sac but does not lose connection with the exterior. The openings of the sac become the external nares, and the cavity of the sac becomes the nasal cavity. Some cells of the olfactory sac differentiate as sensory epithelium and produce nerve fibres entering the forebrain. In most fishes the olfactory sac does not communicate with the oral cavity; in lungfishes and in terrestrial vertebrates, however, canals develop from the olfactory sacs to the oral cavity, where they open by internal nares. A cartilaginous capsule forms around the olfactory organ from cells believed to have been derived from the walls of the sac itself, and thus it is ectodermal in origin.

Gustatory and other organs. Gustatory organs in the form of taste buds develop as local differentiations of the lining of the oral cavity but also, in fishes, in the skin epidermis. They are supplied with nerve endings, as are several other sensory bodies scattered among the tissues and organs of the developing body.

THE EPIDERMIS AND ITS OUTGROWTHS

The major part of the ectodermal epithelium covering the body gives rise to the epidermis of the skin. In fishes and aquatic larvae of amphibians, the many-layered epidermis is provided with unicellular mucous glands. In terrestrial vertebrates, however, the epidermis becomes keratinized; *i.e.*, the outer layers of cells produce keratin, a protein that is hardened and is impermeable to water. During the process of keratinization, many cell components degenerate and the cells die; the layer of keratinized cells is therefore shed from time to time. In reptiles the shedding may take the form of a molt in which the animal literally crawls out of its own skin. It is less well known that frogs and toads also moult, shedding the surface keratinized layer of their skin (which is usually eaten by the animal). In birds and mammals, keratinized cells are shed in pieces that are sloughed off, rather than in extensive layers. In many vertebrates local thickenings of the keratinized layer appear in the form of claws, hooves, nails, and horns.

Keratinization

The epidermis is only the superficial layer of the skin, which is reinforced by the dermis, a connective tissue layer of a much greater thickness. The cells of the dermis are derived from mesoderm and neural-crest cells. In particular the pigment cells found in the dermis of fishes, amphibians, and reptiles are of neural-crest origin. The pigment in the skin of birds and mammals (and also in hairs and feathers) is also produced by neural-crest cells, but in these animals the pigment cells penetrate into the epidermis or deposit their pigment granules there.

The structure of the skin is further complicated by the development of hairs and feathers, on the one hand, and of skin glands, on the other. Hairs and feathers develop from a somewhat similar kind of rudiment. The development starts with a local thickening of the epidermal layer, beneath which a group of mesenchyme cells accumulate. In the case of hairs, the epidermal thickening proliferates downward and forms the root of the hair, from which the shaft then grows outward, emerging on the surface of the skin. In the case of feathers, the epidermal thickening bulges outward to form a hollow fingerlike protrusion with a connective tissue core. Secondarily, the shaft of the feather branches characteristically to produce barbs and barbules. In both cases, however, the final structure—shaft of the hair and shaft barbs and barbules of the feather—consists of keratinized and, thus, dead cells.

The skin of amphibians and mammals (but not of birds and reptiles) is provided with numerous skin glands, which develop as ingrowths from the epidermis. A peculiar type of skin gland is the mammary gland of placental mammals. In the first stage of development, mammary-gland rudiments resemble hair rudiments; they are thickenings of the epidermis, with condensed mesenchyme on their inner surfaces. In some mammals (rabbit, man) two continuous epidermal thickenings called mammary lines stretch along either side of the belly of the embryo. Parts of the line corresponding in number and position to the future glands enlarge while the rest of the thickening disappears. The initial thickenings proliferate inward and produce a system of ramified cords, solid at first but hollowed out later, which become the lactiferous, or milk-bearing, ducts of the gland. Further branching at the tips of the ducts gives rise to smaller ducts and to the secretory end sacs, or alveoli, of the gland.

V. Mesodermal derivatives

THE BODY MUSCLES AND AXIAL SKELETON

Derivatives of somites

The somites, formed in the early stages of development from the upper edges of the mesodermal mantle adjoining the notochord, are complex rudiments that subdivide and give rise to very diverse body structures. The coelomic cavity, present initially, becomes obliterated by the side-to-side flattening of the somites, so that the thinner, outer parietal layer of the somite comes in close contact with its thicker visceral layer. The visceral layer of the somite very early subdivides into two parts. The upper, dorsolateral part called the myotome (Figure 11) remains compact, giving rise to the body muscles. The lower, medioventral part of the somite, called the sclerotome, breaks up into mesenchyme, which contributes to the axial skeleton of the embryo—that is, the vertebral column, ribs, and much of the skull. The parietal layer of the somite, at a later stage, is converted into mesenchyme that, together with components of the neural crest, gives rise to the dermis of the skin and, for this reason, is called the dermatome.

The cells of the myotome are elongated in a longitudinal direction and become differentiated as muscle fibres. The myotomes, originally situated dorsally, expand on either side, penetrating between the skin on the outside and the lateral plates of the mesoderm on the inside, until they meet midventrally; the whole body is thus enclosed in a layer of developing muscle. As the somites and myotomes are segmented, so are the muscles derived from them. Metamerism, or segmentation, a feature in the embryos of all vertebrates, remains preserved only in the adults of fishes and of terrestrial vertebrates that have elongated bodies (salamanders, snakes); it becomes largely erased in four-footed animals that depend on their limbs for locomotion.

The mesenchyme derived from the sclerotomes condenses as cartilage around the notochord and the spinal cord. It forms the cartilaginous vertebral column and ribs. In the head region it produces a part of the cartilaginous skull, mainly its posterior and ventral parts; anteriorly the somitic mesenchyme is supplemented by mesenchyme from the neural crest. Cartilaginous capsules of the olfactory organ and the ear fuse with the cartilaginous capsule surrounding the brain; to this complex are also added cartilages associated with the jaws and gill skeleton. Cartilage in the vertebral column and in the skull is replaced later in the bony fishes and in the terrestrial vertebrates by bone. At a still later stage, dermal bones are added, which, while they have no precursors in the cartilaginous skeleton, develop in the adjoining mesenchyme.

THE APPENDAGES: TAIL AND LIMBS

The tail in vertebrates is a prolongation of the body beyond the anus. It develops in early stages from the tail bud, immediately dorsal to the blastopore. Material for the tip of the tail is situated slightly forward from the edge of the blastopore. The elongation of the back of the body is greater than that of the belly; as a result the tip of the tail bud is carried beyond the blastopore and thus beyond the anus, which, in the developed embryo, marks the position of the blastopore. The consequence is that a section of the dorsal surface of the embryo comes to lie on the ventral surface of the tail; *i.e.*, becomes inflected. The tail bud is formed from parts that have already been differentiated to a certain extent; prolongations of the neural tube and of the notochord are involved, and endoderm extends into the tail rudiment as the postanal gut, which, however, soon degenerates. The bud is also encased in ectodermal epidermis. In amphibians the somites of the tail are not derived from the chordamesodermal mantle but from the inflected posterior portion of the neural plate, which loses its nervous nature and becomes subdivided into segments corresponding to the somites of the trunk. In higher vertebrates the cells in the interior of the tail bud have an undifferentiated appearance and form a growth zone, at the expense of which parts of the tail (neural tube, notochord, somites) are extended backward as the tail elongates.

Formation of the tail bud

The paired limbs of vertebrates derive their first rudiments from the upper edge of the lateral plate mesoderm. The parietal layer becomes thickened, and cells escape from the epithelial arrangement and form a mesenchymal mass adjoining the ectodermal epithelium at the surface of the body. The ectodermal epithelium over the mass of mesenchyme likewise becomes thickened. In higher vertebrates, the accumulation of mesodermal cells and the thickening of the epidermis occur along the entire length of the trunk, from neck to anus, but in the middle of the trunk they soon disappear, and only the most anterior and the most posterior sections develop further into the rudiments of the forelimbs and hindlimbs, respectively. In fishes, the rudiments of the pectoral and pelvic fins are more extended anteroposteriorly in earlier than in final stages.

The mesodermal masses of the limb rudiments proliferate, and, covered with thickened epidermis, form on the surface of the body conical protrusions called the limb buds, which, once formed, possess all the materials necessary for limb development. Limb buds may be transplanted into various positions on the body or on the head and there develop into clearly recognizable limbs, conforming to their origin, whether a forelimb or hindlimb, a wing or a leg in birds. This specificity of the limb is carried by the mesodermal part of the rudiment, but a complex interaction between the mesodermal mesenchyme and the ectodermal epidermis is necessary for the normal development of the limb. In four-limbed vertebrates (tetrapods), the tips of the limb buds become flattened and broadened into hand or foot plates. The edge of the plate is indented, forming the rudiments of the digits. Meanwhile, local areas of the mesodermal mesenchyme in the interior of the limb rudiment condense; these are the rudiments of the various components of the limb skeleton. In fishes, small outgrowths from the myotomes enter the limb rudiment to form the muscles of the fins. In tetrapods, however, the limb muscles develop from the same mass of mesenchyme that gives rise to the skeleton. Thus the muscles of the body and the muscles of the limbs have different origins—the first develop from the myotomes (thus from the somites), and the second develop from the lateral plate mesoderm via the limb buds.

Limb buds

The nerves supplying the limbs grow into the limb rudiments from the spinal cord and the spinal ganglia. The nerves are guided in some way by the limb rudiments, for, if limb rudiments are displaced by transplantation to an abnormal position, the nerves still find their way and establish normal relationships to the limb muscles. Limb rudiments transplanted to sites very far from their normal positions induce local nerves to enter the limb, thereby making it motile.

EXCRETORY ORGANS

The kidneys of vertebrates consist of a mass of tubules that develop from the stalks of somites called nephrotomes. In some primitive vertebrates such as cyclostomes, the nephrotome in each segment gives rise to only one tubule (Figure 11), but, in the great majority of verte-

brates, mesenchyme from adjacent nephrotomes fuses into a common mass that differentiates into a number of nephric tubules irrespective of the original segmentation of the mesoderm. Under primitive conditions each tubule opens by a funnel (the nephrostome) into the coelomic cavity; the opposite ends of the tubules fuse to form the collecting ducts of the kidney. A collection of capillaries (the glomerulus) becomes associated with the nephric tubule, forming its filtration apparatus. The glomerulus may be situated in the coelomic cavity opposite the nephrostome or, in all the more advanced animals, intercalated into the nephric tubule, forming with the latter a renal corpuscle of the kidney. In adults of all vertebrates above the amphibians, the nephrostomes disappear (or are never formed), so that the tubule begins with the renal corpuscle. Parts of the kidney in vertebrates can be distinguished as the pronephros (most anteriorly, at the forelimb level), the mesonephros (in the midtrunk region), and the metanephros (in the pelvic region). The three sections of the kidney develop at different stages, starting with the pronephros and ending with the metanephros. In their morphology and mode of development, the anterior parts show more primitive conditions than the posterior ones. The pronephros, developing early in embryo formation, is the functional kidney of fish and amphibian larvae. Its collecting duct opens into the hindmost part of the intestine, called the cloaca, and later also serves as the collecting duct of the mesonephros. In reptiles, birds, and mammals, the pronephros is nonfunctional, although even in these animals its duct persists as the mesonephric duct. The mesonephros develops later and replaces the pronephros as the functional kidney of adult fishes and amphibians and of the embryos of reptiles, birds, and mammals. The tubules of the mesonephros link up with the duct derived from the pronephros. The pronephric duct in fact stimulates the development of mesonephric tubules, and, in its absence, the mesonephros does not develop at all.

Development of the metanephros

The metanephros is found only in reptiles, birds, and mammals. It replaces the mesonephros of the early embryonic stages and continues as the functional kidney in the postembryonic and adult life of these animals. The metanephros develops from mesenchyme derived from the nephrotomes of the posterior part of the trunk and lying dorsal to the mesonephric duct. The actual differentiation is initiated by a dorsal outgrowth of the mesonephric duct, called the ureteric bud. The ureteric bud grows in the direction of the mesenchyme and becomes the ureter. Having penetrated the mass of mesenchyme, it starts to branch, producing the collecting tubules of the kidney; the mesenchyme, meanwhile, in response to the influence of the duct and its branches, aggregates to form the excretory tubules of the kidney. The influence of the ureter is indispensable for the development of the metanephric excretory tubules, for, if the ureter fails to develop or, in its outgrowth, stops short of reaching the kidney-producing mesenchyme, no kidney develops.

CIRCULATORY ORGANS

The rudiment of the heart in vertebrates develops from the ventral edges of the mesodermal mantle in the anterior part of the body, immediately adjoining the pharyngeal region. A group of mesodermal cells breaks away from the ventral edge of the lateral plate, takes a position just underneath the pharyngeal endoderm, and becomes arranged in the form of a thin-walled tube, which will become the endocardium, or lining of the heart (Figure 12). In vertebrates with complete cleavage, the endocardial tube is single and medial from its start. In higher vertebrates with meroblastic cleavage—reptiles, birds, and mammals—the embryo in early stages of development is flattened out on the surface of the yolk sac; therefore, what are morphologically the ventral edges of the mesodermal mantle lie far apart on the perimeter of the blastodisc. As a result of this arrangement, two endocardial tubes are formed, one on either side of the embryo. Subsequently, when the embryo becomes separated from the yolk sac, the two endocardial tubes meet in the midline ventral to the pharynx and fuse, producing a single heart rudiment. After the formation of the endocardium, the coelomic cavity in the lateral plate mesoderm adjoining the heart rudiment expands slightly and envelops the endocardial tube or tubes (Figure 20). The heart muscle layer, or myocardium, develops from the visceral (splanchnic) layer of the lateral plate in contact with the endocardial tube; the parietal (somatic) layer of the lateral plate forms the pericardium, or covering of the heart. The portion of the coelom surrounding the heart is separated from the rest of the body cavity and becomes the pericardial cavity.

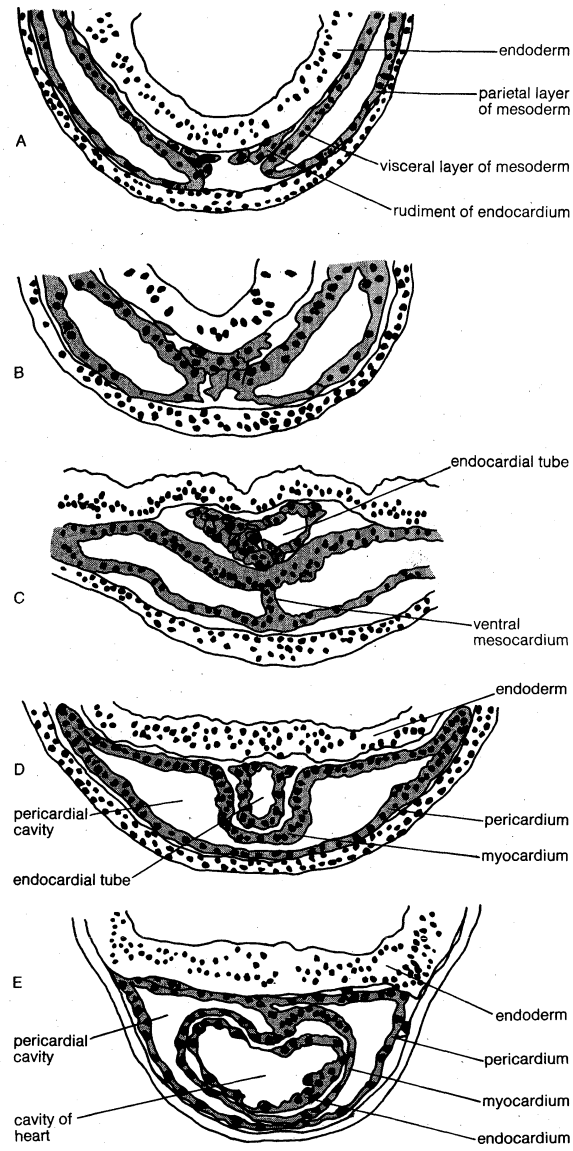


Figure 12: Development of the heart in amphibian embryos.
From O. Hertwig, *Handbuch der Vergleichendenden und Experimentellen Entwicklungslehre der Wirbeltiere* (1906); Gustav Fischer Verlag

The endocardial tube branches anteriorly into two tubes, the ventral aortas; a similar branching of the endocardial tube posteriorly forms the two vitelline veins, which carry blood from the midgut endoderm or from the yolk sac (when present) to the heart.

In its earliest development, the heart rudiment shows a degree of dependence on the adjoining endoderm. The whole of the endoderm can be removed in newt embryos in the neural-tube stage. In such endodermless embryos, the heart fails to develop, even though the mesoderm destined to form the heart rudiment is left intact.

Development of the heart

The heart is initially a straight tube stretching in an anteroposterior direction. Rather early in development, however, it becomes twisted in a characteristic way and subdivided into four main parts: the most posterior, the sinus venosus; the atrium, which comes to lie at the an-

teriorly directed bend of the tube; the ventricle, occupying the apex of the posteroventrally directed inflexion; and, most anteriorly, the conus arteriosus. In the course of development in the more advanced vertebrates, the atrium and ventricle become partially or completely subdivided into right and left halves. In amphibians, only the atrium is separated into two halves, by a partition starting from the posterior end. In reptiles, a partition separates the atria and part of the ventricle. In birds and mammals, the subdivision of the heart is complete, with two atria and two ventricles.

The complete subdivision of the heart is important for separating the pulmonary, or lung, blood supply from the general body circulation. But, if this separation developed early in the embryo, it would create difficulties, since the lungs of the embryo are not functional; the enrichment of the blood with oxygen occurs instead in the placenta. The partition between the atria in mammalian embryos remains incomplete, so that blood returning from the body and from the placenta enters into the right half of the heart but is shunted (through the interatrial foramen) into the left half of the heart and thence again into general circulation. At birth, however, the interatrial foramen is closed by a membraneous flap, and oxygen-depleted blood from the body enters the right atrium, is channelled into the right ventricle, and thence to the lungs for oxygenating.

In an adult vertebrate, blood vessels extend to all parts of the body. It would seem that channels for the supply of blood are provided in proportion to the local demand of the tissues; progressively developing organs or parts with particularly intensified function always receive an increased blood supply. The rudiments of blood vessels are always aggregations of mesenchyme cells. In any blood vessel the endothelial tube is formed first, and the muscular and elastic layers are added later.

The main blood channels in vertebrates develop in certain favoured situations; namely (1) between the endoderm and lateral plate mesoderm; (2) around the kidneys, especially the pronephros and mesonephros; and (3) in connection with the heart, which is a special case of the first category.

From the paired forward extensions from the heart, the ventral aortas, loops develop between the pharyngeal clefts. These are the aortic arches (Figure 13), which served originally to supply blood to the gills in aquatic vertebrates. The arches are laid down in all vertebrates, six or more being found in cyclostomes and fishes; six are present in the embryos of tetrapods, but the first two are degenerate. The arches of the third pair develop as the carotid arteries, supplying blood to the head. Those of the fourth pair (and, exceptionally, in urodeles also the fifth) join dorsally to form the dorsal aorta, providing blood to most of the body. These are the systemic arches. The arches of the sixth pair are the pulmonary arches; in embryos they carry blood to the dorsal aorta, as well as to the lungs, but in fully developed amniotes (reptiles, birds, and mammals), they carry blood only to the lungs.

The paired posterior extensions of the heart of the early embryo are the vitelline veins, whose branches spread out between the lateral plate mesoderm and the endoderm, especially the endoderm of the yolk sac, when present. On their way to the heart, the vitelline veins pass through the liver and break up into a system of small channels—the hepatic sinusoids. Parts of the vitelline veins lying posterior to the liver become the hepatic portal veins, which carry blood from the intestine to the liver; the parts of the vitelline veins anterior to the liver become the hepatic veins, which carry blood fom the liver to the sinus venosus in lower vertebrates (anamniotes), but become the anterior section of the postcaval vein in amniotes.

Whereas the vitelline veins and, later, the hepatic portal vein carry blood from the endodermal parts of the embryo and from the yolk sac to the heart, the blood from the mesodermal and ectodermal parts is returned to the heart through a system of cardinal veins (Figure 14). These latter veins start their development in the form

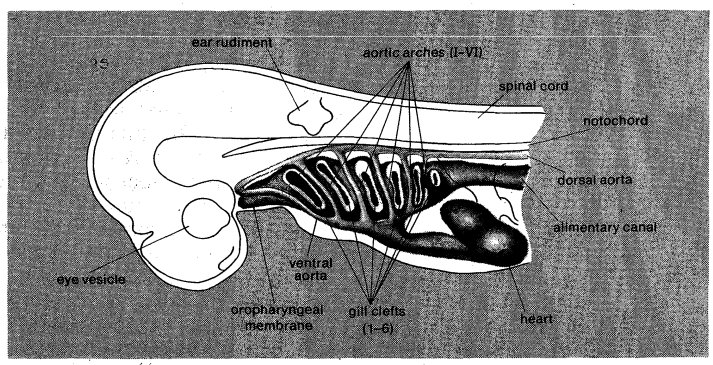


Figure 13: Relation of the aortic arches and the branchial clefts in a dogfish embryo.
From B.I. Balinsky, *An introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

of an irregular sinus around the pronephros, connected by the common cardinal veins (ducts of Cuvier), on either side, to the sinus venosus. Extensions anteriorly and posteriorly give rise to the precardinal and postcardinal veins, respectively. The postcaval vein, present in terrestrial vertebrates, is a late acquisition, both in evolution and in embryogenesis; it is a result of the intercommunication of several venous channels, including the anterior portion of the vitelline veins.

Development of the cardinal veins

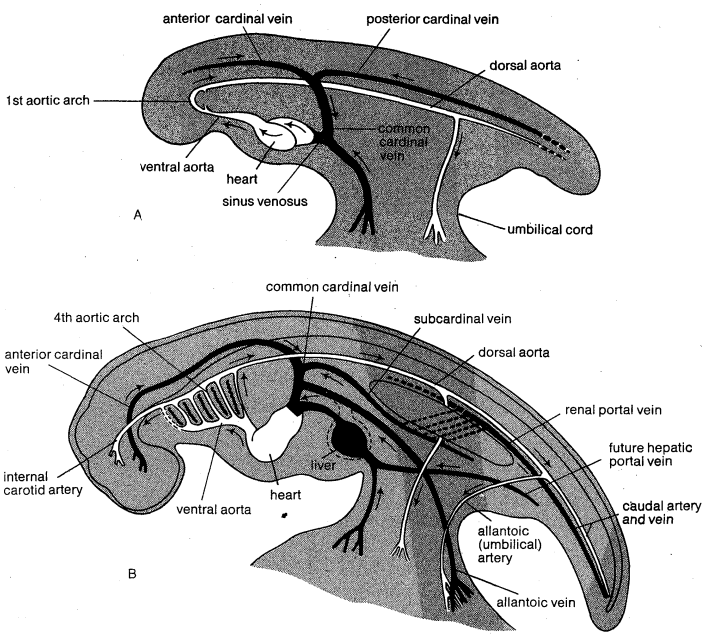


Figure 14: General arrangement of blood vessels in (A) an early amniote embryo and (B) the further development of the circulatory system. The arteries are shown in white, the veins in black.

The first blood cells in vertebrate embryos form in association with the intestinal endoderm on the yolk sac. Groups of mesoderm cells derived from the splanchnic layer of the lateral plate (extra-embryonically in cases in which a yolk sac is present) become so-called blood islands, which are particularly conspicuous on the yolk sac of bird embryos (in the area vasculosa). In bird's eggs, the internal cells of the blood island start producing hemoglobin (gas-carrying component of blood) and become the first red blood cells (erythrocytes) as early as the second day of incubation. The outer cells of the blood islands develop into an endothelial layer and form a network of blood vessels covering part of the surface of the yolk sac. The network acquires a connection to the vitelline veins and vitelline arteries (the latter being branches of the dorsal aorta); thus the blood corpuscles formed in the blood islands can enter the general blood circulation.

At later stages of embryogenesis, blood-cell formation

shifts from the blood islands to the liver and, still later, to the bone marrow.

The lymphatic system, in a manner similar to the blood vessels, develops by the local aggregation of connective tissue to form lymphatic vessels.

REPRODUCTIVE ORGANS

In considering the development of reproductive organs, distinctions must be made between: (1) the origin of sex cells (gametes), (2) the origin and differentiation of the sex glands, or gonads (ovaries and testes), and (3) the origin and development of the supporting parts of the reproductive system (*e.g.*, genital ducts, copulatory organs).

Origin of germ cells

The germ (germinal) cells, which eventually give rise to the gametes, are often segregated from the somatic, or body, cells at a very early stage—during cleavage and before the subdivision of the embryo into ectoderm, mesoderm, and endoderm. In the invertebrate nematodes, the very first of these primordial germ cells is identifiable after as few as five divisions of the egg cell. The germ cell retains the large chromosomes present in the fertilized egg; in the somatic cells the chromosomes become fragmented. Subsequently, the single germ cell gives rise, by mitotic divisions, to all the gametes in the gonad.

In vertebrates, primordial germ cells arise outside the gonads, but they cannot be distinguished in early cleavage stages. In amphibians, cytoplasm at the vegetal pole, rich in ribonucleic acids, becomes incorporated into a number of cells, which, during cleavage and gastrulation, lie among the yolky endoderm cells. Later they migrate into the mesodermal layer and become incorporated into the rudiments of the gonads. In higher vertebrates, primordial germ cells can be recognized in the extra-embryonic endoderm of the yolk sac. In mammals, these cells subsequently migrate into the mesoderm and are located in the gonad rudiments. The mouse embryo, for example, originally has fewer than 100 primary germ cells; during their migration, however, their numbers increase as a result of repeated divisions, to 5,000 or more in the gonads.

Although the primordial germ cells either may appear before the separation of germinal layers or be found originally in the endoderm, the gonads are invariably of mesodermal origin. In vertebrates, the first trace of gonad development is a thickening of the coelomic lining on either side of the dorsal mesentery and medial to the kidney rudiments. The thickening, elongated anteroposteriorly, is known as the germinal ridge (Figure 15). The ridge protrudes into the coelomic cavity, and the fold of thickened epithelium becomes filled with mesenchyme. At this stage the primordial germ cells invade the rudiments of the gonads and become associated with the somatic cells of the germinal ridge. In the functionally differentiated gonads, only the actual gametes and their predecessors (spermatogonia and oogonia) are derived from the primary germ cells; the supporting cells are somatic cells of local mesodermal origin. In the ovaries, the follicle cells surrounding and nourishing the young egg cells (oocytes) are of somatic origin, as are also the connective tissue and blood vessels of the gonad. In the testes, supporting elements called Sertoli cells are somatic cells, as are the interstitial cells, which are scattered between the sperm-carrying tubules of the testes and believed to be the source of male hormones.

In the early stages of their development—even while the gonad rudiment is being invaded by primordial germ cells—the female and male gonads are in an indifferent stage. Only later does tissue differentiation of the gonads begin and male or female gonadal development proceed.

Development of genital ducts

The genital ducts, by which the eggs and sperm are carried away from the gonads, are, in vertebrates, linked with the excretory system. In the male, the seminiferous tubules connect with the nephric tubules of the mesonephros, and the sperm are carried to the exterior by way of the mesonephric duct. In males of lower vertebrates, the mesonephric duct thus serves as a channel both for urine and for sex cells. In amniotes the development of

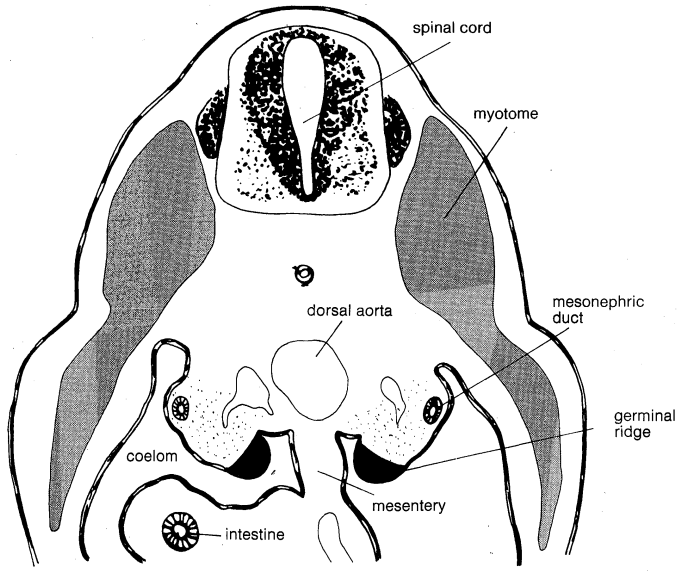


Figure 15: Transverse section of a mouse embryo showing the position of the germinal ridges and their relation to the mesonephric rudiment and the dorsal mesentery.
From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

the metanephros as the urine excreting organ has freed the mesonephric duct to carry products associated only with reproduction. In the female, a separate duct, the paramesonephric duct (Müllerian duct), develops beside the mesonephric duct. At its anterior end it utilizes the funnels of the pronephric tubules as its entrance (ostium). It is remarkable that the paramesonephric duct develops initially in both female and male embryos. The ducts remain in an indifferent stage longer than the gonads. Eventually the sex hormones produced by the differentiating gonads cause a corresponding differentiation of the ducts. The mesonephric ducts, which become reduced in female embryos, remain in male embryos as ducts for conveying sperm (ductus deferens). The paramesonephric ducts, on the other hand, degenerate in male embryos but become the oviducts in female embryos (Figure 16). In mammals, the terminal portions of the paired oviducts differentiate as two uteri, which, in primates and man, fuse to form a single uterus.

In all terrestrial vertebrates except the placental mammals, the genital ducts, as well as the ducts of excretory organs, open into the cloaca. In mammals, however, the

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

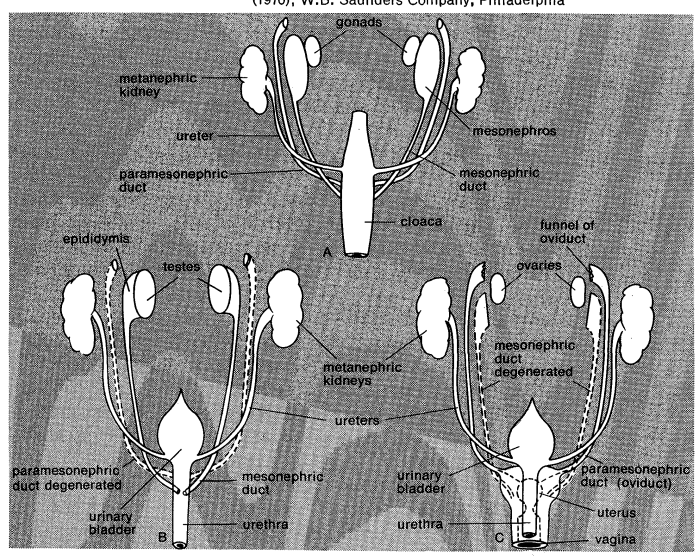


Figure 16: Transformations of the genital ducts in mammalian embryos in transition from an (A) indifferent stage to (B) the male and (C) female condition.

cloaca becomes subdivided into a dorsal part, which conveys the feces, and a ventral part, which receives excretory and genital products. In male mammals the excretory and genital ducts remain connected, having the urethra as their common outlet; in females the urethra serves only for the passage of urine and the uterus opens separately by means of the vagina.

Copulatory organs have developed independently in several groups of vertebrates having internal fertilization. The penis in mammals develops from an outgrowth called the genital tubercle, located at the anterior edge of the urinogenital orifice. The tubercle is laid down in a similar way in embryos of both sexes, and the region of the urinogenital orifice remains in an indifferent state even longer than do the genital ducts. In a comparatively late stage of embryonic life the genital tubercle of male embryos encloses the urethral canal and becomes the penis; in female embryos it remains small and becomes the clitoris.

VI. Endodermal derivatives

THE ALIMENTARY CANAL

The alimentary canal is the chief organ developing from endoderm. The way it forms depends on the type of egg cleavage. In eggs with holoblastic (complete) cleavage, after gastrulation the invaginated mass of endoderm lines the archenteron, the cavity of which becomes the alimentary canal, or gut. In eggs with meroblastic (partial) cleavage—and also in mammals (despite their complete cleavage)—the endoderm is produced in the form of a sheet lying flat over the yolk-sac cavity. Subsequently, folds of endoderm and splanchnic mesoderm appear—first anteriorly, then laterally, and lastly posteriorly—and sink, converging ventrally under the embryo and cutting off the future gut cavity from the cavity of the yolk sac. The most anterior and posterior portions of the gut separate, but the middle part remains in open communication with the yolk sac throughout embryonic life, eventually becoming reduced to the yolk stalk, which passes through the umbilical cord.

The alimentary canal of vertebrates becomes differentiated eventually into the oral cavity, pharynx, esophagus, stomach, and intestine. Whether derived from an archenteron or formed by folding of the endodermal sheet, the canal initially does not possess an opening at its anterior end. This is also the case in some invertebrates (lower chordates and echinoderms), which are grouped together with vertebrates as the Deuterostomia, or animals with secondary mouths.

Formation of vertebrate mouth

In vertebrates, a mouth forms by a rupture at the anterior end, where the endoderm is in contact with ectoderm. The ectoderm of the future mouth region becomes depressed, forming a mouth invagination, or stomodaeum. The ectodermal and endodermal layers separating the cavity of the stomodaeum from the gut fuse to form the oropharyngeal membrane, which thins and ruptures, providing free passage from the exterior to the gut. Because of its mode of origin, the oral cavity is in part lined by ectoderm and in part by endoderm, the two parts becoming indistinguishable. Before the oropharyngeal membrane ruptures, however, a small pocket forms on the dorsal side of the stomodaeal invagination. This, the rudiment of the anterior lobe of the hypophysis, becomes apposed to the ventral surface of the diencephalon and loses its connection with the mouth cavity.

The anal opening in some exceptional cases (urodele amphibians) is derived directly from the blastopore, which persists as a narrow canal after completion of gastrulation. In other vertebrates, however, the anus develops either near the location of the former blastopore or in a corresponding region at the posterior end of the embryo, where the last remnants of mesoderm migrated to the interior. It is thus claimed that the anus in vertebrates is derived, directly or indirectly, from the blastopore. The mode of formation of the opening is somewhat similar to that of the mouth. A slight invagination of the ectoderm occurs, and a cloacal membrane forms, separating the ectodermal invagination from the gut cavity. The membrane ruptures later to provide the anus.

THE PHARYNX AND ITS OUTGROWTHS

The anterior portion of the endodermal gut, lying immediately posterior to the mouth cavity, expands laterally as the pharynx. The lateral pockets of the pharyngeal cavity, called the pharyngeal pouches, perforate the mesodermal layer, reach the ectoderm, and break through to form pharyngeal, or gill, clefts (Figure 13). In fishes and larvae of amphibians, these clefts develop gills and become respiratory organs. Pharyngeal pouches develop in the early embryos of all vertebrates, including the air-breathing terrestrial reptiles, birds, and mammals. The number of pouches has been reduced in the course of evolution from six or more to four in tetrapods, and the posterior pouches may not actually break through.

The consistent development of pharyngeal pouches and clefts indicates their importance in vertebrate development. Many parts of the vertebrate body are derived from, or dependent on, the pharyngeal pouches; for example, the aortic arches—the most important blood vessels of a vertebrate—develop between successive pharyngeal pouches (Figure 14). Skeletal visceral arches also occur between consecutive pharyngeal pouches (they do not develop if the pharyngeal pouches are prevented from developing). In adult terrestrial vertebrates, parts of the visceral arches are transformed into the hyoid apparatus, supporting the tongue, the auditory ossicles, and parts of the larynx and trachea. Furthermore, some of the material of the pharyngeal pouches is utilized for the formation of the parathyroid glands and the thymus; the former are indispensable glands of internal secretion, and the latter are a source, in mammals, of cells that produce antibodies. The pharynx also produces the rudiment of the thyroid gland as a ventral outgrowth.

THE LIVER, PANCREAS, AND LUNGS

Three additional important organs develop from the endoderm: the liver, the pancreas, and the lungs. The liver develops as a ventral outgrowth of the endodermal gut just posterior to the section that will become the stomach. Initially, the liver takes the form of a tubular gland, but it soon acquires a close relationship to the blood sinuses and capillaries, forming lobules around blood vessels rather than around glandular ducts. The pancreas develops from three independent rudiments: two ventral ones, formed just posterior to the liver rudiment, and a dorsal one. The ventral and the dorsal rudiments fuse in most vertebrates to form one organ with a complicated system of ducts opening into the duodenum, a portion of the small intestine. The lungs develop from a ventral hollow outgrowth of the gut, just posterior to the pharyngeal region; the outgrowth branches into a right and left trunk that grow posteriorly beside the esophagus and then expand into hollow sacs, in lower terrestrial vertebrates, or into a system of tubes, in birds and mammals.

Development of lungs

The endodermal parts of the alimentary system are, along their entire length, encased by the splanchnic mesoderm of the lateral plates. The coelomic cavities of the right and left sides fuse ventral to the gut but remain separated dorsally by their respective walls, which form the dorsal mesentery—a double membrane by which the gut is suspended from the dorsal side of the body cavity and through which blood vessels and nerves reach the gut (Figure 11). The layer of splanchnic mesoderm next to the endoderm produces the connective tissue and muscular layers of the gut. During development of the glands of the alimentary canal (*e.g.*, pancreas, salivary glands), the mesoderm forms a connective tissue capsule around the branching tubules of the gland. The development of the tubules is dependent on this mesodermal capsule and cannot proceed without it.

VII. Postembryonic development

After partially developing within the egg membranes or within the maternal body, the newly formed individual emerges. The new animal is then born (ejected from the mother's body) or hatched from the egg. The condition of the new organism at the time of birth or hatching differs in various groups of animals, and even among ani-

mals within a particular group. In sea urchins, for example, the embryo emerges soon after fertilization, in the blastula stage. Covered with cilia, the sea-urchin blastula swims in the water and proceeds with gastrulation. Frog embryos emerge from the egg membranes when the main organs have already begun to develop, but functional differentiation of the tissues is unfinished; for instance, the components of the eyes and ears are far from complete, the mouth is not yet open, and the gut is filled with yolk-laden cells. Certain birds (called precocial) emerge from the egg covered with downy feathers and can run about soon after hatching, whereas others (altricial) hatch naked, with only rudiments of feathers, and are quite unable to move around. Among mammals there is a great range in the degree of development at birth. In marsupials, such as opossums and kangaroos, the young are born incompletely developed and very small; the young are then kept for a long time in the pouch of the mother, all the while firmly attached to the teats and suckling. Many small mammals are helpless at birth. Mice are born naked and blind; puppies and kittens are born covered with fur but with unopened eyes. Newborn human babies have their eyes open but cannot move themselves about for several months. Hoofed mammals, on the other hand, bear young that can stand up and run after their mothers within a few hours of birth.

Egg tooth of birds

In birds the hard shell is broken by the hatchling's beak, which is provided with a sharp tubercle on its top. A similar "egg tooth" appears on the tip of the snout of hatchling reptiles. Many arthropods have a preformed line of fragility that allows part of the eggshell to be burst open like a lid, allowing the young to emerge. Birth in mammals is effected through the contraction of smooth muscles of the uterus.

THE LARVAL PHASE AND METAMORPHOSIS

The organism emerging from the egg or from the maternal body, apart from being incompletely developed, may have an organization more or less different from that of an adult. In some cases the difference is so great that, without knowing the origin of the eggs or without following the young through their full course of development, it would be impossible to know that the young and the adult are of the same animal species. Such young, called larvae, transform into the adult form by a process of metamorphosis. The term larva also applies to young that resemble the adult form but differ from it in some substantial respect, as in possessing organs not present in the adult or in lacking an important structure (apart from sex glands and associated parts, which tend to develop later in life in most animals). Larvae in different animals have special names given to them, such as the tadpole of frogs, the caterpillar of butterflies, and the fry of fishes.

Advantages of larvae

The development of the embryo into a larva rather than directly into an organism similar to the adult has various advantages. At the time of emergence from the egg, the new individual is relatively small, and the organization that enables the adult to lead a particular mode of life may not be suitable for a miniature copy of the adult. The larva may have to procure food for itself and, being small, may not be able to feed in the same way as the adult. It also may not be able to use effectively the same defense mechanisms the adult possesses. The larval stage enables an animal to avoid such hazards; it provides a mode of life and corresponding organization better suited to the smaller size of the newly emerged organism. Another advantage is that the larva may be able to exploit an entirely different environment because its organization is very different from that of the adults. A terrestrial adult may have aquatic larvae, a flying adult may have burrowing larvae, and a parasitic adult may have a free-living larva. A third advantage of a larval stage emerges in animals whose adult stages are sessile or restricted in their movements; the larvae can move freely, either of their own accord or on water currents. In this way the larvae of sedentary animals serve for the dispersal of the species. Lastly, the larval stage is of great advantage for certain internal parasites, which, once inside a host, cannot transfer to another. New hosts are infected instead by the larval stages. (The usual means of attaining this end is for the parasite to produce enormous quantities of eggs and rely on the passive entry of the eggs into the new host with food. A more efficient way, however, is for a mobile larva to enter the new host actively.)

A large number of marine invertebrates possess floating larvae that have hairlike projections (cilia) as their means of locomotion. There are three main types of larvae, characteristic of large subdivisions of the animal kingdom.

The planula larva of coelenterates has an elongated shape and cilia covering its entire surface. The internal organization is simple, hardly beyond differentiation into ectoderm and endoderm in the interior. The larva does not feed but serves only for dispersal.

The trochophore larva is found in many marine invertebrates. Typically, as in polychaetes, it has an alimentary canal with mouth and anus and a ring of ciliated cells arranged anterior to the mouth. It also possesses a sensory organ and rudiments of mesoderm. Cilia around the mouth bring in food—unicellular plants and other small particles. The larva thus not only serves for dispersal but also feeds and grows before it transforms into an adult worm. Other trochophore larvae are found in marine mollusks and in certain marine worms. The larva of echinoderms is similar to the trochophore in possessing a gut and a ciliary band, but the arrangement of the latter is different. The echinoderm larva also feeds and grows as well as serves for dispersal.

Larvae of very different kinds are found in many arthropods. In crustaceans the larva, called nauplius, does not differ substantially in mode of life or means of locomotion from the adult but has fewer appendages than the adult. A typical crustacean nauplius has three pairs of legs and an unpaired simple eye. Additional pairs of appendages and paired compound eyes appear in the course of a sometimes prolonged development. In insects the larva differs from the adult by the absence of wings but, in addition, may have a different mode of life and different way of feeding. Among chordates the tunicates (sea squirts) deserve attention; the larval form is a free-swimming creature, showing unmistakable relation to vertebrates, but the adult is sedentary, with much reduced nervous and muscular systems. The tadpole of a frog differs from the adult in being totally aquatic, in possessing a tail and gills for respiration, and in having a mouth adapted for feeding on plants. The adult frog is adapted to land life, except for reproductive periods, has no tail and no gills, and is an active predator.

Types of metamorphosis

Metamorphosis, the transformation of the larva into an adult, is a more or less complicated process depending on the degree of difference between the two forms. The transformation may be gradual, extend over a long period, and involve a number of intermediate stages; alternatively, the transformation may be achieved in one step. In the latter case, especially if the difference between the larva and adult is great, large parts of the body of the larva, including all the specifically larval organs, disintegrate (necrobiotic metamorphosis). At the same time, organs of the adult are built up, sometimes from reserve groups of cells that remain undifferentiated or nonfunctional in the larva. A good illustration of the distinction between gradual and abrupt metamorphosis occurs among the insects. In more primitive insects, such as cockroaches and grasshoppers, metamorphosis is gradual. The larva, often referred to as a nymph, has more or less the same organization as the adult, or imago; it feeds in a similar way but differs from the adults in lacking wings and in having incomplete sex organs. The wings appear in later stages of larval life; they are small at first but increase with each molt, and they attain full size and functional capacity at the last (imaginal) one. The larva of other insects, such as beetles, butterflies, and wasps, is a grub or caterpillar, a wormlike creature not even remotely resembling the adult. The difference in organization is so profound that the transformation cannot be achieved gradually, and an intermediate resting, or pupal, stage is interposed between the larva and imago. The pupa neither feeds nor moves,

as the larval organs inside are destroyed and replaced with organs of the adult, including wings and sex organs. Eventually, when formation of the adult organs is complete, the pupal skin is cast off, and the adult emerges. The destruction of the larval parts may be far reaching and include even the skin and most of the alimentary canal. The tissues of the adult are formed from groups of reserve cells that were present all along in the larva as imaginal disks.

Necrobiotic metamorphosis is observed in the tunicate larva, in which the tail, including notochord, nerve cord, and muscles, and most of the brain, including eye and statocyst, are destroyed at the same time that the large pharyngeal cavity of the adult develops. A tadpole metamorphosing into an adult frog loses its tail—the cells of which are destroyed and devoured by phagocytic cells—its gills, and its larval mouthparts; concurrently the legs of the adult frog develop progressively, the structure of the mouth and alimentary canal change, and the skin acquires a bony (keratinized) layer and a system of subcutaneous glands (Figure 17).

Hormonal function in metamorphosis. The complicated changes taking place during metamorphosis, especially in the case of necrobiotic metamorphosis, must be performed in a coordinated way. So that no changes are made prematurely and no organ systems are left behind in the general transformation, some common signal for the change must be provided. For both insect and amphibian metamorphoses, which have been the most extensively studied, the signal is a hormonal one, sent in the blood to all the cells and tissues of the body.

Molting

Metamorphosis in an insect is complicated by the fact that the rigid cuticle covering its body is very restrictive; new features can appear only after a molt, when the old cuticle is replaced by a newly formed one. Molting in insects is caused by the action of two hormones. In the brain of insects, several groups of neurosecretory cells produce the first hormone. This brain hormone does not itself affect molting but stimulates the prothoracic gland, a loose mass of secretory cells situated in the thorax in close association with tracheal tubes. In response to the stimulation by the brain hormone, the prothoracic gland releases into the blood a second hormone, the molting hormone, or ecdysone. Under the influence of ecdysone, the tissues of the body produce a new cuticle under the old one, after which the old cuticle is shed (the actual molting). The new cuticle embodies any new developmental features that were scheduled to appear. The kind of feature that emerges after a molt is controlled by a third organ of internal secretion, the corpora allata, secretory tissue situated posterior to the brain, near or around the dorsal aorta. The corpora allata emit the juvenile hormone, which, as long as it circulates in the blood, acts to perpetuate the larval form. As the larva approaches the end of its development, however, the corpora allata stop producing juvenile hormone or reduce its quantity; whereupon, the larva, at the next molt, metamorphoses into an adult. Withdrawal of the juvenile hormone is the immediate cause of metamorphosis, in conjunction with the brain hormone and ecdysone, which are responsible for the shedding of the larval cuticle and for the production of the new cuticle embodying the

From E. Witschi, *Development of Vertebrates* (1956); W.B. Saunders Company, Philadelphia

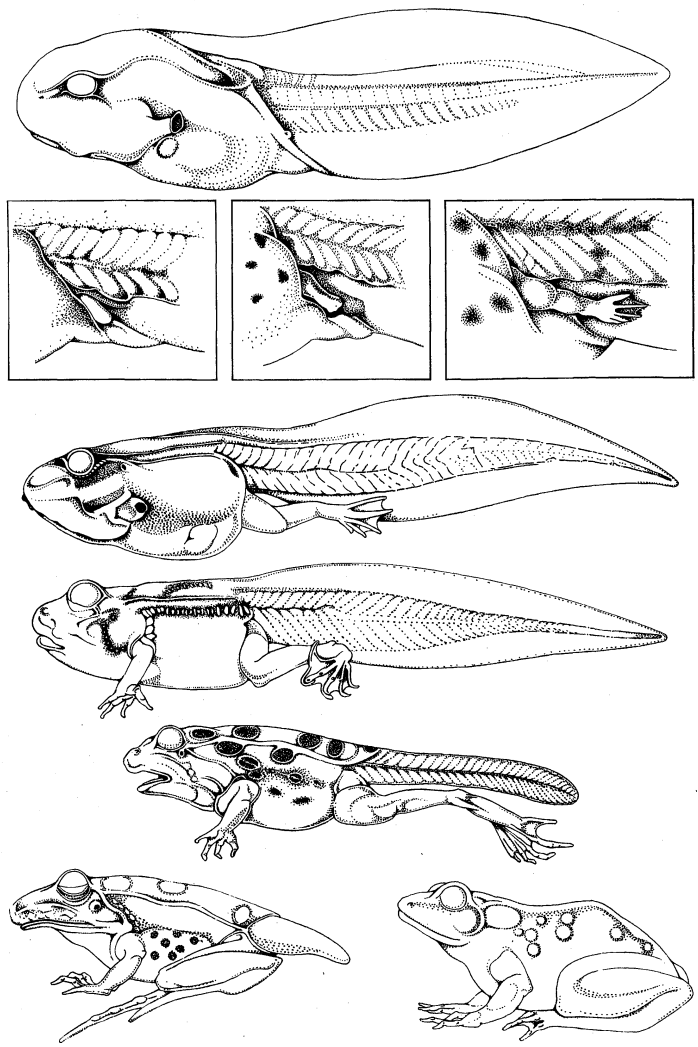


Figure 17: Stages in the metamorphosis of the frog from aquatic tadpole to terrestrial adult frog, including details of development of the hindleg.

features of the imago. Metamorphosis through the stage of the pupa is effected by diminishing levels of juvenile hormone, which determine first the transformation of the larva into a pupa and, with further reduction of the juvenile-hormone level, the final step of transformation of the pupa into the adult.

The metamorphosis of a tadpole into a frog also depends upon two hormones: one initiating the process and the other directly influencing the tissues involved in the change. The first hormone is the thyrotropic hormone, produced by the hypophysis. It has no immediate effect on the tissues of the body but activates the thyroid gland to produce several substances, the most important of which is thyroxine. Thyroxine and other iodine-containing compounds circulate in the blood and cause changes that, in their entirety, constitute the process of metamorphosis. It is remarkable that different tissues react in different ways to the presence of thyroxine. The muscles of the tadpole's tail degenerate, whereas the muscles of the trunk and legs are not affected; in fact, the growth and development of limbs are stimulated as a part of metamorphosis. The effect of the hormone depends on the nature of the reacting cells and tissues—*i.e.*, on their competence—just as the embryonic inductor in the earlier stages of development influences only cells with the competence for a particular kind of reaction.

DIRECT DEVELOPMENT

If an animal after birth or emergence from an egg differs from the adult in comparatively minor details (apart from not having functional sex organs), the development is said to be direct. There is no larval stage and no metamorphosis. Direct development does not mean, however, that no changes occur between birth and adulthood. One very obvious change is the growth of the animal.

Factors influencing rate of growth

The rate of growth—not absolute increase—is highest in the early stages of postembryonic life; subsequently, growth continues to slow, ceasing completely at the attainment of adulthood. The rate of growth is dependent on many factors, both external (feeding, temperature) and internal. Of the internal factors, the most important are hormones, especially the growth hormone produced by the hypophysis. If the growth hormone is produced in insufficient quantities, the result is dwarfism; if it is produced in excessive quantities, the result is gigantism.

In the case of direct development, the most important change is the attainment of sexual maturity, which is achieved in several steps and involves the action of several hormones. The gonad rudiments and rudiments of the supporting parts of the reproductive system remain inactive long after birth. At the approach of adulthood, however, two sets of hormones come into action: hypophyseal hormones stimulate the gonads to activity, and gonadal hormones (produced by the gonads) cause the supporting sex organs and other sex characters to become fully developed. To become functional, the gonads must be acted upon by secretions from the hypophysis. In immature females the follicle-stimulating hormone, which alone causes the egg follicles and the oocytes to grow, and the luteinizing hormone stimulate the follicle cells to produce the female sex hormone, estrogen, which effects the development of the uterus, the milk glands, and other characteristics of the female sex. In the male, the same hypophyseal hormones are produced, with the result that the testes start to produce sperm and to secrete the male sex hormone, androgen. It appears that the luteinizing hormone is the more active in the male sex, being able to cause both spermatogenesis and androgen secretion. Androgen, in turn, brings about the development of the penis, the descent of the testicles, the appearance of typical male hair growth, and other secondary sex characteristics.

VIII. Maturity and death

Sexual maturity and the ensuing reproductive activity mark the pinnacle of development and morphogenesis and, for many animals, herald the end of life. The biological goal of the entire process is achieved with the launching of the next generation, and the life cycle that runs from the formation of gametes by one generation to the formation of gametes by the next generation is completed. In many animals the females die after laying their eggs; the males may have died earlier, after pairing. Indeed, some males (spiders, praying mantises) are eaten by the females immediately after copulation.

The developmental period can only truly be said to end with the termination of an organism, for much activity continues to unfold new developmental sequences, not all of them progressive and favourable, to be sure. Senescence, or a decline in abilities, signals advancing age in mammals but is not a general occurrence in the animal kingdom. Far more animals continue to function at near-peak capacity well into old age. And even among those species—salmons, eels, many moths—whose members die after a single reproductive act, death is relatively swift and not accompanied by a prolonged period of deterioration.

In most animals the reproductive potential is not exhausted in a single act of gamete production, but the sexually mature individuals remain alive and reproduce repeatedly. In these cases life may extend long beyond the first attainment of reproductive ability and be accompanied by further growth of the individuals, as occurs in most fishes, amphibians, and reptiles, and also in mollusks and certain other invertebrates. In the case of prolonged life-spans, however, reproductive activity may cease with advancing age, and a senile involution take place, as is observed mainly in mammals and, particularly, in man. The changes taking place may be described as regressive development. In most animals, however, the end of life is not preceded by any overt traces of senility. As a general rule, then, the attainment of reproductive ability may be said to be the final phase of progressive development among animals.

A gradual loss of alertness and vigour is typical of the aging pattern of primates and is especially important to man. For additional information, the following articles will be helpful: LIFE-SPAN, for a consideration of the expectation of life for many organisms, including man; AGING, for the mental and physical changes that accompany the passage of time; and DEATH, for the culminative act of a lifetime of development.

BIBLIOGRAPHY. General textbooks covering the whole subject of animal development include: B.I. BALINSKY, *An Introduction to Embryology*, 3rd ed. (1970); C.W. BODEMER, *Modern Embryology* (1968); and F.G. GILCHRIST, *A Survey of Embryology* (1968).

Theoretical considerations of embryonic development are found in: H. SPEMANN, *Experimentelle Beiträge zu einer Theorie der Entwicklung* (1936; Eng. trans., *Embryonic Development and Induction*, 1938, reprinted 1962), a classic exposition of the experimental method in embryology; J.T. BONNER, *The Evolution of Development* (1958); J. BONNER, *The Molecular Biology of Development* (1965); and C.H. WADDINGTON, *Principles of Embryology* (1956).

Descriptions of the morphological aspects of embryonic development are found in: L.B. AREY, *Developmental Anatomy: A Textbook and Laboratory Manual of Embryology*, 7th ed. (1965); and B.M. PATTEN, *Foundations of Embryology*, 2nd ed. (1964).

Physiological and analytical studies of embryonic development are treated in: L.J. BARTH, *Development: Selected Topics* (1964); E.M. DEUCHAR, *Biochemical Aspects of Amphibian Development* (1966); R.A. FLICKINGER (comp.), *Developmental Biology* (1966); J. NEEDHAM, *Biochemistry and Morphogenesis* (1942); L. SAXEN and S. TOIVONEN, *Primary Embryonic Induction* (1962); and J.P. TRINKAUS, *Cells into Organs: The Forces that Shape the Embryo* (1969).

Specific treatments of embryonic adaptations in a wide range of animals are considered in: W.J. HAMILTON, J.D. BOYD, and H.W. MOSSMAN, *Human Embryology*, 3rd ed. (1962), deals with the prenatal development of form and function in man; O.E. NELSEN, *Comparative Embryology of the Vertebrates* (1953); and M. KUME and K. DAN (eds.), *Invertebrate Embryology* (1968; orig. pub. in Japanese, 1957).

Books dealing with metamorphosis and senescence include the following: W. ETKIN, "Metamorphosis," in J.A. MOORE (ed.), *Physiology of the Amphibia*, ch. 8 (1964); V.B. WIGGLESWORTH, *The Physiology of Insect Metamorphosis* (1954); and A. COMFORT, *Ageing: The Biology of Senescence*, rev. ed. (1964).

(B.I.B.)

Development, Biological

Development in its most general meaning refers to any process of progressive change. In this sense, most modern philosophical outlooks would consider that development of some kind or other characterizes all things, in both the physical and biological worlds. Such points of view go back to the very earliest days of philosophy.

GENERAL FEATURES

Philosophical basis. Among the pre-Socratic philosophers of Greek Ionia, half a millennium before Christ, some, like Heracleitus, believed that all natural things are constantly changing. In contrast, others, of whom Democritus is perhaps the prime example, suggested that the world is made up by the changing combinations of atoms, which themselves remain unaltered, not subject to change or development. The early period of post-Renaissance European science may be regarded as dominated by this latter atomistic view, which reached its fullest development in the period between Newton's laws of physics and Dalton's atomic theory of chemistry in the early 19th century. This outlook was never easily reconciled with the observations of biologists, and in the last hundred years a series of discoveries in the physical sciences have combined to swing opinion back toward the Heracleitan emphasis on the importance of process and development. The atom, which seemed so unalterable to Dalton, has proved to be divisible after all, and to maintain its identity only by processes of interaction between a number of component subatomic particles, which themselves must in certain aspects be regarded as processes rather than matter. Albert Einstein's theory of relativity showed that time and space are united in continuum, which implies that all things are involved in time; that is to say, in development.

The philosophers who charted the transition from the nondevelopmental view, for which time was an accidental and inessential element, were Henri Bergson and, in particular, Alfred North Whitehead. Karl Marx and Friedrich Engels, with their insistence on the difference between dialectical and mechanical materialism, may be regarded as other important innovators of this trend, although the generality of their philosophy was somewhat compromised by the political context in which it was placed and the rigidity with which their later followers have interpreted it.

Philosophies of the Heracleitan type, which emphasize process and development, provide much more appropriate frameworks for biology than do philosophies of the atomistic kind. Living organisms confront biologists with changes of various kinds, all of which could be regarded as in some sense developmental; however, biologists have found it convenient to distinguish the changes and to use the word development for only one of them. Biological development can be defined as the series of progressive, nonrepetitive changes that occur during the life history of an organism. The kernel of this definition is to contrast development with, on the one hand, the essentially repetitive chemical changes involved in the maintenance of the body, which constitute "metabolism," and on the other hand, with the longer term changes, which, while nonrepetitive, involve the sequence of several or many life histories, and which constitute evolution.

The blurring of distinctions in nature

As with most formal definitions, these distinctions cannot always be applied strictly to the real world. In the viruses, for instance, and even in bacteria, it is difficult to make a distinction between metabolism and development, since the metabolic activity of a virus particle consists of little more than the development of new virus particles. In certain other cases, the distinction between development and evolution becomes blurred: the concept of an individual organism with a definite life history may be very difficult to apply in plants that reproduce by vegetative division, the breaking off of a part that can grow into another complete plant. The possibilities for debate that arise in these special cases, however, do not in any way invalidate the general usefulness of the distinctions as conventionally made in biology.

The scope of development. All organisms, including the very simplest, consist of two components, distinguished by a German biologist, August Weismann, at the end of the 19th century, as the "germ plasm" and the "soma." The germ plasm consists of the essential elements, or genes, passed on from one generation to the next, and the soma consists of the body that may be produced as the organism develops (Figure 1). In more

From *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed. (1968); Rand McNally & Company

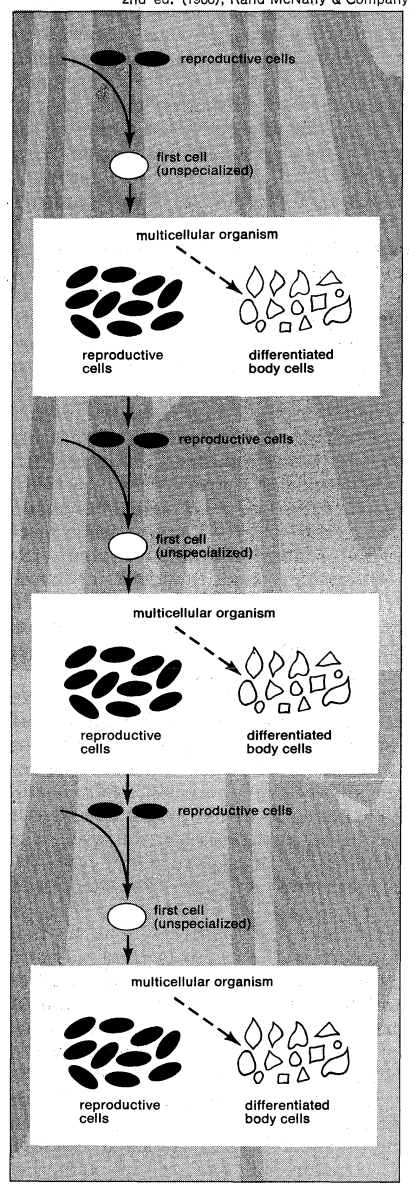


Figure 1: Weismann's concept of the continuity of the germ plasm.

modern terms, Weismann's germ plasm is identified with DNA (deoxyribonucleic acid), which carries, encoded in the complex structure of its molecule, the instructions necessary for the synthesis of the other compounds of the organism and their assembly into the appropriate structures. It is this whole collection of other compounds (proteins, fats, carbohydrates, and others) and their arrangement as a metabolically functioning organism that constitutes the soma. Biological development encompasses, therefore, all the processes concerned with implementing the instructions contained in the DNA. Those instructions can only be carried out by an appropriate executive machinery, the first phase of which is provided by the cell that carries the DNA into the next generation: in animals and plants by the fertilized egg cell; in viruses by the cell infected. In life histories that have more than

a minimal degree of complexity, the executive machinery itself becomes modified as the genetic instructions are gradually put into operation, and new mechanisms of protein synthesis are brought into functional condition. The fundamental problem of developmental biology is to understand the interplay between the genetic instructions and the mechanisms by which those instructions are carried out.

Genotype versus phenotype

In the language of genetics the word genotype is used to indicate the hereditary instructions passed on from one generation to another in the genes, while phenotype is the term given to the functioning organisms produced by those instructions. Biological development, therefore, consists in the production of phenotypes. The point made in the last paragraph is that the formation of the phenotype of one generation depends on the functioning of part of the phenotype of the previous generation (*e.g.*, egg cell), as the mechanism that begins the interpretation of the instructions contained in the new organism's genotype.

Types of development. In the entire realm of organisms, many different modes of development are found, the most important categories of which can be discussed as pairs of contrasting types.

Quantitative and qualitative development. Development may amount to no more than a quantitative change (usually an increase) in a system that remains essentially unaltered. Qualitative development involves an alteration in the nature of the system. Pure examples of the first type are difficult to find. Approximations to it occur when an animal or plant has attained a structure with the full complement of organs; it then appears to increase only in size, that is to say, quantitatively. This would be a period of simple growth. A closer examination nearly always shows that the system is also undergoing some qualitative change, however. A human infant at birth, for example, already has its full complement of organs, but the ensuing developmental period up to adulthood involves not only growth but also processes of maturation that involve qualitative as well as quantitative changes. Perhaps the most uncomplicated examples of quantitative development occur in certain simple plants and animals. Flatworms, for example, may become reduced in size when starved but increase in size again when provided with suitable nutrition; they thus undergo quantitative changes. Even in these cases, however, it is found the constituent organs do not always merely become reduced in size but may actually suffer the loss of certain parts.

Progressive and regressive development. The normal processes of development in the majority of plants and animals may be considered progressive since they lead to increases in size and complexity and to the addition of new elements to the system. As already indicated, some organisms, when placed in adverse conditions, may undergo regressive changes, both in size and complexity.

The adaptability of regression

Such regressive changes are a part of the normal life history of certain organisms. Characteristically, these are species in which the organism at an early stage develops a relatively complex structure that enables it to be motile, and later adopts a form of life for which motility is no longer a necessity. A good example is that of the barnacles, a group of marine crustaceans in which the egg at first develops into a motile larva that soon settles down and becomes firmly attached to a solid underwater surface. The barnacle then loses many of the organs characteristic of the motile phase and develops into its familiar stationary form.

There are a number of other examples, particularly in groups in which the adults adopt a parasitic form of life, especially within the digestive system or other tissues of a host animal, from which they have only to absorb their nutriment without having to move or to possess suitable organs for capturing prey. In such cases the early developmental period is characterized by progression toward more complex forms followed by a period of regression in which many of these organs may be lost. During this regressive period certain components of the organism (*i.e.*, those concerned with functioning as a sessile or parasitic form) may undergo progressive development at the same time as the other organs are regressing.

Single-phase and multiphase development. The most familiar organisms, including man, undergo a single-phase development; the organs that appear at early stages persist throughout the whole of life. There are many kinds of animals that develop one or more larval stages adapted to a life different from that of the adult. Perhaps the best known of these is the common frog. The egg first develops into a tadpole, which is provided with a large muscular tail by which it swims. The tadpole eventually undergoes a change of form, or metamorphosis. This involves the regression and resorption of the tail and the growth of the limbs. During this time the rest of the body of the tadpole undergoes less profound changes; the organs persist but undergo relatively far-reaching progressive changes. In other animals, the alteration between the larval and the adult forms may be much more drastic. The egg of a sea urchin, for instance, at first develops to a small larva (the pluteus), which is completely unlike that of the adult. During metamorphosis nearly all the structures of the pluteus disappear; the five-rayed adult develops from a very small rudiment within the larva. In other groups of marine invertebrates, there may be successive larval stages before the adult form appears.

Alternation of generations

Plants in general appear to exhibit a type of development related in a general way to the multiphased development just discussed in animals, although rather different from it in essence. This is called the "alternation of generations" (Figure 2). The majority of higher plants

From H.A. Moore and J.R. Carlock, *The Spectrum of Life;* copyright © 1970 by Harper & Row, Publishers, Inc.

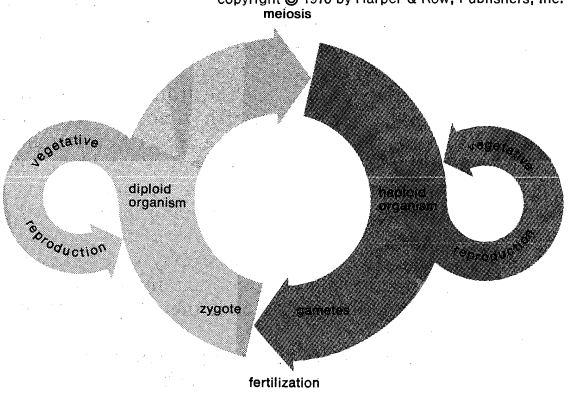


Figure 2: Generalized life cycle showing alternation of generations.

possess two sets of similar chromosomes in each of their cells, that is to say they are diploid (2n), as are most higher animals. But in sexual reproduction, diploid cells undergo a reduction division so as to form precursors of the sex cells, which are haploid—*i.e.*, they contain only one set of chromosomes. In animals these cells develop directly into the sex cells—egg and sperm—which unite in fertilization. In plants the haploid cells undergo some developmental processes before the functioning sex cells are produced. The products of this development are spoken of as the "haploid generation." In most higher plants the haploid development is quite reduced, so that the haploid individuals contain only a few nuclei—those associated with the pollen tube on the male side and a few associated with the egg on the female side. In some lower plants, however, such as mosses and ferns, the haploid development may be much more extensive and give rise to quite sizable separate plants. In such cases a species contains two kinds of individuals, produced by different types of developmental processes controlled, however, by the same genotype. This may be compared with the multiphasing development of larval forms in animals. The situation in plants, however, is characterized by the two forms of the organism having different chromosomal constitutions—haploid and diploid—whereas the larval forms and the adult of an animal species have the same chromosomal constitution.

Structural and functional development. These two categories cannot be regarded as a pair of opposites as were the previous pairs in this list; rather, they are two aspects of all processes of biological development and can be separated only conceptually, and for purposes of convenience of description. Function is the capacity of the biological system to carry out operations. At the level of the organism, these operations include walking, swimming, eating, digesting, etc.; at the cell level, typical functions are respiring, contracting, conducting nervous impulses, secreting hormones, etc.; and at the molecular level, all functions depend on the production of enzymes, coded by particular genes. Structure encompasses all parts of the organism capable of carrying out functions localized within the body of the organism and arranged in some particular spatial pattern. Contractile cells, for example, are grouped together to form muscle, and other cells are grouped together to form elements of the skeleton; both the muscles and the skeletal elements have definite spatial relations to each other.

These two aspects of development—function and structure—are not opposed to each other in any way. On the contrary, it is obvious that the higher level functions are clearly dependent on the proper structural relations and functions of cell systems. Even at the basic cellular or molecular levels, secretion or nervous conduction essentially depends upon the proper structural relation of the subcellular elements. It is, however, often convenient to focus discussion on one or other of these two aspects of development; for instance, a study may be made into the developmental processes that bring about the production of hemoglobin or insulin by a certain kind of cell, without at the moment being concerned with structural problems. Or again, the focus may be on the results of a certain process by which a mass of cells develops into a typical hand with five digits. In such an inquiry the structural aspects are paramount.

Normal and abnormal development. If a number of fertilized eggs of a given species are provided with conditions that enable them to develop at all, they will, with extraordinary regularity, develop into exceedingly similar adult organisms. The range of conditions they can tolerate is rather wide, and the similarity of the end products surprisingly complete. There are, indeed, good grounds for recognizing what must be considered normal development. The situation is perhaps more marked in animals than in plants, since the plants produced from a given batch of seed under a variety of environmental conditions often present considerably greater variation than is commonly found among animals. Even among plants, however, the differences produced by different conditions of cultivation are usually no more than quantitative differences in size and number of such organs as leaves and flowers, so that an individual can be described as well or poorly developed rather than as normally or abnormally developed. It is only in relatively few cases that a plant develops in quite different ways under two different conditions, neither of which can be considered abnormal or normal. In certain aquatic plants, for instance, the shape of the emergent leaves is different from the leaves that develop underwater. In such cases the plant actually has two normal forms of development.

The latitude of normal development

It is possible, of course, to produce abnormal organisms by submitting a developing system to stimuli not usually encountered in normal environment, such as certain chemicals. The presence of unusual genes also may result in deviations from the normal processes of development. In the vast majority of cases such abnormalities can be regarded as resulting from failure to carry out fully the normal processes of development. Functional abnormality in the adult consists in the failure of the system to produce a certain enzyme or functional cell type; a structural abnormality consists in the unusual appearance of certain component elements or in their arrangement in incompletely realized patterns. It is extremely rare to find examples in which the abnormality consists in the addition of a new enzyme not produced in normal development, or the formation of a new structural pattern of the elements.

One very important type of development that, from some points of view, can be considered as an exception to the rule that abnormal development is nearly always retrogressive, is carcinogenesis, the production of tumours. Carcinogenesis involves a change in the developmental behaviour of a group of cells. Initially, it often involves a loss of some of the functional and structural characteristics that previously appeared in the cells. It is commonly followed, however, by the assumption of new properties, which however untoward they may be for the host animal, must be considered as a progressive type: the cells often grow faster and multiply sooner than the noncancerous cells, for example. Furthermore, the cells may undergo a sequence of changes in character and in their arrangement within the tumour. All these features can be regarded in a developmental sense as progressive.

Cancerous growth as progressive

In view of the great rarity of cases of abnormal development that lead to progressive changes, it seems to follow that the organs produced during the normal development of any given species actually exhaust all the potentialities of its genotype for the production of orderly functional structures. It appears that the only abnormal developments that can be produced are either displacements of normal organs, or inadequacies in carrying out normal processes, or the initiation of progressive but quite disorderly processes, as in the production of tumours (see MALFORMATION, BIOLOGICAL).

GENERAL SYSTEMS OF DEVELOPMENT

Development of single-celled organisms. In viruses, activities consist in the production, aided by the machinery of a host cell, of units for building new virus or phage particles: development is simply the assemblage of these constituent units.

In the next higher grade of biological organization, the organism consists of a single cell. Many single-celled algae produce special forms of cells that correspond to the sex cells, or gametes; these cells may unite in fertilization, the resulting fertilized egg, or zygote, undergoing a short period of development. In many other single-celled organisms, however, reproduction takes place by the simple division of an original cell into two daughter cells. In such forms, development normally is part of the process of subdivision. It involves the remodelling of the parent cell into two smaller cells, which are then separated by the division. Something similar must, of course, be involved in the division of cells of higher organisms also. In many single-celled organisms, however, the cell contains a number of defined parts, which are arranged in very definite ways, so that the process of remodelling is very striking and easily observed. This is so, for instance, with ciliated protozoans, in which the cortex is provided with a large number of hairlike cilia or other appendages, arranged in precise patterns, and often with such other structures as a mouth or a gullet. These structures are reproduced in two identical but smaller copies during cell division. This does not necessarily imply that no other developmental processes are possible. The process of regeneration of parts removed occurs quite independently of cell division, for example.

Open and closed systems of development. There is a marked difference between the general system of development in multicellular plants and multicellular animals. In a plant, certain groups of cells retain throughout the whole life of the plant an embryonic capability to give rise to many types of cells. These regions, known as meristems, occur at the growing tips of branches and roots and as a cylindrical sheath around the stem. They consist of rapidly dividing cells capable of assembling into groups that form buds from which may arise new stems, leaves, flowers, or roots.

Continued embryonic development in plants

By contrast, most animals have no special regions that retain an embryonic character. In most forms, the whole egg, and the whole collection of cells immediately derived from it, take part in the developmental processes and form parts of the developing embryo. In some forms that go through a number of larval stages, the development of certain cells is interrupted at an early stage, and they are set aside and resume their development to form

a later type of larva, or to form the adult after the larval stages are completed. An example would be the imaginal buds of some insects. The cells of these buds cannot be regarded as retaining a fully embryonic character comparable to that of the plant meristems, since they cannot perform all the developmental processes but only those involved in the production of the particular late-larval or adult structure for which they have been set aside. In general, then, plants remain embryonic in character, capable as it were of starting again from the beginning to carry out the entire developmental process. Their development is, in this sense, "open." Most animals, on the other hand, lack persistently embryonic cells of this kind, and their development may be characterized as "closed." (There may be certain exceptions to this in very simple forms, such as flatworms, in which certain cells called neoblasts seem able to participate in any type of development; these cells are usually scattered throughout the body, and the major developmental processes that bring into being the general form of the organism cannot be attributed to them, as the development of the plant can be attributed to the meristems.)

Blastogenesis versus embryogenesis. Some animals possess a second system of development, in contrast to the "closed" embryonic system emphasized in the last section. In its most fully developed form, this system consists in remodelling a portion of the parental body into a new organism without any involvement of eggs or sperm. In an adult hydra, a microscopic aquatic animal, a portion of the body may begin to grow exceptionally fast; its cells differentiate into the various cell types and become molded into the constituent organs to build up a new individual identical to the parent. The group of cells responsible for this behaviour is, in its early stages, referred to as a bud, or blastema. Before they become activated these cells may appear quite indistinguishable from the other cells of the body and betray no embryonic capability comparable to the meristems of plants.

In some higher organisms, including certain insects, reptiles, and amphibians, incomplete but still fairly extensive new developments of a similar kind may take place. They require the stimulus of an injury, however, which may involve the removal of part of the normal body. The usual result is a new development to regenerate, or replace, the missing part. The first stage in such regenerative processes consists in the formation of a blastema, that is, a group of rapidly dividing cells that shows little sign of cellular specialization. The evidence indicates that they may not arise, as was once thought, from persisting embryonic cells scattered within the adult body, but instead are formed of cells near the position of the injury. These cells lose their normal adult character and become capable of developing into most of the tissues required to replace the parts removed by the injury (see REGENERATION, BIOLOGICAL).

The developmental aspect of regeneration

Development from a blastema, or blastogenesis, presents many contrasts to embryogenesis, the normal form of development from a fertilized egg. In blastogenesis, tissues that, during embryonic development, appear in sequence one after another, may be formed simultaneously and without any obvious sequential relations. Very little, however, is as yet understood about the mechanisms by which the various tissues within the blastema become differentiated from one another. It may well be that these mechanisms are more similar to those found in embryonic development than appears at first sight.

CONSTITUENT PROCESSES OF DEVELOPMENT

Growth. As was pointed out earlier, developing systems normally increase in size, at least during part of their development. "Growth" is a general term used to cover this phenomenon. It comprises two main aspects: (1) increase in cell numbers by cell division and (2) increase in cell size. These two processes may in some examples occur quite separately from each other; for instance, cells in certain rapidly growing tissues (*e.g.*, the connective tissue or blood-forming systems in vertebrates) may increase greatly in number, while the cells remain approximately the same size. Alternatively, in some organs (*e.g.*, the salivary glands of insects) the cells may increase greatly while remaining the same in number, each cell becoming enlarged, or hypertrophied. In such greatly enlarged cells there is often duplication of the genes, involving an increase in the DNA content of the nucleus, although no cell division takes place, and the nucleus continues as a single body, although with a multiplied, or "polyploid," set of chromosomes.

In very many cases, however, the growth of an organ depends on increases both in cell number and in cell size. The relative importance of these two processes has yet to be properly investigated. One case that has been well studied is the size of the wings of the fruit fly *Drosophila*. The number of cells in the wing can be easily determined, since each bears a single hair that can be seen and counted in simple microscopic preparations. It has been found that there is an accommodation of factors: if there is an ususually large number of cells, these may be somewhat smaller than usual, so that the total size of the wing remains relatively unchanged.

Perhaps the major theoretical difficulty in the concept of growth is that it is a quantitative notion attached to an ill-defined entity. Growth is an increase in size; but size of what? If a cell or organ increases in volume merely by the absorption of water, or by the laying down of a mineral substance such as calcium carbonate, is this to be regarded as growth or not?

Morphogenesis. As was pointed out earlier, morphogenesis refers to all those processes by which parts of a developing system come to have a definite shape or to occupy particular relative positions in space. It may be regarded as the architecture of development. Morphogenetic processes involve the movement of parts of the developing system from one place to another in space, and therefore involve the action of physical forces, in contrast to processes of differentiation (see below), which require only chemical operations. Although in practice the physical and chemical processes of development normally proceed in close connection, for purposes of discussion it is often convenient to make an artificial separation between them.

There is an enormous variety of different kinds of structures within living organisms. They occur at all levels of size, from an elephant's trunk to organelles within a cell, visible only with the electron microscope. There is still no satisfactory classification of the great range of processes by which these structures are brought into being. The following paragraphs constitute a tentative categorization that seems appropriate for the present state of biological thought on this topic.

Morphogenesis by differential growth. After their initiation, the various organs and regions of an organism may increase in size at different rates. Such processes of differential growth will change the overall shape of the body in which they occur. Processes of this kind take place very commonly in animals, particularly in the later stages of development. They are of major importance in the morphogenesis of plants, where the overall shape of the plant, the shape of individual leaves, and so on, depends primarily on the rates of growth of such component elements as the stems, the lateral shoots, and the vein and intervein material in leaves. In both animals and plants, such growth processes are greatly influenced by a variety of hormones. It is probable that factors internal to individual cells also always play a role.

Although differential growth may produce striking alterations in the general shape of organisms, these effects should probably be considered as somewhat superficial, since they only modify a basic pattern laid down by other processes. In a plant, for instance, the fundamental pattern is determined by the arrangement of the lateral buds around the central growing stem; whether these buds then grow fast or slowly relative to the stem is a secondary matter, however striking its results may be.

Morphogenetic fields. Many fundamental processes of pattern formation (*e.g.*, the arrangement of lateral buds in growing plants) occur within areas or three-dimensional masses of tissue that show no obvious indications of where the various elements in the pattern will arise

The appearance of pattern

until they actually appear. Such masses of tissue, in which a pattern appears, have been spoken of as "fields." This word was originally used in the early years of the 20th century by German authors who suggested an analogy between biological morphogenetic fields and such physical entities as magnetic or electromagnetic fields. The biological field is a description, but not an explanation, of the way in which the developing system behaves. The system develops as though each cell or subunit within it possessed "positional information" that specifies its location within the field and a set of instructions that lays down the developmental behaviour appropriate to each position.

There have been several attempts to account for the nature of the positional information and of the corresponding instructions. The oldest and best known of these is the gradient hypothesis. In many fields there is some region that is in some way "dominant," so that the field appears as though organized around it. It is suggested that this region has a high concentration of some substance or activity, which falls off in a graded way throughout the rest of the field. The main deficiency of the hypothesis is that no one has yet succeeded in identifying satisfactorily the variables distributed in the gradients. Attempts to suppose that they are gradients of metabolic activity have, on investigation, always run into difficulties that can only be solved by defining metabolic activity in terms that reduce the hypothesis to a circular one in which metabolic activity is defined as that which is distributed in the gradient.

Recently, a new suggestion has been advanced concerning position information. Most processes within cells normally involve negative feedback control systems. These systems have a tendency to oscillate, or fluctuate regularly (in fact, any aspect of cell metabolism may be basically oscillatory in character). The cycle of cell growth and division may be only one example of a much more widespread phenomenon. The substances involved in these oscillations are likely to include diffusible molecules capable of influencing the behaviour of nearby cells. It is easy to envisage the possibility that there might be localized regions with oscillations of higher frequency or greater amplitude that act as centres from which trains of waves are radiated in all directions. It has been suggested that positional information is specified in terms of differences in phase between two or more such trains of transmitted oscillations.

Certain types of field phenomenon may involve an amplification of stochastic (random) variations. In systems containing a number of substances, with certain suitable rates of reaction and diffusion, chance variation on either side of an initial condition of equilibrium may become amplified both in amplitude and in the area involved. In this way, the processes may give rise to a pattern of differentiated areas, distributed in arrangements that depend on the boundary conditions.

Morphogenesis by the self-assembly of units. Complex structures may arise from the interaction between units that have characteristics such that they can fit together in a certain way. This is particularly appropriate for morphogenesis at the simple level of molecules or cells. Units such as the atoms of carbon, hydrogen, oxygen, nitrogen, and so on, can assemble themselves into orderly molecular structures, and larger molecules, such as those of tropocollagen, or protein subunits in general, can assemble themselves into complexes whose structure is dependent on localized and directional intermolecular forces. It seems that such comparatively large entities as the units that come together to form the head structures of bacteriophages or bacterial flagella are capable of orderly self-assembly, but the chemical forces that give rise to the interunit bonds are still little understood.

Molecular assembly

Processes that fall into the same general category as self-assembly may occur within aggregates of cells. The units that self-assemble are the cells themselves. Interaction and aggregation may be allowed to occur in assemblages of cells of one or more different kinds. In such cases it is commonly found that the originally isolated cells tend to adhere to one another, at first more or less at random and independently of their character, but later they become rearranged into a number of regions consisting of cells of a single kind. When the cells in the initial collection differ in two different characteristics, for instance in species and organ of origin, the assortment in some cases brings together cells from the same organ, in other cases cells from the same species. Mixtures of chick and mouse cells, for instance, reassort themselves into groups derived from the same organ, whereas cells from two different species of amphibia sort out into groups from the same species more or less independently of organ type.

This morphogenetic process probably has only a restricted application to the formation of structures in normal development, in which only in a few tissues (*e.g.*, the connective system) do cells ever pass through a free stage in which they are not in intimate contact with other cells, and cells of different origin do not normally become intermingled so as to call for processes of reassortment. To explain normal morphogenetic processes of plants and animals one must look to the results that can be produced by the differential behaviour of cells that remain in constant close contact with one another. Several authors have shown how striking morphogenetic changes could be produced within a mass of cells that remain in contact, but that undergo changes in the intensity of adhesion between neighbouring cells, in the area of surface in the proportion to cell volume, and so on.

Differentiation. Differentiation is simply the process of becoming different. If, in connection with biological development, morphogenesis is set aside as a component for separate consideration, there are two distinct types of differentiation. In the first type, a part of a developing system will change in character as time passes; for instance, a part of the mesoderm, starting as embryonic cells with little internal features, gradually develops striated myofilaments, and with a lapse of time develops into a fully formed muscle fibre. In the second type, space rather than time is considered; for instance, other cells within the same mass of embryonic mesoderm may start to lay down an external matrix around them and eventually develop into cartilage. In development, differentiation in time involves the production of the characteristic features of the adult tissues, and is referred to as histogenesis. Differentiation in space involves an initially similar (homogeneous) mass of tissue becoming separated into different regions and is referred to as regionalization.

Tissues and regions

Histogenesis involves the synthesis of a number of new protein species according to an appropriate timetable. The most easily characterized are those proteins formed in a relatively late stage of histogenesis, such as myosin and actin in muscle cells. The synthesis of proteins is under the control of genes, and the problem of histogenesis essentially reduces to that of the genetic mechanisms that direct protein synthesis.

Regionalization is concerned with the appearance of differences between various parts of what is at first a homogeneous, or nearly homogeneous, mass. It is a prelude to histogenesis, which then proceeds in various directions in the different regions so demarcated. The processes by which the different regions acquire distinct contrasting characteristics must be related to some of the processes discussed under morphogenesis. Unlike morphogenesis, regionalization need not involve any change in the overall spatial shape of the tissues undergoing it. Regionalization falls rather into the type of process for which field theories have been invoked.

CONTROL AND INTEGRATION OF DEVELOPMENT

Phenomenological aspects. One of the most striking characteristics of all developmental systems is a tendency to produce a normal end result in spite of injuries or abnormalities that may have affected the system in earlier stages. In many cases, perhaps in most, only injuries inflicted during a certain restricted period of development can be fully compensated for. During such periods the system is said to be capable of regulation or the restoration of normality.

Developmental regulation is often discussed in terms of homeostasis, or regulatory mechanisms. Many systems, including biological ones, exhibit a tendency to return to initial equilibrium once it is diverted from it. A developing system is, by definition, always changing in time, moving along some defined time trajectory, from an initial stage, such as a fertilized egg, through various larval stages to adulthood, and finally to senescence. The regulation that occurs in such systems is a regulation not back to an initial stable equilibrium, as in homeostasis, but to some future stretch of the time trajectory. The appropriate word to describe this process is homeorhesis, which means the restoration of a flow.

A second major phenomenological characteristic of development is that the end state attained is not unitary but can be analyzed into a number of different organs and tissues. The overall time trajectory of this system can, therefore, also be analyzed into a number of component trajectories, each leading to one or another of the end products that can be distinguished in the later stages. A major discovery of the early experiments on developing systems was that, in many cases at least, the different time trajectories diverge from one another relatively suddenly during some short period of development, which usually occurs well before any visible signs of divergence can be seen microscopically or by any other available means of analysis. The most dramatic and influential example of this was provided by studies on the development of the amphibian egg at the time of gastrulation, or formation of a hollow ball of cells. At this time the lower hemisphere of the embryo will be pushed inward (invaginated) to develop into the mesoderm and endoderm, and the upper hemisphere will remain on the surface, expanding in area to cover the whole embryo. Approximately one-third of the upper hemisphere will develop into the nervous system and the remainder into the skin. During the period when these morphogenetic movements of invagination and expansion are occurring, a process takes place by which a portion of the upper hemisphere enters a trajectory toward neural tissue and another part enters a trajectory leading to epidermal development.

Determination

This process of determination of developmental pathways happens relatively quickly, during a period when the cells of the two different regions appear superficially alike. The occurrence of the determination can in fact be demonstrated only experimentally. Before it occurs, any part of the hemisphere can develop either into neural tissue or into skin. After it has happened, each part can develop only into one or the other of these alternatives.

It is clear than an adequate theory of development has to account not only for the processes by which a developing system moves along its appropriate time trajectory, but also for the nature of the processes by which the trajectories diverge from one another and become fixed or determined in the developing cells.

The determined state can be transmitted through many cell generations. An example of this transmission can be seen in *Drosophila* flies. The imaginal buds of *Drosophila* are small packets of cells that become separated from the main body of the embryo in the early stages of development. They persist throughout larval life and then enter into the differentiation of adult characteristics when stimulated to do so by the hormones secreted at the time of pupation. These pupation hormones disappear from the body of the adult insect, and imaginal buds transplanted into the body cavity of an adult undergo many cell generations, but they do not show any signs of differentiating into the specific tissues of the corresponding adult organ. After many generations of proliferation, however, the cells can be transplanted back into a larva ready to pupate; they thus submit to the pupation hormones and differentiation occurs. Through many generations of proliferation the cells have retained the determination as to which adult organ they will develop into when the pupation hormones become available.

Attempts to identify the determining agent have not yet been successful. Experiments on amphibian eggs, however, have given rise to one important general conclusion; namely, that the process of determination can take place only during a certain period of development, in which the cells of the upper half of the amphibian egg are poised between the two alternatives of development into neural tissue or into skin. They are said at this time to be "competent" for one or the other of these types of development. While they are in this state, and only while they are in it, a variety of external agents can switch them into one or the other of the possible pathways. Such a situation may be contrasted with one in which the cells were neutral, or featureless, and required then an external agent to transmit to them the quality of becoming nervous tissue or of becoming skin. This would mean that the reacting cells required information or instructions to be added to them from outside. Such a situation is not characteristic of biological development. Both in highly developed organisms such as amphibians and in simpler ones such as bacteria, the external agents act only as a releaser that switches on one or another process for which all of the necessary information is already incorporated in the cells concerned.

Analytical aspects. The existence of these developmental phenomena was realized in the first third of this century. During this period, biologists had no clear notion of the fundamental concepts needed to explain development. Developmental biologists, or embryologists, attempted to account for their observations by means of ill-defined notions, such as "potencies" or "organ-forming substances," or by referring to cellular properties that are real enough but obviously in themselves complex and essentially secondary in nature, such as cellular adhesiveness, the capacity of cell surfaces to differentially absorb certain substances, and so on. It was only gradually that developmental biologists came to realize the importance of the demonstration by genetics that nearly all the instructions required for the building of a new organism are contained in the genes that come together during fertilization, and that the small additional amount of information, contained primarily in the ovum, is itself a product of genetic instructions provided in the body of the mother in which the ovum is produced. The fundamental problems of the theory of development are, therefore, to understand how these units interact with one another to form more complex mechanisms that bring about the cellular or tissue behaviours of the different types of developing systems.

The genes as repositories of developmental instructions

In the development of the neural system of vertebrates, for example, a great many genes must be active in controlling the synthesis of particular proteins. In the formation of the wing of a *Drosophila*, the activity of some 20 or 30 genes has been definitely demonstrated, and certainly many more are involved. The action of all these genes, however, must be considered to form a network involving many types of feedback and other interactive loops, the overall result of which is a product in which many components are present in precisely defined concentrations; and further, the developmental process leading to this end result must be buffered or stabilized, in the sense that if the process is diverted from its normal course at an early stage, it returns to some later stage of the normal trajectory. Such a buffered time trajectory has been called a chreod, a word meaning "a necessary path." The realization that the basic units of development are genes and that a chreod involves the action of tens, if not hundreds, of genes forces one to conceive of chreods as structures within a multidimensional space. The realization that biological development is fundamentally an expression of the controlled activities of genes has finally resolved one of the old philosophical controversies about the nature of development, between preformation and epigenesis. The former supposed that, at the initiation of development, for instance in the fertilized egg, the system already contained some representative of every organ that would eventually put in an appearance. The vindicated theory of epigenesis, on the other hand, supposed that later appearing entities were produced during the course of development.

The modern interpretation of epigenesis is that the initial stage of development does contain certain entities with well-defined properties, namely the genes. These do

not, however, represent directly the later formed organs, which arise by the gradual interaction and progressive unfolding of the properties of groups of genes.

One of the major problems confronting modern developmental biology—namely, the nature of "determination"—requires an understanding of how genes are "primed" to enter into activity when an appropriate stimulus is given. The state of priming presumably has to apply to quite a large number of genes, though perhaps not to all that will be involved in the chreod, since some may be brought into activity by the operation of the earlier active ones. The priming, moreover, has to be able to persist through cell division and be capable of transmission through many generations of cell proliferation. Few concrete suggestions as to the mechanism have yet been made. One is that the primed genes are already producing the ribonucleic acid molecules, called messenger RNA's, which direct protein synthesis in the cell, but that these messengers are in some way inactivated or prevented from activating the protein-synthesizing machinery; this is known as the "masked messenger" hypothesis. Arguments in favour of this hypothesis are, however, circumstantial rather than direct. In some cases, for instance that of the *Drosophila* imaginal buds, there is direct evidence against it. Another hypothesis, perhaps more attractive, but much vaguer, is that the determination or priming involves the intervention of some of the large amounts of reiterated DNA known to be present in the cells of higher organisms. At the present time, however, biology lacks any convincing theory of determination in terms of gene action.

The "masked messenger" hypothesis

It appears at first sight that more is known about actual differentiation than initial determination. Actual differentiation must involve the controlled synthesis of particular proteins, coded for by specific genes. Certainly, a great deal is known about the mechanisms that control the action of genes in directing the synthesis of proteins in simple organisms such as viruses and bacteria. It is tempting to suppose that similar systems operate in controlling the synthetic activities of genes in higher organisms. Unfortunately, no single case of an exactly similar controlling system has ever been discovered in higher organisms, in spite of an intense search for it. It may in fact be suggested that until there is a fuller understanding of the mechanism of "priming" genes at the time of determination, there can scarcely be an adequate account of the way in which the activity of these genes is controlled at later stages.

DEVELOPMENT AND EVOLUTION

Evolution is carried out by a process dependent on mutation and natural selection. Expositions of this thesis, however, tend to overlook the fact that mutation occurs in the genotype, whereas natural selection acts only on the phenotype, the organism produced. It follows from this that the theory of evolution requires as one of its essential parts a consideration of the developmental or epigenetic processes by which the genotype becomes translated into the phenotype. The consequences of such considerations are discussed in the following sections.

Effect on life histories. *Length and timing of the reproductive phase.* Natural selection results in the production by one generation of offspring that are able to survive and reproduce themselves to form a further generation. The time unit appropriate to natural selection is therefore the generation interval. There will always be some natural selective pressure for the shortening of the generation interval, simply out of a natural economy, and for an increase of the number of offspring produced by any reproducing individual. One of the ways in which such an increase could be assured would be the lengthening of the reproductive phase in the life history; another would be an increase in the number of offspring produced.

These are, of course, not the only natural selective pressures that operate. It is clear enough that, in evolution, they have often been overcome by other pressures. There is another natural selective pressure of more general importance. This is the pressure to restrict the length of the reproductive period, and indeed to remove reproductive individuals, in order to make room for the maturation of a new generation in which new genetic combinations can be tried out for their fitness. A species whose individuals were immortal would exhaust its possibilities for future evolution as soon as its numbers saturated all the ecological niches suitable for its way of life. Death is a necessary condition for the trying out of new genetic combinations in later generations. It is usually brought about, in great part at least, by combinations of two processes: restriction of the period of effective reproduction to a certain portion of a life history, and as a necessary consequence of this, the absence of natural selection for genetic mutations that would be effective in preserving life after reproduction has ceased. In some organisms—for instance, long-lived trees—there may be no restriction of reproduction to a particular period in the life history, but their development involves the gradual accumulation of larger and larger quantities of nonliving materials, such as dead wood, which presents a growing handicap, in the face of which the organism cannot indefinitely maintain itself against the inevitable hazards of existence. It is still something of a question whether these natural selective forces are sufficient in themselves to account for the phenomena of senescence, aging, and eventual death, which are found in various forms throughout nearly the whole biological kingdom (see AGING).

The evolutionary significance of death

As was mentioned above, evolution has produced a number of the types of multiphasic development, in which the life history involves a succession of larval stages. Such types of development offer the possibility of changing the relative importance of the various stages in relation to the exploitation of resources and reproduction by the species. There are, for instance, many types of animals (particularly insects) in which nearly the whole life history is passed in a larval stage in which most of the feeding and growth of the organism is carried out, the final adult stage being short and used almost entirely for reproduction. Another evolutionary strategy has been to transfer the reproductive phase from the final stage of the life history to some earlier larval stage. This again has occurred in certain insects. If such a process is carried to its logical evolutionary conclusion, the final previously adult stage of the life history may totally disappear, the larval stage of the earlier evolutionary form becoming the adult stage of the later derivative of it. An example in which this process is at least partially accomplished is in the axolotl, a salamander that reproduces in a larval stage and in nature rarely if ever metamorphoses into the adult, but can be persuaded to do so if injected with extra supplies of the hormone thyroxin. It has been suggested that such processes of neoteny (the retention of some juvenile characteristics in adulthood) have played a decisive role in certain earlier phases of evolution, evidence of which is now lost. It has been argued that the whole vertebrate phylum may have originated from modifications of one of the larval stages of an invertebrate group.

The role of neoteny in evolution

Recapitulation of ancestral stages. The modifications of life histories just mentioned are aspects of a more general situation; namely, that the only variations that can become available for natural selection to operate on are those that can be produced by alterations of the developmental or epigenetic system of an existing organism. Any new mutant gene can cause a change only in a pre-existing set of developmental interactions; the phenotypes to which it can give rise are limited by the nature of the system that it will modify. One immediate result of this situation is that the development of a later evolved form will retain many features from the development of its ancestors: most evolutionary developments are likely to be additions to the previous organization. Since there is evolutionary pressure to reduce the length of time between generations, the addition of a new feature to development is likely to be accompanied by a speeding up of the older stages, and probably omission of certain of them.

To repeat, the development of a late-evolved form retains those aspects of earlier life histories that are essential for the building up of later developmental stages that

may be important for natural selection. In the vertebrates, for instance, highly evolved types such as mammals and birds produce during their early development remnants of the primitive kidneys (pronephros and mesonephros) that functioned as excretory organs in their evolutionary ancestors. Although these organs no longer perform their physiological functions in later organisms, they play an essential role during the formative processes of embryonic development. Some structures characteristic of evolutionary ancestors may be retained for relatively short evolutionary periods after they have lost their original function simply because there is not sufficient natural selective pressure to bring about their elimination when they no longer have any obvious function, either physiologically or epigenetically; the human appendix is an example.

Adaptability and the canalization of development. A developing organism is subjected to natural selection by its particular environment. The environment is not the same for all individuals of a population, nor does it necessarily remain the same throughout evolutionary periods of time. An organism can be regarded as having to meet environmental changes that are unpredictable. There are basically two different types of strategy employed, in various proportions in different organisms, to meet this situation. One, perhaps the more obvious, is to evolve a high capacity for modification by environmental circumstances in ways that increase fitness in the environment in question; this is the strategy of increasing adaptability. It is probably true to say that all organisms show some capacity for adaptation, either short-term (physiological) or longer term (developmental), to their environments. In most organisms, however, particularly in most higher organisms, there is considerable development of the alternative strategy, which is to build up well-buffered or channelled developmental chreods, which lead to the production of a relatively predictable invarient end result in the face of very diverse environments. The second strategy is likely to be followed in situations in which the environment is likely to change markedly during the course of the organism's life.

The tendency to channel development

Whether or not this is the main reason for the evolution of channelled, or canalized, developmental systems, a considerable degree of canalization is very common. It is relatively rare to find instances in which the form of an animal is highly dependent on the early environment, although such dependence is common enough among plants. Much more frequently, situations such as that typified by the house mouse are encountered: the mouse develops into an almost identical form whether it lives in the tropics or in a cold-storage depot.

This canalization of development severely restricts the phenotypic effects that can be produced by mutations. In particular, many new mutations occurring in a single dose in a diploid organism are found to be recessive, or ineffective in causing any alteration in the phenotype. As this discussion makes clear, canalization should not be considered as a relation involving only the normal and mutated forms of a particular gene, but rather the result of the interaction of many genes.

Genetic assimilation. A long-standing controversy in biology has been concerned with whether phenotypic modifications produced by abnormal environments are heritable in the sense that they can be produced by later generations in the absence of the original environmental stress. The hypothesis that they are heritable was advanced by the French evolutionist Lamarck in the 18th century and is generally known as the "inheritance of acquired characters." It found some supporters among biologists, some of whom used it as an argument against the Darwinian theory of evolution. In a broad sense, all characters are to some extent inherited, in that they depend on the genotype of the organism, and to some extent acquired, since development is also affected by the environment. In a stricter sense, however, Lamarck's hypothesis suggests that there is some inherent biological property that enables organisms to pass on physical modifications to their descendants, independently of a Darwinian mechanism of selection.

The combination of adaptability and canalization in development can explain such phenomena in strictly Darwinian rather than Lamarckian terms. The abnormal environment acting during development may succeed in modifying even a well-canalized development system. If the modification is of an adaptive kind and increases the fitness of the individuals in the unusual environment, it will be favoured by natural selection. The development of the selected individuals will, however, also show some properties of canalization, that is to say, resistance to further environmental changes. This invariance may be sufficient to prevent offspring of the selected individuals from reverting completely to the original phenotype even if they are removed from the abnormal environment. After selection for an adaptive modification in an abnormal environment has proceeded for many generations, a form may be produced whose canalization is strong enough to maintain the new phenotype almost unaltered when the environment reverts to what it was before the abnormality occurred. This process, which has been demonstrated in a number of laboratory experiments, is known as genetic assimilation. It produces exactly the same results as those emphasized by advocates of the Lamarckian inheritance of acquired characters, but it produces them by an orthodox Darwinian mechanism operating on developmental systems that have the common properties of canalization and adaptability. It provides the most convincing explanation for the evolution of organisms that are physiologically or functionally adapted to the demands their way of life will make.

The Darwinian interpretation of apparent Lamarckism

BIBLIOGRAPHY. Classical works that laid the foundation for the modern interpretation of development in terms of gene activities include: T.H. MORGAN, *Embryology and Genetics* (1934); R. GOLDSCHMIDT, *Physiologische Theorie der Vererbung* (1927); C.H. WADDINGTON, *Organisers and Gene* (1940); D.W. THOMPSON, *On Growth and Form*, rev. ed. (1942). A few more recent books, which include the theoretical as well as empirical aspects of development are: C.W. WARDLAW, *Organization and Evolution in Plants* (1965); J.T. BONNER, *Morphogenesis* (1952); E.W. SINNOTT, *The Problem of Organic Form* (1963). Probably the best general account of modern experimental work is A. KUHN, *Entwicklungsphysiologie*, 2nd ed. (1965). Less up to date is C.H. WADDINGTON, *Principles of Embryology* (1956). Shorter and less comprehensive works are: J.D. EBERT, *Interacting Systems in Development* (1965); and C.H. WADDINGTON, *New Patterns in Genetics and Development* (1962). A number of essays on the theory of development may be found in *Towards a Theoretical Biology*, 3 vol., ed. by C.H. WADDINGTON (1968–70), see particularly the essays by Goodwin, Thom, Wolpert, and Waddington.

(C.H.W.)

Development, Human

This article is concerned with the physical growth of a child from birth into maturity. Development before birth receives its principal treatment in the article EMBRYOLOGY, HUMAN.

Growth is far from being a simple and uniform process of becoming taller or larger. As a child gets bigger, there are changes in shape and in tissue composition and distribution. In the newborn infant the head represents about a quarter of the total length; in the adult it represents about one-seventh. In the newborn infant the muscles constitute a much smaller percentage of the total body mass than in the young adult. In most tissues, growth consists both of the formation of new cells and the packing in of more protein or other material into cells already present; early in development cell division predominates and later cell filling.

TYPES AND RATES OF HUMAN GROWTH

Different tissues and different regions of the body mature at different rates, and the growth and development of a child consists of a highly complex series of changes. It is like the weaving of a cloth whose pattern never repeats itself. The underlying threads, each coming off its reel at its own rhythm, interact with one another continuously, in a manner always highly regulated and controlled. The fundamental questions of growth relate to these processes of regulation, to the program that controls the loom, a

The complex patterns of growth

subject as yet little understood. Meanwhile, height is in most circumstances the best single index of growth, being a measure of a single tissue (that of the skeleton; weight is a mixture of all tissues, and this makes it a less useful parameter in a long-term following of a child's growth). In this article, the height curves of girls and boys are considered in the three chief phases of growth; that is (briefly) from conception to birth, from birth until puberty, and during puberty. Also described are the ways in which other organs and tissues, such as fat, lymphoid tissue, and the brain, differ from height in their growth curves. There is a brief discussion of some of the problems that beset the investigator in gathering and analyzing data about growth of children, of the genetic and environmental factors that affect rate of growth and final size, and of the way hormones act at the various phases of the growth process. Lastly, there is a brief look at disorders of growth. Throughout, the emphasis is on ways in which individuals differ in their rates of growth and development.

Growth as distance and velocity

The changes in height of the developing child can be thought of in two different ways: the height attained at successive ages and the increments in height from one age to the next, expressed as rate of growth per year. If growth is thought of as a form of motion, the height attained at successive ages can be considered the distance travelled, and the rate of growth, the velocity. The velocity or rate of growth reflects the child's state at any particular time better than does the height attained, which depends largely on how much the child has grown in all preceding years. The blood and tissue concentrations of those substances whose amounts change with age are thus more likely to run parallel to the velocity rather than to the distance curve. In some circumstances, indeed, it is the acceleration rather than the velocity curve that best reflects physiological events.

In general, the velocity of growth decreases from birth onward (and actually from as early as the fourth month of fetal life; see below), but this decrease is interrupted shortly before the end of the growth period. At this time, in boys from about 13 to 15 years (Figure 3), there is marked acceleration of growth, called the adolescent growth spurt. From birth until age four or five, the rate of growth in height declines rapidly, and then the decline, or deceleration, gets gradually less, so that in some children the velocity is practically constant from five or six up to the beginning of the adolescent spurt. A slight increase in velocity is sometimes said to occur between about six and eight years.

Prenatal growth. This general velocity curve of growth in height begins a considerable time before birth. Figure 1 shows the distance and velocity curves for body length in the prenatal period and first two postnatal years. The peak velocity of length is reached at about four months after the mother's last menstruation. (Age in the fetal period is usually reckoned from the first day of the last menstrual period, an average of two weeks before actual fertilization, but, as a rule, the only locatable landmark.)

Growth in weight of the fetus follows the same general pattern as growth in length, except that the peak velocity is reached much later, at approximately 34 weeks after the mother's last menstrual period.

There is considerable evidence that from about 34 to 36 weeks onward the rate of growth of the fetus slows down because of the influence of the maternal uterus, whose available space is by then becoming fully occupied. Twins slow down earlier, when their combined weight is approximately the 36-week weight of the single fetus. Babies who are held back in this way grow rapidly as soon as they have emerged from the uterus. Thus there is a significant negative association between weight of a baby at birth and weight increment during the first year; in general, larger babies grow less, the smaller more. For the same reason there is practically no relation between adult size and the size of that person at birth, but a considerable relation has developed by the time the person is two years old. This slowing-down mechanism enables a genetically large child developing in the uterus of a small mother to be delivered successfully. It operates in many species of animal; the most dramatic demonstration was by crossing reciprocally a large Shire horse and a small Shetland pony. The pair in which the mother was a Shire had a large newborn foal, and the pair in which the mother was Shetland had a small foal. But both foals were the same size after a few months, and when fully grown both were about halfway between their parents. The same has been shown in cattle crosses.

Rapid growth following small size at birth

Poor environmental circumstances, especially of nutrition, result in lowered birth weight in the human being. This seems chiefly to be caused by a reduced rate of growth in the last two to four weeks of fetal life, for weights of babies born in 36 or 38 weeks in various parts of the world in various circumstances are said to be similar. Mothers who, because of adverse circumstances in their own childhood, have not achieved their full growth potential may produce smaller fetuses than they would have, had they grown up in better circumstances. Thus two generations or even more may be needed to undo the effect of poor environmental circumstances on birth weight.

The great rate of growth of the fetus compared with that of the child is largely due to the fact that cells are still multiplying. The proportion of cells undergoing mitosis (the ordinary process of cell multiplication by splitting) in any tissue becomes progressively less as the fetus gets older, and it is generally thought that few if any new nerve cells (apart from the cells in the supporting tissue, or neuroglia) and only a limited proportion of new muscle cells appear after six postmenstrual months, the time when the velocity in linear dimensions is dropping sharply.

Prenatal and postnatal growth contrasted

The muscle and nerve cells of the fetus are considerably different in appearance from those of the child or adult. Both have little cytoplasm (cell substance) around the nucleus. In the muscle there is a great amount of intercellular substance and a much higher proportion of water than in mature muscle. The later fetal and the postnatal growth of the muscle consists chiefly of building up the cytoplasm of the muscle cells; salts are incorporated and the contractile proteins formed. The cells become bigger, the intercellular substance largely disappears, and the concentration of water decreases. This process continues quite actively up to about three years of age and slowly

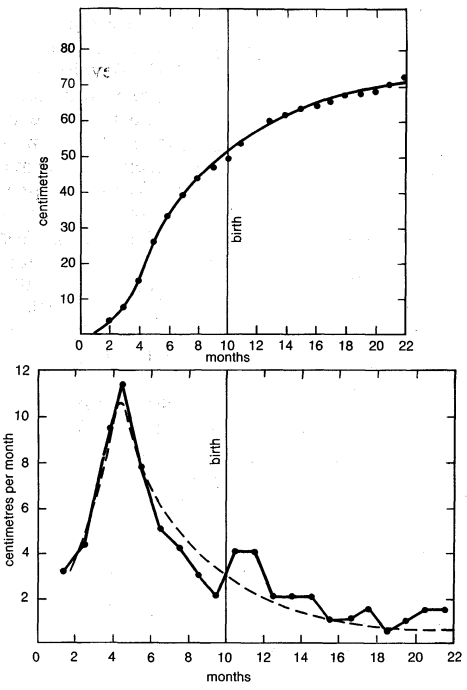


Figure 1: Curves indicating (above) growth attained and (below) velocity of growth during prenatal and early postnatal periods. The dashed line in the lower chart represents a smoothing out of the curve between points.

From D. Thompson, *Growth and Form*, 2nd ed. (1963); Cambridge University Press

thereafter; at adolescence it briefly speeds up again, particularly in boys, under the influence of androgenic (male sex) hormones. In the nerve cells cytoplasm is added and elaborated, and extensions grow that carry impulses from and to the cells—the axons and dendrites, respectively. Thus postnatal growth, for at least some tissues, is chiefly a period of development and enlargement of existing cells, while early fetal life is a period of division and addition of new cells.

Types of growth data. Growth is in general a regular process. Contrary to what is said in some of the older textbooks, growth in height does not proceed by fits and starts, nor does growth in upward dimensions alternate with growth in transverse ones. The more carefully measurements are taken, with precautions, for example, to minimize the decrease in height that occurs during the day for postural reasons, the more regular does the succession of points in a graph of growth become. Many attempts have been made at finding mathematical curves that fit, and thus summarize, human growth data. What is needed is a curve or curves with relatively few constants, each capable of being interpreted in a biologically meaningful way. Yet the fit to empirical data must be adequate within the limits of measuring error. The problem is difficult, partly because the measurements usually taken are themselves biologically complex. Stature, for example, consists of leg length and trunk length and head height, all of which have rather different growth curves. Even with relatively homogeneous dimensions such as the length of the radius bone in the forearm, or width of an arm muscle, it is not clear what purely biological assumptions should be made as the basis for the form of the curve. The assumption that cells are continuously dividing leads to a different formulation from the assumption that cells are adding constant amounts of nondividing material or amounts of nondividing material at rates varying from one age period to another.

Fitting a curve to the individual values, however, is the only way of extracting the maximum information from an individual's measurement data. More than one curve is needed to fit the postnatal age range. It seems that two curves may suffice, at least for many measurements such as height and weight—one curve for the period from a few months after birth to the beginning of adolescence and a different type of curve for the adolescent spurt.

Such curves have to be fitted to data on single individuals. Yearly averages derived from different children each measured only once do not, in general, give the same curve. Thus the distinction between the two sorts of investigation is important. When the same child at each age is used, the study is called longitudinal; when different children at each age are used, it is called cross-sectional. In a cross-sectional study all of the children at age eight, for example, are different from those at age seven. A study may be longitudinal over any number of years; there are short-term longitudinal studies extending from age four to six, for instance, and full birth-to-maturity longitudinal studies in which the children may be examined once, twice, or more times every year from birth until 20 or over. Mixed longitudinal studies are those in which children join and leave the group studied at varying intervals. Both cross-sectional and longitudinal studies have their uses, but they do not give the same information, and the same statistical methods cannot be used for the two types of study. Cross-sectional surveys are obviously cheaper and more quickly done and can include much larger numbers of children. Periodic cross-sectional surveys are valuable in assessing the nutritional progress of a country or a socio-economic group and the health of the child population as a whole. But they never reveal individual differences in rate of growth or in the timing of particular phases such as the adolescent growth spurt. It is these individual rate differences that throw light on the genetic control of growth and on the correlation of growth with psychological development, educational achievement, and social behaviour.

Advantages in longitudinal and cross-sectional studies

Longitudinal studies are laborious and time-consuming; they demand great perseverance on the part of those who make them and those who take part in them; and they demand high technical standards, since in the calculation of a growth increment from one occasion to the next opportunities for two errors of measurement occur. In spite of these problems, longitudinal studies are the indispensable base on which the diagnosis and treatment of disorders of growth rest, for the clinical approach is a longitudinal one; and each child treated with human growth hormone, or with other hormones that affect growth, represents an attempt to alter an individual pattern of growth velocity.

Averages simply computed from cross-sectional data inevitably produce velocity curves that are flatter and broader than the curve for an individual and hence not a proper basis for clinical standards. It is possible to construct curves, however, whose 50th percentile (or average) represents the actual growth of a typical individual, by taking the shape of the curve from individual longitudinal data and the absolute values for the beginning and end from large cross-sectional surveys. Figures 2 and 3 show height-attained and height-velocity curves for the "typical" boy and girl in Britain in 1965, determined in this way. By "typical" is meant that boy or girl who has the mean (average) birth length, grows always at the mean velocity, has the peak of the adolescent growth spurt at the mean age, and finally reaches the mean adult height at the mean age of cessation of growth. Practically no individual follows the 50th percentile curve of Figure 2, but most have curves of the same shape. Standards for height for clinical use are constructed around these curves.

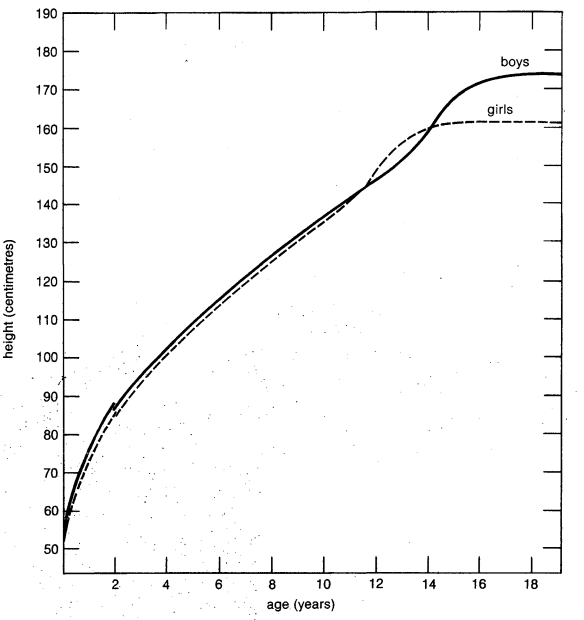


Figure 2: Individual height curves for British boys and girls in 1965 (see text).

Boys' and girls' height curves. Figures 2 and 3 show the height curves from birth to maturity. Up to age two, the child is measured lying on his back. One examiner holds his head in contact with a fixed board, and a second person stretches him out to his maximum length and then brings a moving board into contact with his heels. This measurement, called supine length, averages about one centimetre more than the measurement of standing height taken on the same child, hence the break in the line in Figure 2 at age two. This occurs even when, as in the best techniques, the child is urged to stretch upwards to the full and is aided in doing so by a measurer's applying gentle upward pressure to his mastoid processes.

Measuring the child's height

Figure 2 shows the typical girl as slightly shorter than the typical boy at all ages until adolescence. She becomes taller shortly after age 11 because her adolescent spurt takes place two years earlier than the boy's. At age 14 she is surpassed again in height by the typical boy, whose

adolescent spurt has now started, while hers is nearly finished. In the same way, the typical girl weighs a little less than the boy at birth, equals him at age 8, becomes heavier at age 9 or 10, and remains so until about age 14½.

The velocity curves given in Figure 3 show these processes more clearly. At birth the typical boy is growing slightly faster than the typical girl, but the velocities become equal at about seven months, and then the girl

From *Archives of Disease in Childhood*, vol. 41, pp. 454, 613 (1966); British Medical Association

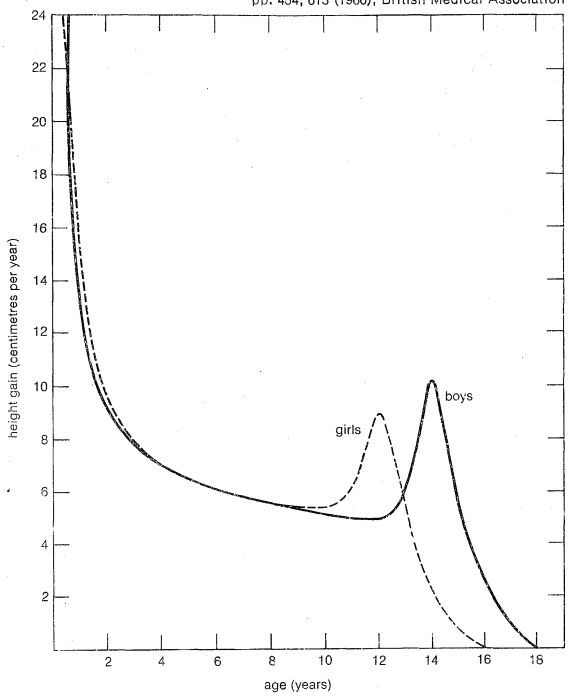


Figure 3: Typical velocity curves for supine length or height in British boys and girls in 1965 (see text).

grows faster until four years. From then until adolescence no differences in velocity can be detected. The sex difference is best thought of, perhaps, in terms of acceleration, the boy decelerating harder than the girl over the first four years.

Different tissues and parts of the body. The majority of skeletal and muscular dimensions follow approximately the growth curve described for height, and so also do the dimensions of the internal organs such as the liver, the spleen, and the kidneys. But some exceptions exist, most notably the brain and skull, the reproductive organs, the lymphoid tissue of the tonsils, adenoids, and intestines, and the subcutaneous fat.

In Figure 4 these differences are shown; the size attained by various tissues is given as a percentage of the birth-to-maturity increment. Height follows the "general" curve. The reproductive organs, internal and external, have a slow prepubescent growth, followed by a large adolescent spurt; they are less sensitive than the skeleton to one set of hormones and more sensitive to another.

Growth of the brain, the skull, and the face

The brain, together with the skull covering it and the eyes and ears, develops earlier than any other part of the body and thus has a characteristic postnatal curve. At birth it is already 25 percent of its adult weight, at age five about 90 percent, and at age 10 about 95 percent. Thus if the brain has any adolescent spurt at all, it is a small one. A small but definite spurt occurs in head length and breadth, but all or most of this is due to thickening of the skull bones and the scalp, together with development of the air sinuses.

The dimensions of the face follow a path somewhat closer to the general curve. There is a considerable adolescent spurt, especially in the lower jaw, or mandible, resulting in the jaw's becoming longer and more projecting, the profile straighter, and the chin more pointed. As always in growth, there are considerable individual differ-

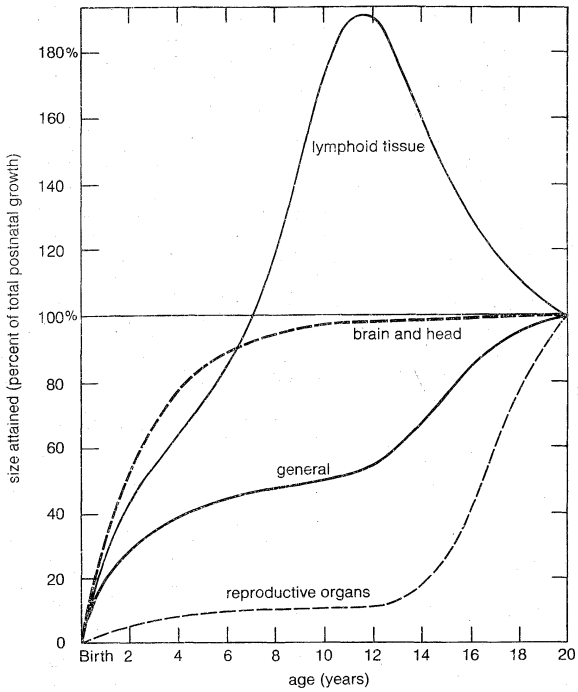


Figure 4: Major types of postnatal growth of parts and organs of the body (see text).
From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications, after Scammon

ences, to the point that a few children have no detectable spurt at all in some face measurements.

The eye probably has a slight adolescent spurt, which is probably responsible for the increase in frequency of short-sightedness in children that occurs at the time of puberty. Though the degree of myopia increases continuously from at least age six to maturity, a particularly rapid rate of change occurs at about 11 to 12 in girls and 13 to 14 in boys, and this would be expected if there was a rather greater spurt in the axial dimension (the dimension from front to back) of the eye than in its vertical dimension.

The lymphoid tissue has quite a different growth curve from the rest. It reaches its maximum amount before adolescence and then, probably under the direct influence of sex hormones, declines to its adult value.

Development of fat

The subcutaneous fat layer also has a curve of its own, of a slightly complicated sort. Its thickness can be measured either by X-rays or, more simply, at certain sites in the body, by picking up a fold of skin and fat between the thumb and forefinger and measuring its thickness with a special, constant-pressure caliper. Figure 5 shows the distance curves of skin folds taken halfway down the back

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

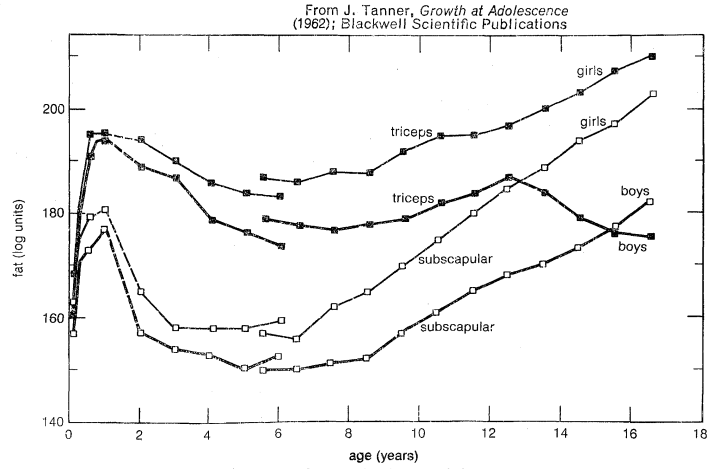


Figure 5: Amount of subcutaneous fat on the back of the arm (triceps) and the back of the body (subscapular) from birth to 16 years (see text).

of the upper arm (the triceps) and on the body just below the angle of the shoulder blade, or scapula. Subcutaneous fat begins to be laid down in the fetus at about 34 weeks postmenstrual age, increases from then until birth and from birth onward until about nine months. (This is in the average child; the peak may be reached as early as six months or as late as 12 or 15.) After nine months, when the velocity of fat gain is zero, the fat usually decreases (that is, it has a negative velocity) until age six to eight, when it begins to increase once more. Girls have a little more fat than boys at birth, and the difference becomes more marked during the period of loss, since girls lose less than boys. From eight years on, the curves for girls and boys diverge more radically, as do the curves for limb and body fat. At adolescence the limb fat in boys decreases, while the body fat shows a temporary slowing down of gain but no actual loss. In girls there is a slight halting of the limb-fat gain at adolescence, but no loss; the trunk fat shows only a steady rise until adolescence.

Development at puberty. *Growth.* At puberty, a considerable alteration in growth rate occurs. There is a swift increase in body size, a change in shape and composition of the body, and a rapid development of the gonads, or sex glands—the reproductive organs and the characters signalling sexual maturity. Some of these changes are common to both sexes, but most are sex-specific. Boys have a great increase in muscle size and strength, together with a series of physiological changes making them capable of doing heavier physical work than girls and of running faster and longer. These changes all specifically adapt the male to his primitive primate role of dominating, fighting, and foraging. Such adolescent changes occur generally in primates (that is, men, apes, and monkeys) but are more marked in some species than in others. Man lies at about the middle of the primate range, as regards both adolescent size increase and degree of sexual differentiation.

The adolescent spurt in height

During the adolescent spurt in height, for a year or more, the velocity of growth approximately doubles; a boy is likely to be growing again at the rate he last experienced about age two. The peak velocity of height (P.H.V., a point much used in growth studies) averages about 10.5 centimetres per year in boys and 9.0 centimetres in girls (about 4 and 3.4 inches, respectively), but this is the "instantaneous" peak given by a smooth curve drawn through the observations. The velocity over the whole year encompassing the six months before and after the peak is naturally somewhat less. During this year a boy usually grows between 7 and 12 centimetres (2.75 and 4.75 inches) and a girl between 6 and 11 centimetres (2.35 and 4.35 inches). Children who have their peak early reach a somewhat higher peak than those who have it late.

The average age at which the peak is reached depends on the nature and circumstances of the group studied more, probably, than does the height of the peak. In moderately well-off British or North American children at present the peak occurs on average at about 14.0 years in boys and 12.0 years in girls. Though the absolute average ages differ from population to population, the two-year sex difference always persists.

Practically all skeletal and muscular dimensions take part in the spurt, though not to an equal degree. Most of the spurt in height is due to acceleration of trunk length rather than of length of legs. There is a fairly regular order in which the dimensions accelerate; leg length as a rule reaches its peak first, followed by the body breadths, with shoulder width last. The earliest structures to reach their adult status are the head, hands, and feet.

The spurt in muscle, of both limbs and heart, coincides with the spurt in skeletal growth, for both are caused by the same hormones. Boys' muscle widths reach a peak velocity of growth that is greater than that reached by girls. But, since girls have their spurt earlier, there is actually a period, from about 12½ to 13½, when girls on average have larger muscles than boys of the same age, as well as being taller (Figure 2). Simultaneously with the spurt there is a loss of fat, as described above.

The marked increase in muscle size in boys at adolescence leads to an increase in strength, illustrated in Figure 6. Before adolescence, boys and girls are similar in

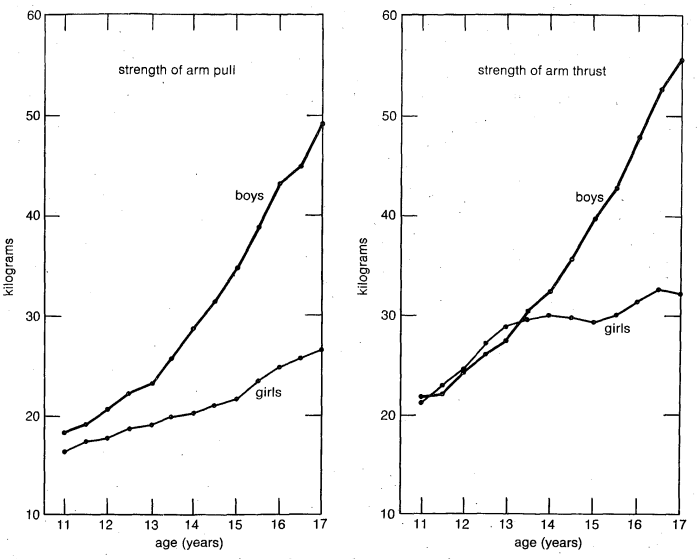


Figure 6: Strength of arm pull and thrust for boys and girls.

strength for a given body size and shape; after, boys have much greater strength, probably due to development of more force per gram of muscle as well as to absolutely larger muscles. They also develop larger hearts and lungs relative to their size, a higher systolic blood pressure (the pressure resulting from a heart contraction), a lower resting heart rate, a greater capacity for carrying oxygen in the blood with more hemoglobin, and a greater power for neutralizing the chemical products of muscular exercise such as lactic acid. In short, the male becomes at adolescence more adapted for the tasks of hunting, fighting, and manipulating all sorts of heavy objects, as is necessary in some forms of food-gathering.

Adolescent increase in strength

It is as a direct result of these anatomical and physiological changes that athletic ability increases so much in boys at adolescence. The popular notion of a boy's "outgrowing his strength" at this time has little scientific support. It is true that the peak velocity of strength is reached a year or so later than that of height, so that a short period may exist when the adolescent, having completed his skeletal and probably also his muscular growth, still does not have the strength of a young adult of the same body size and shape. But this is a temporary phase; considered absolutely, power, athletic skill, and physical endurance all increase progressively and rapidly throughout adolescence.

Reproductive system. The adolescent spurt in skeletal and muscular dimensions is closely related to the rapid development of the reproductive system that takes place at this time. The course of this development is outlined diagrammatically in Figure 7. The bar marked "breast" in the chart for the girls and the bars marked "penis" and "testis" in the chart for the boys represent the periods of accelerated growth of these organs. Other bars indicate the genitalia rating and the advent and development of the pubic hair. The sequence and timings that are given represent in each case average values for British boys and girls; the North American average is within two or three months of this. To give an idea of the individual departures from the average, figures for the range of age at which the various events begin and end are inserted under the first and last point of the bars. The acceleration of penis growth, for example, begins on average at about age 12½ years, but sometimes as early as 10½ and sometimes as late as 14½. The completion of penis development usually occurs at about age 14½, but in some boys is at 12½ and in others at 16½. There are a few boys, it will be noticed, who do not begin their spurts in

Development of reproductive organs

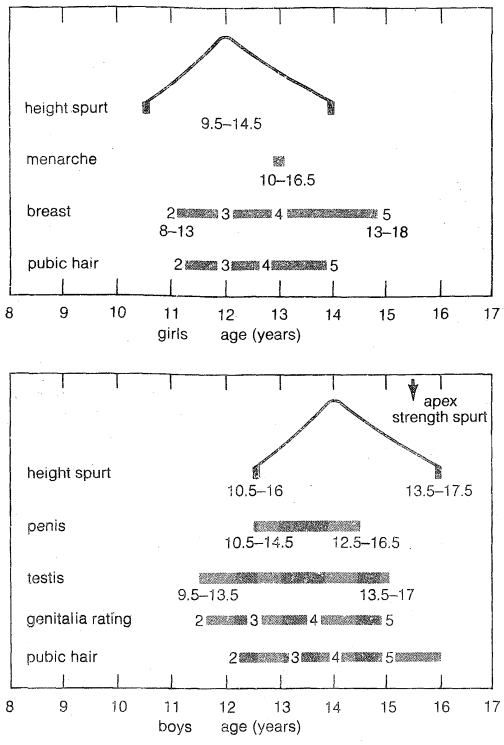


Figure 7: Sequence of events at adolescence for average British boys and girls. Range of ages within which each event may begin and end is given by the figures placed directly below its start and finish. The numbers 2 to 5 within the bars indicate stage of maturity, 5 being full maturity and 2 the beginning of adolescence (see text)
W. Marshall and J. Tanner, *Archives of Disease in Childhood* (February 1970); British Medical Association

height or penis development until the earliest maturers have entirely completed theirs. At ages 13, 14, and 15 there is an enormous variability among any group of boys, who range all the way from practically complete maturity to absolute preadolescence. The same is true of girls aged 11, 12, and 13.

The psychological and social importance of this difference in the tempo of development, as it has been called, is great, particularly in boys. Boys who are advanced in development are likely to dominate their contemporaries in athletic achievement and sexual interest alike. Conversely the late developer is the one who all too often loses out in the rough and tumble of the adolescent world, and he may begin to wonder whether he will ever develop his body properly or be as well endowed sexually as those others whom he has seen developing around him. An important part of the educationist's and the doctor's task at this time is to provide information about growth and its variability to preadolescents and adolescents and to give sympathetic support and reassurance to those who need it.

The first sign of puberty in the boy

The sequence of events, though not exactly the same for each boy or girl, is much less variable than the age at which the events occur. The first sign of puberty in the boy is usually an acceleration of the growth of the testes and scrotum with reddening and wrinkling of the scrotal skin. Slight growth of pubic hair may begin about the same time but is usually a trifle later. The spurts in height and penis growth begin on average about a year after the first testicular acceleration. Concomitantly with the growth of the penis, and under the same stimulus, the seminal vesicles, the prostate, and the bulbo-urethral glands, all of which contribute their secretions to the seminal fluid, enlarge and develop. The time of the first ejaculation of seminal fluid is to some extent culturally as well as biologically determined but as a rule is during adolescence and about a year after the beginning of accelerated penis growth.

Axillary (armpit) hair appears on average some two years after the beginning of pubic hair growth; that is, when pubic hair is reaching stage 4 (Figure 7). There is enough variability and dissociation in these events, so that a very few children's axillary hair actually appears first. In boys, facial hair begins to grow at about the time that the axillary hair appears. There is a definite order in which the hairs of moustache and beard appear: first at the corners of the upper lip, then over all the upper lip, then at the upper part of the cheeks, in the midline below the lower lip, and, finally, along the sides and lower borders of the chin. The remainder of the body hair appears from about the time of first axillary hair development until a considerable time after puberty. The ultimate amount of body hair that an individual develops seems to depend largely on heredity, though whether because of the kinds and amounts of hormones secreted or because of variations in the reactivity of the end organs is not known.

Breaking of the voice occurs relatively late in adolescence. The change in pitch accompanies enlargement of the larynx and lengthening of the vocal cords, caused by the action of the male hormone testosterone on the laryngeal cartilages. There is also a change in quality that distinguishes the voice (more particularly the vowel sounds) of both male and female adults from that of children. This is caused by the enlargement of the resonating spaces above the larynx, as a result of the rapid growth of the mouth, nose, and maxilla (upper jaw).

In the skin, particularly of the armpits and the genital and anal regions, the sebaceous and apocrine sweat glands develop rapidly during puberty and give rise to a characteristic odour; the changes occur in both sexes but are more marked in the male. Enlargement of the pores at the root of the nose and the appearance of comedones (blackheads) and acne, while likely to occur in either sex, are considerably more common in adolescent boys than girls, since the underlying skin changes are the result of androgenic (male sex hormone) activity.

During adolescence the male breast undergoes changes, some temporary and some permanent. The diameter of the areola, which is equal in both sexes before puberty, increases considerably, though less than it does in girls. In some boys (between a fifth and a third of most groups studied) there is a distinct enlargement of the breast (sometimes unilaterally) about midway through adolescence. This usually regresses again after about one year.

First signs of puberty in the girl

In girls the start of breast enlargement—the appearance of the "breast bud"—is as a rule the first sign of puberty, though the appearance of pubic hair precedes it in about one-third. The uterus and vagina develop simultaneously with the breast. The labia and clitoris also enlarge. Menarche, the first menstrual period, is a late event in the sequence. It occurs almost invariably after the peak of the height spurt has been passed. Though it marks a definitive and probably mature stage of uterine development, it does not usually signify the attainment of full reproductive function. The early cycles may be more irregular than later ones and in some girls, but by no means all, are accompanied by discomfort. They are often anovulatory; that is, without the shedding of an egg. Thus there is frequently a period of adolescent sterility lasting a year to 18 months after menarche, but it cannot be relied on in the individual case. Similar considerations may apply to the male, but there is no reliable information about this. On average, girls grow about six centimetres (about 2.4 inches) more after menarche, though gains of up to twice this amount may occur. The gain is practically independent of whether menarche occurs early or late.

Normal variations. Children vary a great deal both in the rapidity with which they pass through the various stages of puberty and in the closeness with which the various events are linked together. At one extreme one may find a perfectly healthy girl who has not yet menstruated though her breasts and pubic hair are characteristic of the adult and she is already two years past her peak height velocity; and at the other, a girl who has passed all the stages of puberty within the space of two years.

In girls the interval from the first sign of puberty to complete maturity varies from 18 months to six years. The period from the moment when the breast bud first

appears to menarche averages 2½ years but may be as little as six months or as much as 5½ years. The rapidity with which a child passes through puberty seems to be independent of whether puberty is occurring early or late. Menarche invariably occurs after peak height velocity is passed, so that the tall girl can be reassured about future growth if her menstrual periods have begun.

In boys a similar variability occurs. The genitalia may take between two and five years to attain full development, and some boys complete the whole process before others have moved from the first to the second stage.

The height spurt occurs relatively later in boys than in girls. Thus there is a difference between the average boy and girl of two years in age of peak height velocity but of only one year in the first appearance of pubic hair. Indeed, in some girls the acceleration in height is the first sign of puberty; this is never so in boys. A small boy whose genitalia are just beginning to develop can be unequivocally reassured that an acceleration in height is soon to take place, but a girl in the corresponding situation may already have had her height spurt.

Sex dimorphism. The differential effects on the growth of bone, muscle, and fat at puberty increase considerably the difference in body composition between the sexes. Boys have a greater increase not only in stature but especially in breadth of shoulders; girls have a greater relative increase in width of hips. These differences are produced chiefly by the changes of puberty, but other sex differentiations arise before that time. Some, like the external genital difference itself, develop during fetal life. Others develop continuously throughout the whole growth period by a sustained differential growth rate. An example of this is the greater relative length and breadth of the forearm in the male when compared with whole arm length or whole body length.

Part of the sex difference in pelvic shape antedates puberty. Girls at birth already have a wider pelvic outlet. Thus the adaptation for childbearing is present from an early age. The changes at puberty are concerned more with widening the pelvic inlet and broadening the much more noticeable hips. It seems likely that these changes are more important in attracting the males' attention than in dealing with its ultimate product.

Physical and behavioral interaction. Children vary greatly in their tempo of growth. The effects are most dramatically seen at adolescence, but they are present at all ages from birth and even before.

Developmental and chronological ages

The concept of developmental age, as opposed to chronological age, is an important one. To measure developmental age, there is need of some way of determining how far along his own path to maturity a given child has gone. Thus there is need of a measure in which everyone at maturity ends up the same (not different as in height). The usual measure used is skeletal maturity or bone age. This is measured by taking an X-ray of the hand and wrist (using the same radiation exposure that a child inevitably gets, and to more sensitive areas, by spending a week on holiday in the mountains). The appearances of the developing bones can be rated and formed into a scale; the scale is applicable to boys and girls of all genetic backgrounds, though girls on average reach any given score at a younger age than boys; and blacks on average, at least in the first few years after birth, reach a given score younger than do whites. Other areas of the body may be used if required. Skeletal maturity is closely related to the age at which adolescence occurs; that is, to maturity measured by some sex character developments. Thus the range of the chronological age within which menarche may normally fall is about 10 to 16½, but the corresponding range of bone age for menarche is only 12 to 14½. Evidently the physiological processes controlling progression of skeletal development are in most instances closely linked with those that initiate the events of adolescence. Furthermore, children tend to be consistently advanced or retarded during their whole growth period, at any rate after about age three.

Effects of late and early maturity

There is little doubt that being an early or a late maturer may have repercussions on behaviour and that in some children these repercussions may be considerable. There is little enough solid information on the relation between emotional and physiological development, but what there is supports the common-sense notion that emotional attitudes are clearly related to physiological events. Boys who are advanced in development, not only at puberty but before as well, are more likely than others to be leaders. Indeed, this is reinforced by the fact that muscular, powerful boys on average mature earlier than others and have an early adolescent growth spurt. The athletically built boy not only tends to dominate his fellows before puberty but, by getting an early start, is in a good position to continue that domination. The unathletic, linear boy, unable, perhaps, to hold his own in the preadolescent rough-and-tumble, gets still further pushed to the wall at adolescence, as he sees others shoot up while he remains nearly stationary in growth. Even boys several years younger now suddenly surpass him in size, athletic skill, and perhaps in social graces also.

Larger size and earlier maturation. The rate of maturing and the age of onset of puberty are dependent on a complex interaction of genetic and environmental factors. Where the environment is good, most of the variability in age at menarche in a population is due to genetical differences. In France in the 1950s, the mean difference for identical twins was two months, while that between nonidentical twin sisters was eight months. In many societies puberty occurs later in the poorly off, and, in most societies investigated, children with many siblings grow more slowly than children with few.

Modern tendency toward earlier maturity

During the last hundred years there has been a striking tendency for children to become progressively larger at all ages. This is known as the "secular trend." The magnitude of the trend in Europe and America is such that it dwarfs the differences between socio-economic classes.

The data from Europe and America agree well: from about 1900, or a little earlier, to the present, children in average economic circumstances have increased in height at age five to seven by about one to two centimetres (0.4 to 0.8 inch) per decade, and at 10 to 14 by two to three centimetres (0.8 to 1.2 inches) each decade. Preschool data show that the trend starts directly after birth and may, indeed, be relatively greater from age two to five than subsequently. The trend started, at least in Britain, as early as 1850.

Most of the trend toward greater size in children reflects a more rapid maturation; only a minor part reflects a greater ultimate size. The trend toward earlier maturing is best shown in the statistics on age at menarche. A selection of the best data is illustrated in Figure 8. The trend is between three and four months per decade since 1850 in average sections of western European populations. Well-off persons show a trend of about half of this

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

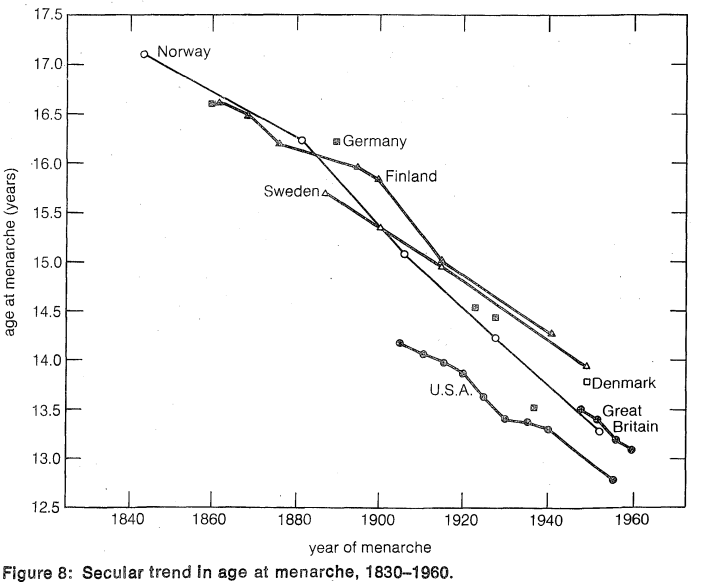


Figure 8: Secular trend in age at menarche, 1830–1960.

magnitude, having never been so retarded in menarche as the worse off. The causes of the secular trend are probably multiple. Certainly better nutrition is a major one and perhaps in particular more protein and calories in early infancy. A lessening of disease may also have contributed. Hot climates used to be cited as a potent cause of early menarche, but it seems that their effect, if any, is considerably less than that of nutrition. Some authors have supposed that the increased psychosexual stimulation consequent on modern urban living has contributed, but there is no positive evidence for this.

HORMONES AND GROWTH

The main hormones concerned with growth are pituitary growth hormone, thyroid hormone, the sex hormones testosterone and estrogen, and the pituitary gonadotropic (sex-gland-stimulating) hormones.

Pituitary growth hormone. Pituitary growth hormone, a protein with molecular weight of 21,600 and of known amino-acid composition, is secreted by the pituitary gland throughout life. Exactly what its function is in the adult is not clear, but in the child it is necessary for growth; without it dwarfism results. During fetal life it seems not to be necessary, though normally present. It is not secreted at a constant rate all day but in small bursts of activity. Secretion by the pituitary is controlled by a substance sent to it from an adjacent part of the brain. The normal stimulus for secretion is not certain, but a sharp and "unnatural" lowering of blood sugar will cause growth hormone to be secreted, and this is used as a test. The hormone decreases the amount of fat and causes protein to be laid down in muscles and viscera. Children who lack it are fat as well as small; when given it by injection, they lose fat and grow rapidly.

Action of growth hormone

The hormone is peculiar in being species-specific; that is, only growth hormone from human glands is active in man. Supplies of the hormone for treating children who need it are obtained at autopsy, and supply has been limited by this. The hormone has been made synthetically, but the processes for its manufacture have not yet been adopted for commercial production.

Thyroid hormone. Thyroid hormone from the thyroid gland in the neck is necessary for normal growth, though it does not itself stimulate growth, for example in the absence of pituitary growth hormone. Without thyroid hormone, however, cells do not develop and function properly, especially in the brain. Babies who lack thyroid hormone at birth are small and have insufficiently developed brains; they are known as cretins. Frequently, if the condition is diagnosed and they are treated with thyroid hormone at once, they recover completely; the longer they go without treatment, the more likely it is that the brain damage will be permanent.

Thyroid lack may also develop later in childhood, when it causes a slowing of growth rate; full catch-up follows prompt treatment.

Sex hormones. Testosterone, secreted by the interstitial cells of the testis, is important not only at puberty but before. Its secretion by the fetal testis cells is responsible for the development of certain parts of the male genital apparatus and for fixing the brain into the male pattern, which cannot later, it seems, be altered. If testosterone is not secreted at a particular and circumscribed time, the genitalia and the brain develop into the female form.

Effects of sex hormones on growth

Only small amounts of testosterone circulate between birth and puberty, but at puberty the interstitial cells develop greatly in response to pituitary luteinizing hormone (see below), and testosterone is secreted in large amounts, bringing about most of the changes of male puberty. It acts on a widespread series of receptors—for example, the cells of the penis, the muscles, the skin of the face, the cartilages of the shoulder, and certain parts of the brain. Most of the adolescent growth spurt is due to testosterone.

The female sex hormones, collectively called estrogens, are first secreted in quantity at puberty by cells in the ovary. They cause growth of the uterus, vagina, and breast; they act also on the bones of the hip, causing the specifically female widening. The adolescent growth spurt in the female is probably caused by testosterone-like substances (androgens) secreted by the adrenal gland in both male and female.

Pituitary gonadotropins and initiation of puberty. The pituitary secretes two other hormones concerned in development: one, follicle-stimulating hormone (FSH), causes growth of the main portions of the ovary in the female and the sperm-producing cells in the testis of the male; the other, luteinizing hormone (LH), causes growth and secretion of the testosterone-secreting cells of the male and has an action in controlling the menstrual cycle in the female. The pituitary is caused to secrete gonadotropins by substances called releasing factors that come to it from adjacent areas of the brain, where they are made. Certain children develop all the changes of puberty, up to and including sperm production or ovulation, at an early age, either as the result of a brain lesion or as an isolated developmental, sometimes genetic, defect. The youngest mother on record was such a child; she gave birth to a full-term healthy infant by cesarean section at the age of five years and eight months. The existence of precocious puberty and the results of accidental ingestion by small children of male or female sex hormones indicate that breasts, uterus, and penis will respond to hormonal stimulation long before puberty. Evidently an increased end-organ sensitivity plays at most a minor part in puberal events.

Youngest mother on record

The signal to start the sequence of events is given by the brain, not the pituitary. Just as the brain holds the information on sex, so it holds information on maturity. The pituitary gland of a newborn rat successfully grafted in place of an adult pituitary begins at once to function in an adult fashion and does not have to wait until its normal age of maturation has been reached. It is the hypothalamus in the brain, not the pituitary, that must mature before puberty begins. Small amounts of sex hormones circulate from the time of birth, and these appear to inhibit the prepuberal hypothalamus from producing gonadotropin releasers. At puberty the hypothalamic cells become less sensitive to sex hormones. The small amount of sex hormones circulating then fails to inhibit the hypothalamus; gonadotropins are released, and these stimulate the production of testosterone by the testis or estrogen by the ovary. The level of the sex hormone rises until the same feedback circuit is re-established but now at a higher level of gonadotropins and sex hormones. The sex hormones are now high enough to stimulate the growth of secondary sex characters and to support mating behaviour.

DISORDERS OF GROWTH

Short stature. There are a number of causes of short stature (a term much preferable to dwarfism). The most common of them is simple inheritance with or without a greater-than-usual delay in the tempo of growth. Given a reasonably good environment, a child's height is chiefly controlled by the genes that he inherited from his parents. The control depends on many genes, each of small effect, and, in the average case, depends equally on the father and the mother. Thus two small parents, unless they are small because stunted by chronic malnutrition in their childhood, can expect to have mostly small children. Charts are now available giving the normality of a child's height when its parents' heights are known.

Secondly, the child may be simply delayed in growth. This can be established by estimating his bone age or by seeing whether his height is within normal limits for the bone age rather than the chronological age. In boys with delayed growth, testosterone or one of its derivatives can be given if psychological problems become great at an age when the majority of boys are entering puberty; however, unless the bone age is extremely retarded, one runs the risk of diminishing the final adult height through maturing the skeleton too rapidly, so that the treatment should be given only by persons with considerable experience of it.

Neither hereditary small stature nor delayed maturation is a disorder, of course, but the extreme of normal variation.

Growth hormone deficiency. There may be a deficiency of growth hormone alone, with other pituitary functions normal, or there may also be deficiencies of the other hormones secreted by the pituitary, such as thyroid-stimulating hormone, adrenocorticotropic hormone (ACTH), and gonadotropins.

The isolated growth hormone deficiency is more common in boys than in girls and usually is manifested by a reduced growth rate from birth onward. The size at birth is typically close to normal. Usually the child is fat and has a delayed bone age; the diagnosis is clinched by an inability to secrete growth hormone on a stimulation test. Treatment with human growth hormone is usually successful, a rapid rate of growth being induced in the first year of treatment. Even if the same dose is maintained, the response becomes less as the child approaches normal size for his age. In rare cases antibodies to HGH develop, and treatment has to be abandoned.

Panhypopituitarism, deficiency of all the pituitary hormones, most commonly follows the removal of a craniopharyngioma, a benign tumour of a part of the brain close to the pituitary gland. For this condition other hormones, notably thyroid and cortisone, may have to be given as well as HGH. Results are usually good.

Hormonal causes of short stature

Hypothyroidism. Lack of a properly functioning thyroid gland will, among other things, cause a decreased rate of growth. Bone age is delayed more than in HGH deficiency. Treatment with thyroid hormone is usually effective.

Steroid (cortisone) excess. Hormones of the type of cortisone have a powerful growth-stopping effect, and short stature follows administration of cortisone and similar drugs, if these are given in more than extremely small doses. This may have to be done for really intractable asthma or rheumatoid arthritis in a child but should be avoided to the utmost. Cushing's disease is a state of excess secretion of cortisone by the adrenal gland; this causes short stature, delayed bone age, and fatness.

Gonadal dysgenesis (Turner's syndrome). The sex chromosomes in females are normally XX, in contrast to the male XY configuration. In Turner's syndrome, in girls, the configuration is XO (that is, there is only one sex chromosome, an X-chromosome), or XX/XO. Short stature is characteristic, for unknown reasons. The ovaries fail to develop, so that no enlargement of the breasts or uterus occurs at puberty and the affected persons are sterile. The diagnosis is made by examination of the chromosomes.

Low birth weight and short stature. A proportion of children born with abnormally low birth weight, usually accompanied by a characteristic facial appearance, lack of fat, and often asymmetry of the body, remain small throughout their lives. They have normal amounts of thyroid and growth hormone and undergo a normal puberty. Injections of growth hormone are ineffective. It seems likely that the basic defect is a lack of the normal complement of cells. The disorder is called Silver's syndrome or Russell's syndrome.

Other causes of short stature

Psychosocial short stature. Certain children who are subjected to psychological stress by an inimical home environment appear to react by not growing normally. This is not simple starvation or refusal to eat; on the contrary, the eating pattern often is bizarre and obsessive, food being eaten from time to time voraciously or taken from a dog's dish or a garbage can. These children tend to be fat, like those with growth hormone deficiency. Giving human growth hormone is ineffective, but removal of the children from their surroundings usually results in a catch-up in growth.

Malabsorption, starvation. Chronic starvation causes short stature, as does chronic malabsorption.

Bone disorders. The most common bone disorders causing short stature are achondroplasia and its near but distinct relative hypochondroplasia. In both there is shortening of the legs and arms relative to the trunk; in particular, the upper arm and the thigh are shortened. In achondroplasia there is also a characteristic facial appearance with prominent forehead and snub nose; in hypochondroplasia the face is normal. The disorder is one of bone, and no treatment is currently effective. The state is inherited, but most cases are due to fresh mutations and do not have a family history. There are numerous rarer bone diseases causing short stature.

Metabolic and other diseases. A number of metabolic diseases, all relatively rare, cause short stature; they include certain diseases of the kidneys, the connective tissue, and the liver. Some children with congenital heart disease and some with bronchiectasis (inflammation and dilation of the bronchial tubes) are short. Children with various forms of mental defect are frequently small, for unknown reasons.

Tall stature. The majority of tall children simply inherit their size from tall parents. Two forms of pathological gigantism occur, however. One is caused by oversecretion of pituitary growth hormone, often due to a pituitary tumour, and the other, called cerebral gigantism, is of unknown cause. In the latter condition the tall size is associated with a characteristic head and face, clumsiness, and often some degree of mental defect.

BIBLIOGRAPHY. B.T. DONOVAN and J.J. VAN DER WERF TEN BOSCH, *Physiology of Puberty* (1965), a standard monograph dealing with events and control of puberty for the most part in experimental animals; F. FALKNER (ed.), "Child Development: An International Method of Study," *Annales Paediatrici*, suppl. no. 72 (1960), a description with illustrations of the methods of taking body measurements of children; *Human Development* (1966), a text with chapters by many contributors varying somewhat in their approach and covering both psychological and physical development; W.A. MARSHALL and J.M. TANNER, "Variations in Pattern of Pubertal Changes in Girls," *Arch. Dis. Child.*, 44:291–303 (1969), and "Variations in the Pattern of Pubertal Changes in Boys," *Arch. Dis. Child.* 45:13–23 (1970), descriptive papers; J.M. TANNER, *Education and Physical Growth* (1961), a standard textbook on physical growth, written for teachers and used for colleges of education but on a level also appropriate for the layman; *Growth at Adolescence*, 2nd ed. (1962), a more advanced textbook than the above, suitable chiefly for pediatricians and biologists; "Earlier Maturation in Man," *Scient. Am.*, 218:21–27 (1968), a popular article on earlier maturing; "Growth and Endocrinology of the Adolescent," in L. GARDNER (ed.), *Endocrine and Genetic Diseases of Childhood* (1969), an up-to-date and advanced description in a standard medical textbook; J.M. TANNER and G.R. TAYLOR, *Growth* (1965), readings for high school students and laymen distinguished by beautiful illustrations; J.M. TANNER, R.H. WHITEHOUSE, and M. TAKAISHI, "Standards from Birth to Maturity for Height, Weight, Height Velocity and Weight Velocity, British Children 1965," *Arch. Dis. Child.*, 41:454–471 (1966), an advanced and somewhat difficult paper which forms the basis for most of the standards used in present day pediatrics.

(J.M.T.)

Development, Plant

Development refers to the sequence of changes undergone by an organism from the time of its origin until its senescence and death. It encompasses all aspects of growth and differentiation of parts. The potentiality for development is inherited in the genes in the form of information coded in molecules of deoxyribonucleic acid (DNA). Although both plants and animals share this chemical basis of inheritance and of translation of the code into structural units called proteins, plant development differs from that of animals in several important ways. Higher plants sustain growth throughout life and, in this sense, are perpetually embryonic; animals, on the other hand, generally have a determinate period of growth, after which they are considered mature. Furthermore, both growth and organ formation in plants are influenced by their possession of a rigid cell wall and a fluid-filled space called the vacuole, two features unique to the plant cell. Conversely, certain features of animal cells are absent in plants. Notable is the lack of cellular movements and fusions that play an important part in tissue and organ development in higher animals.

Uniqueness of plant development

This article is divided into the following major sections:

I. General features
Life cycles
Body plans
Preparatory events

II. Early development: from zygote to seedling
Embryo formation
Germination and early growth
III. Later development: the sporophyte plant body
Continuation of organ formation
The shoot system and its derivatives
The root system and its derivatives
IV. Correlations in plant development
Coordination of shoot and root development
Determination of mature form
Seasonal adaptations
Senescence and death

I. General features

LIFE CYCLES

The life cycle of all tracheophytes (vascular plants), bryophytes (mosses and liverworts), and many algae and fungi is based on an alternation of generations, or different life phases: the gametophyte, which produces gametes, or sex cells, alternating with the sporophyte, which produces spores. Gametophytes develop from the spores and, like them, are normally haploid; *i.e.*, each cell has one set of chromosomes. Sporophytes develop from a fertilized egg, or zygote, that results from the fusion of gametes (fertilization) formed by the gametophytes; they are accordingly diploid (*i.e.*, each cell has two sets of chromosomes). Although the two generations are phases of one life cycle, they have independent developmental histories; each begins life as a single cell and passes through a juvenile period before reaching maturity and giving rise to the alternate phase.

In various algae and fungi the two generations are alike in form (*i.e.*, are isomorphic), and, despite the difference in chromosome number, their development follows essentially identical pathways. More commonly, however, the alternating generations have different forms (*i.e.*, are heteromorphic); this is true for the bryophytes and for all vascular plants, both angiosperms (flowering plants) and gymnosperms (conifers and allies). General rules for vascular plants are that the sporophyte generation is physically the larger, has a more complex developmental history, produces a greater range of cell types, and expresses a more diverse biochemistry; the gametophyte is often diminutive, reduced in the case of the angiosperms to a mere few cells. In the bryophytes, the gametophyte generation, rather than the sporophyte, is the more conspicuous.

Conveying the genes through alternate generations

Although the gametophyte generation in vascular plants is small and has limited physiological capabilities, its cells must convey genes capable of directing the sporophytic developmental pattern, because the pattern is transmitted through the gametes to the zygote (Figure 1). The expression of "sporophytic" genes must therefore be repressed in the gametophyte, probably from the time of spore formation (sporogenesis). Correspondingly, events associated with gamete formation (gametogenesis) or fertilization must somehow free the sporophytic genes and thus permit the zygote to enter the sporophytic developmental pattern. Although it might be supposed that the "switch" is associated with the difference in chromosome number between the haploid spore (a single set) and the diploid zygote (a double set), this has been shown not to be the determining factor.

The alternation of generations illustrates an important principle, namely that cell lineages arising from single parental cells containing the same genetic potentiality may pursue mutually exclusive developmental patterns. Channelling, or canalizing, events of this nature occur repeatedly in the course of development of an individual plant, beginning with the pattern of cell division from the very first cleavage of the zygote cell.

BODY PLANS

Collectively plants manifest a wide range of body plans, ranging from the single cell (or unicell), with a single nucleus, through various types of colonial and filamentous forms to massive multicellular structures. (Algae, including the single-celled forms, have a great deal in common in structure and biochemistry with vascular plants. Bacteria and fungi stand somewhat apart, but because of their various plantlike qualities they are taken as plants for the purposes of this article.)

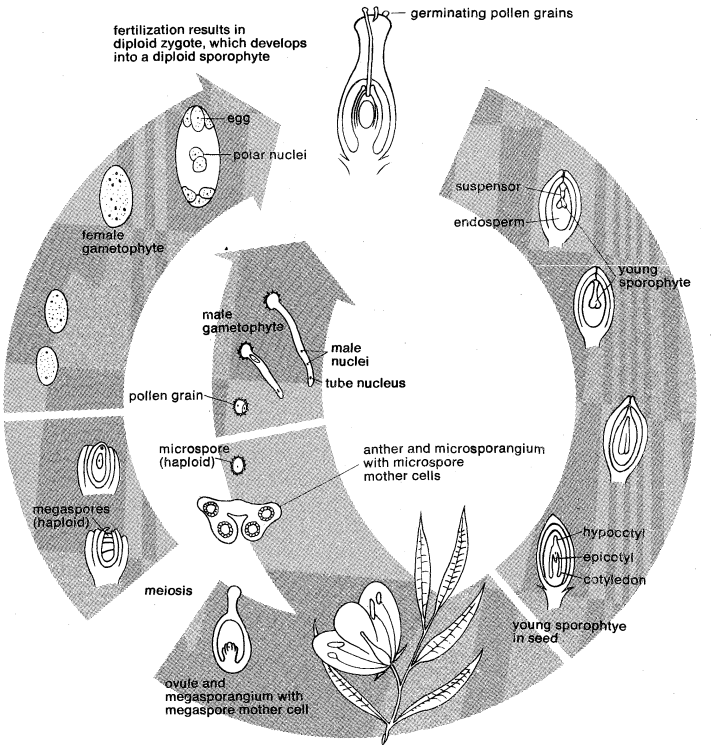


Figure 1: Stages in the fertilization and seed formation of a dicot.
From C.A. Villee, *Biology*, 6th ed. (1972); W.B. Saunders Co.

For the unicell, development is the same as cell differentiation. Although many unicellular fungi and algae show little differentiation other than that connected with reproduction, others undergo elaborate structural changes that illustrate many principles basic to development in multicellular plants. An important example is the green alga *Acetabularia*. This alga first produces a rootlike system and stalk and then, later, a flattened umbrella-like cap. The developmental potentialities of this unicell, with its single nucleus, are, however, limited; in order for there to be any advance beyond the state seen in *Acetabularia*, with the development of greater body mass and a division of labour among different parts, an increase in the number of participating nuclei seems obligatory.

One method of providing more nuclei is by nuclear division without a corresponding cell division; the result is a coenocytic structure. Plants with this type of multinucleate organization show considerable diversity; examples are found in both algae and fungi. Growth occurs by the extension of the cell wall in certain zones, usually at the tips of filaments, and structural differentiation results from branching and the specialization of parts for particular functions. The aggregation of coenocytic filaments can lead to the development of a three-dimensional body, or thallus, but plants with this type of organization have not achieved great size.

Multicellularity: an advance in development

A more significant type of body plan, one based on the multicellular filament, is found in its most simple form in certain algae known as diatoms, in which chains of cells of indefinite length arise, although the cells show no evidence of interaction. More advanced is the condition of many other algae, in which there occur branches that may either be identical with the original filament or show structural or physiological specialization. This condition occurs in certain green algae, in which the main branches creep and the lateral branches grow erect; such diversification represents an important developmental innovation and, possibly, the evolutionary beginnings of organ specialization in plants.

Three-dimensional body forms may evolve from the association of cells in colonies. Cells among the colonial

green algae are of definite number; each component cell resembles a free-living unicell, but all are united by cellular connections, or plasmodesmata, which may be important in coordinating the development of the colony. Colonies are often of precise geometric shape, forming either a circular plate or a sphere. In elaborate ones, certain cells are specialized for reproduction, and others are concerned primarily with movement.

Another developmental pathway resulting in more massive body structure is by the association of filaments. The reproductive structures of many fungi are composed of large numbers of closely interwoven filaments, which, although not physically connected, do interact in some way to produce structures such as the mushroom cap. Several filamentous body plans are found among the red algae. In one, a single main (axial) filament grows at one end, but, just behind its tip, cells divide to produce a number of lateral filaments that grow parallel with the axial filament. The older parts of the thallus, therefore, seem to be an aggregate of filaments. More massive structures are produced when there are several axial filaments; and, by branching, particularly when accompanied by fusion, dense tissues resembling the basic undifferentiated tissue (parenchyma) of higher plants are formed. In these algae, cellular connections occur between daughter cells of a filament, and others may develop secondarily between cells of neighbouring filaments.

Three-dimensional body form

The transition from a filamentous to a three-dimensional form appears most notably in the brown algae. In certain brown algae, growth is by an axial filament, but, behind the tipmost cell, divisions produce a denser tissue lacking evidence of filamentous organization. In the sporophytes of kelp, one of the largest and most complex of the algae, cell division often is restricted to areas comparable to the growing tips of vascular plants, and, although a filamentous organization may be evident in the centre of the thallus, the surrounding cortical regions are composed of a tissue that is essentially undifferentiated. (The gametophytes of kelp, however, have a simple filamentous organization).

Among nonvascular plants, true parenchyma is found in the bryophytes, in both the gametophyte and sporophyte phases. The development of the moss gametophyte illustrates the transition from a filamentous to a highly organized three-dimensional growth form. The moss spore germinates into a filamentous plant, the protonema, which later produces a leafy shoot. This type of transition from simple to more complex growth form is accompanied by the synthesis of new kinds of ribonucleic acids (RNA's), presumably through the activation of genes that were not expressed during the early growth of the gametophyte.

Much of the remainder of this article is concerned with the development of the complex body forms of vascular-plant sporophytes, which do not normally pass through any filamentous stages. It may be noted, however, that, in the course of evolution, the capacity for this type of growth has not been lost, since it may be adopted by cells grown in tissue cultures in the laboratory.

PREPARATORY EVENTS

The sporophytes of all vascular plants produce cells called spore mother cells—since they will give rise to spores—in spore cases (sporangia). Spore mother cells are usually surrounded, during development, by a special nutritive tissue. In the more primitive groups each sporangium holds many mother cells. This is true also in the pollen-producing sporangia of gymnosperms and angiosperms but not in the egg-producing sporangia (ovules), which usually have only one mother cell.

In certain lower vascular plants, typified by the club moss *Selaginella,* the gametophyte is formed entirely—or almost entirely—within the spore wall. Two kinds of gametophytes develop from the two kinds of spores produced by the sporophyte in different sporangia; the larger spore (megaspore) gives rise to the female gametophyte, the smaller spore (microspore) to the male. This condition is referred to as heterospory. The gametophytes, or prothalli, of other club mosses and most horsetails and ferns are sexually undifferentiated and arise from one kind of spore, a condition termed homospory.

In these groups the gametophytes develop as free-living and independent plants that ultimately produce the gametes. In general, the male gametes (antherozoids) are produced in globose structures (antheridia) that are either stalked or sunken in the gametophyte. The antherozoids, always many in number, develop from mother cells enclosed in the jacket of the antheridium. Each antherozoid can move by using its whiplike hairs, or flagella, two or three (in the lycopods) or many (in the horsetails and ferns). The female gametes are formed singly in flask-shaped structures (archegonia) that also are either stalked or sunken in the gametophyte. The neck of the flask is closed by neck canal cells, which later break down to permit the entry of the male gamete. The egg itself lies in the basal part, or venter, of the flask, with a ventral canal cell above it. When the male gametes, or antherozoids, are released by the rupture of the antheridium, they swim in a water film to the archegonia and effect fertilization.

Among the gymnosperms the male gametophyte is much reduced and is a parasite on the sporophyte for only a short time. Cell cleavages within the spore wall cut off a prothallial cell, which will give rise to the vegetative (*i.e.*, nonreproductive) part of the plant, and an antheridial cell, which divides into a tube cell and a generative cell. The male gametophyte so formed and contained within the spore wall is the pollen grain. After transfer to the ovule by wind, the pollen grain germinates to form a tube, and the generative cell divides into two cells, one of which forms the male gametes by further division. The gametes bear numerous spirally arranged flagella. The female gametophyte meanwhile develops entirely within the parent sporangium in the ovule. The size of the single functional spore increases greatly as the spore nucleus divides repeatedly to produce numerous free nuclei. Cell-wall formation then begins at the periphery, extending inward until the whole area is divided into cells. Up to four archegonia are formed, sunken in the tissue of the gametophyte, each with a female gamete, or egg.

Fertilization: a change of generations

The end of the gametophyte phase and the beginning of the sporophyte phase occur at fertilization, when one of the male gametes fuses with the female gamete to form the zygote, which will then develop as the sporophyte. (Development of the sporophyte can, in some cases, be triggered by means other than fertilization, in which case the organism is said to arise parthenogenetically.)

The male gametophyte of angiosperms is reduced to three cells, one so-called vegetative cell and two male gametes. The division producing the gametes may occur either before dispersal of the pollen grain or later, during the growth of the pollen tube. The female sporangium has one or two coats, or integuments, except for an opening (micropyle) at one end; the sporangium with an integument is called the ovule. The female gametophyte, known in this group as the embryo sac, develops from the parent spore while it is still retained in the sporangium. Three cell divisions result in eight nuclei, which arrange themselves so that three lie at each end and two lie in the centre. The cytoplasm then cleaves and three cells are formed at each pole, leaving two nuclei in a large central cell. The three cells at the micropylar pole (end toward the micropyle) form the egg apparatus. Two of these cells, called synergids, correspond to the neck cells of an archegonium; the third is the egg cell. The three cells at the opposite pole, the antipodals, play a part in embryo nutrition in certain genera. The two polar nuclei in the central cell ultimately unite, becoming the fusion nucleus. The pollen grain is transferred by various agencies (wind, water, animals) to the stigma of the female flower, and, as in the gymnosperms, it germinates to produce a tube. This tube grows through intervening tissues, through an opening (micropyle) of the egg, and enters a cell near the micropyle (synergid), in which the two male gametes are discharged. The unique feature of this phase of angiosperm development is that two fertilizations occur. One male gamete fuses with the

egg to give the diploid zygote; the other makes its way to the fusion nucleus in the central cell, already diploid, and by a second fusion gives a triploid primary endosperm nucleus, which is later concerned in the formation of the nutritive tissue, or endosperm.

II. Early development: from zygote to seedling

EMBRYO FORMATION

Cleavage of the zygote. In vascular plants embryo formation, or embryogenesis, usually occurs within a few hours after fertilization, with the first cell division that cleaves the zygote, or fertilized egg, into two daughter cells. Thereafter, rapid cell division provides the building blocks of the primary organs of the embryo sporophyte: the first root, first leaves, and the shoot apex. Temporary structures concerned with embryo nutrition —suspensor and foot—may also be produced. These organs originate in a polarization established at the time of zygote cleavage, but the details of their development vary widely among the different groups.

In the club mosses the zygote divides in a plane at right angles to the axis of the archegonium. The daughter cell toward the neck forms a short filament of cells, the suspensor; the inner cell gives rise to the other organs of the embryo, the shoot, root, and foot. The axis of the embryo is inclined to that of the archegonium and may be almost at right angles. This is in contrast to the behaviour of the true mosses, in which the embryo is oriented along the length of the archegonium, with the foot directed inward and the structures that are equivalent to the shoot, namely the spore capsule and its stalk, directed toward the neck.

A polarity like that of the mosses appears in the horsetails, in which the zygote divides by transverse and longitudinal walls to form a group of four cells. Of these, the two cells toward the neck give rise to the shoot system; the inner two produce the foot and root.

Embryogenesis in gymnosperms and angiosperms

The details of early embryogenesis in gymnosperms vary considerably. In the cycads and ginkgoes, the initial cleavage establishes a polarity opposite to that in the horsetails, the inner cell giving rise to the shoot and the outer producing the root. Many conifers are unique in that the zygote undergoes a period of free-nuclear division without cell formation, producing usually four or eight nuclei, which move to the end of the zygote, away from the neck cells, where cleavage begins. In the pines a further division gives four tiers of four cells. The intermediate tiers extend greatly to form a suspensor; each of the four cells at the lower pole may act as the parent cell of an embryo, a condition sometimes referred to as polyembryony.

In contrast, there is no free-nuclear stage in angiosperm embryogenesis. The zygote cleaves by a wall more or less at right angles to the axis of the embryo sac. The daughter cell next to the micropyle (basal cell) produces a suspensor and contributes to the root; the inner (terminal) cell gives rise to the shoot system. (Angiosperm embryogenesis is more fully described in the following section dealing with the origin of primary organs.)

Notwithstanding the variation in the different groups, the pattern of development established in the early cell cleavages is consistent. The primary polarization of the zygote must necessarily be imposed by the adjacent tissues of the sporophyte, but thereafter the fate of daughter cells depends on control established within the young sporophyte itself.

Although it is often possible to specify the origin of the cell lineages contributing to the various organs and tissue layers, a geometric regularity in cell division is generally maintained through only the first few division cycles in the embryo. The final form of the embryo is thus determined not through the specification of a precise scheme of cell division, as in the development of colonial algae, but through an overall control in which cell and tissue interactions play an important part.

Origin of the primary organs. Angiosperm embryogenesis can be described in terms of a much studied flowering plant called shepherd's purse (*Capsella bursa-pastoris*). The zygote divides into two cells, the terminal cell and the basal cell. The terminal cell divides by a wall formed at right angles to the first cleavage wall and then again by a wall formed at right angles to this; a quadrant of cells is thus formed (Figure 2A–D). The partition of the quadrant cells in a transverse plane then produces an octant stage (Figure 2E). By transverse divisions, the basal cell forms a filament, the suspensor, of up to ten cells, the end cell of which swells to form an absorbing organ. The attachment cell, or hypophysis, adjoins the octants derived from the terminal cell (Figure 2F).

After Soueges and Schaffner in P. Maheshwari, *Introduction to the Embryology of the Angiosperm* copyright © 1950; used with permission of McGraw-Hill Book Company

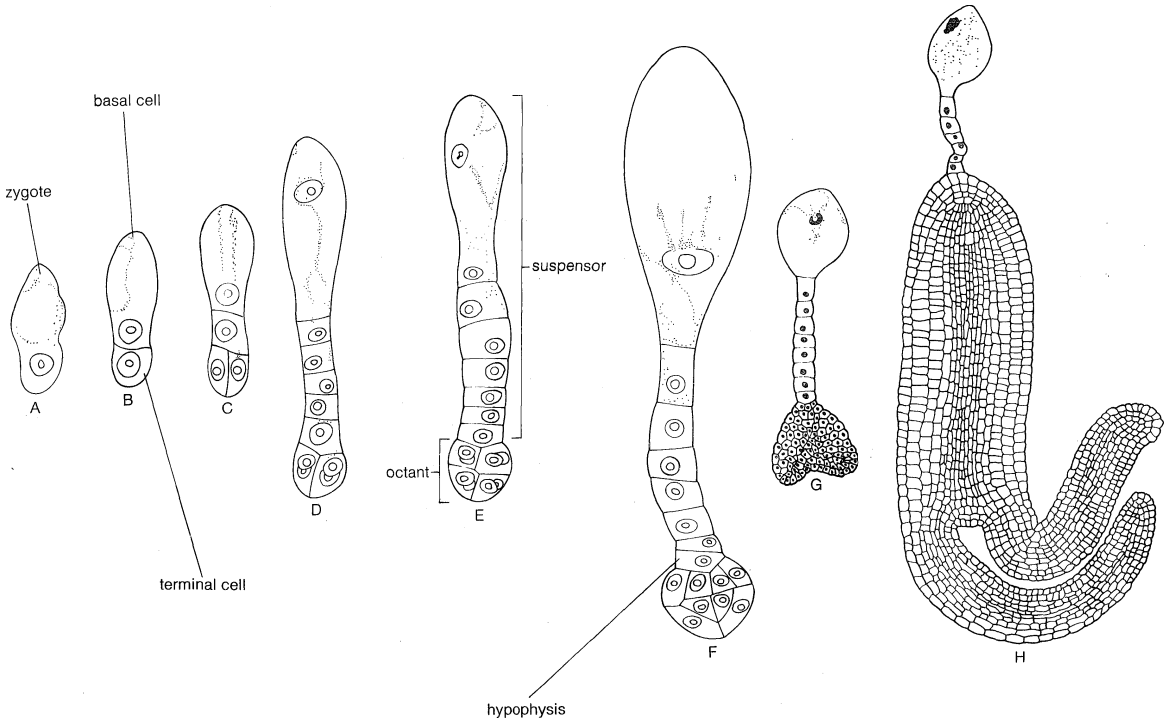


Figure 2: Development of the embryo in *Capsella bursa-pastoris* (see text; G and H less enlarged than A–F).

At this time, the prospective future of each of the zones of the embryo can be specified. Four cells of the octant group will ultimately produce the seed leaves (cotyledons) and the shoot apex; the other four will form the hypocotyl, the part of the embryo between the cotyledons and the primary root (radicle). The hypophysis will give rise to the radicle and the root cap; the cells of the suspensor will degenerate as the embryo matures.

Establishing the embryo

The zones of the embryo destined to form the principal organs are established by this first sequence of divisions, and tissue layers are defined during the ensuing divisions. The octant cells divide by curved walls parallel to the surface; in this way the outer layer responsible for producing the epidermis of the shoot system is defined. Divisions of a more irregular nature in the inner zone ultimately define the tissues from which the central cylinder and vascular core of the main axis of the shoot will develop. Simultaneously, the hypophysis forms a group of eight cells by three successive divisions, the planes of which are mutually at right angles. Of these eight cells, the outer four produce the root cap and epidermis; the inner four contribute to the radicle.

The embryo is at first globular (Figure 2F), but it soon becomes heart-shaped by a combination of numerous cell divisions and enlargement in two zones of the outer hemisphere (Figure 2G). In this manner two cotyledons form. The volume of tissue between the cotyledons is the prospective shoot apex. The characteristic form of the apex is not established until after germination.

As the cotyledons become extended, the embryo bends, because of physical restraints, to conform with the cavity of the embryo sac (Figure 2H). From the heart-shaped phase onward, the core of the hypocotyl and the radicle appears as a cylinder of narrow and elongated cells. This is the parent tissue of the vascular system of the seedling. The surrounding tissue contributes the cortex layer of the stem and root.

The embryogenesis of *Capsella* illustrates only one of several patterns found among flowering plants. Among dicotyledons, the planes of division of the terminal cell, the form of the suspensor, and the contribution made by the basal cell to the embryo all provide evidence used in determining the embryogenetic plan.

Monocotyledons, flowering plants the seeds of which contain only one cotyledon, share with dicotyledons such as *Capsella* the main features of early embryogenesis, including the possession of a suspensor and, in most cases, a fairly regular progression of cell divisions to the octant stage. Thereafter the symmetrical growth pattern is lost through the development of the single cotyledon. In the lily family (Liliaceae), generally accepted as a primitive family of monocotyledons, the cotyledon is derived from an octad of cells arising from the terminal cell. The hypocotyl and stem apex are derived from the proximal cell of a short filament formed by the basal cell, and the root comes from the pair of cells next to it. The suspensor forms from the distal cell or cells of the filament. In the more advanced families of monocotyledons, including the grasses (Gramineae) and orchids (Orchidaceae), embryogenesis is much less regular. The grass embryo possesses structures that do not occur in any other flowering plants, namely, the scutellum, an organ concerned with the nutrition of the seedling, and the coleoptile and coleorhiza, protective sheaths of the young shoot and the radicle. The scutellum arises from octant cells, which also contribute to the cotyledon. The basal cell forms part of the coleoptile and also gives rise to the shoot apex and the tissues of the root and coleorhiza. The embryo is asymmetrical, with the shoot apex lying on one side in a notch, ensheathed by the coleoptile.

In marked contrast, embryogenesis of the orchids is more simple. Except when a suspensor is formed, early cleavages follow no well-defined plan, and the product is an ovoid mass of tissue called the proembryo. No cotyledon, stem apex, or root apex is organized in this early period; these organs do not appear until after germination has occurred.

Nutritional dependence of the embryo. During their early growth, the embryos of all vascular plants exist as virtual parasites depending for nutrition on either the gametophyte or the previous sporophyte generation through the agency of the gametophyte or, in the special case of the angiosperms, upon an initially triploid tissue, the endosperm, which is itself nourished by the parent sporophyte.

Embryos as parasites

The early nutrition of the sporophyte in ferns, horsetails, and club mosses such as *Lycopodium* is clearly provided by the gametophyte. In these groups the young sporophyte produces a multicellular structure, the foot, which remains embedded in the tissues of the gametophyte throughout early development withdrawing nutrients. Ultimately, both shoot and root of the sporophyte grow out from the gametophyte, but, even after the first leaf has begun to photosynthesize and thus to produce its own food, the gametophyte may persist.

In *Selaginella*, the gametophytes are sexually distinct. The female gametophyte develops within the wall of the megaspore. The archegonia are exposed after the megaspore wall splits, but the gametophyte never escapes completely. After fertilization, the zygote cleaves, and the outer cell produces a long suspensor that pushes the embryo deeply into the tissues of the gametophyte. A foot is then formed, as in *Lycopodium*, and further development of the embryo continues at the expense of reserves transferred to the megaspore from the preceding sporophyte generation.

There are superficial similarities between the nutritional history of the embryo in gymnosperms and in *Selaginella*, for, in each, the female gametophyte, dependent upon reserves derived from the sporophyte, acts as an intermediary between one sporophyte generation and the next.

In the pines, the female gametophyte develops within the tissues of the nucellus and acquires abundant food reserves. The proembryo forms after a period of free-nuclear division in the zygote, and the tier of cells above the basal four then elongates to form a suspensor, which pushes the embryonic group deep into the gametophyte (Figure 3). Secondary suspensor cells may form from the basal tier to continue the process. During embryogenesis, the gametophyte continues to grow and to accumulate food materials, which are transferred to the embryo or remain as reserves in the seed.

The female gametophyte of angiosperms never acquires copious reserves, although starch is frequently present in the central cell and sometimes in the egg itself. The unique feature, here, is that the embryo is nutritionally dependent upon the endosperm, a tissue that, in the genetical sense, constitutes a third organism—neither gametophyte nor sporophyte. Furthermore, as a tissue the endosperm manifests several other special characteristics. The nuclei have three chromosome sets and, therefore, three times the deoxyribonucleic acid (DNA) of haploid cells. As nuclear division ends, the amount of DNA per nucleus increases still further, a condition comparable with that in various plant- and animal-gland nuclei, presumably connected with the nutritional function of the endosperm. Nuclear division takes place at first without cell-wall formation so that a coenocyte is produced; later, partitioning of the cytoplasm results in a cellular tissue.

The food reserves

The reserves accumulated in the endosperm include carbohydrates (especially starch), lipids, and proteins. As reserves accumulate, the nuclei of the endosperm cells may undergo deformation and degeneration. In many plants the growing embryo consumes the endosperm before seed maturation; in others, the tissue persists in the seed, providing a reserve for the developing seedling after germination. Endosperm is not formed in certain angiosperms. In such cases the embryo depends on the transfer of nutrients directly from the sporophyte.

Tissues other than the endosperm may become specialized for the early nutrition of the embryo. The antipodal cells of the female gametophyte sometimes acquire glandular properties, as may cells of the nucellus surrounding the embryo sac. In some species the embryo itself develops a suspensor that penetrates the tissues of the parent sporophyte and acts as an absorbing organ.

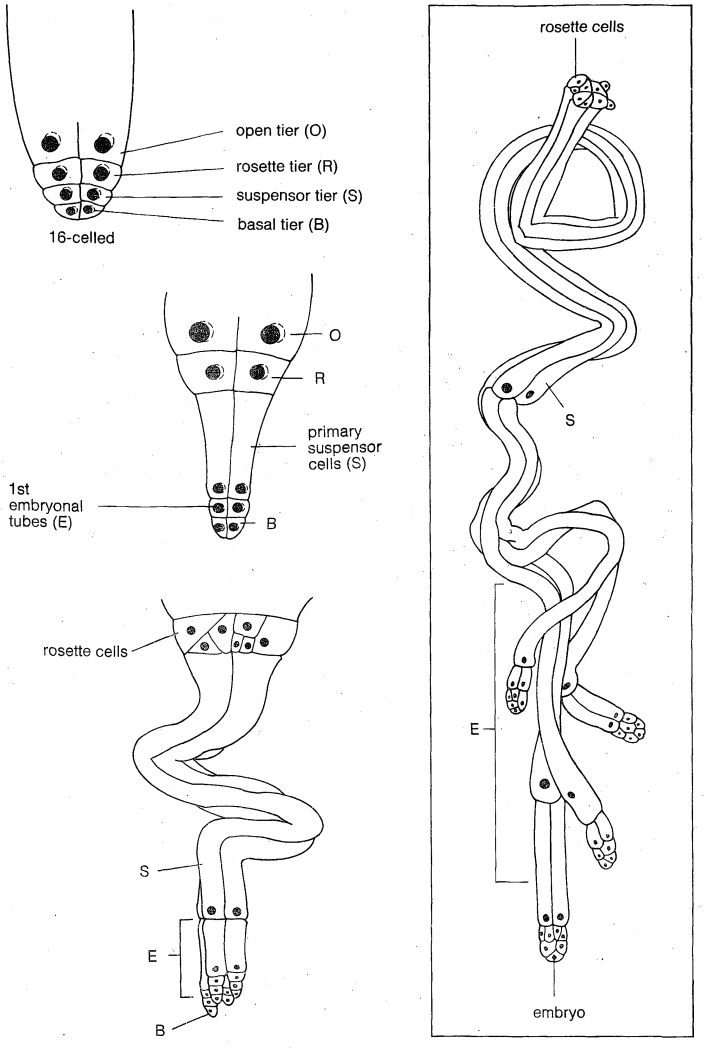


Figure 3: Fertilization and early embryology in *Pinus*.
From *Comparative Morphology of Vascular Plants* by Adriance S. Foster and Ernest M. Gifford, Jr. W.H. Freeman and Company. Copyright © 1959

Dormancy of the embryo. Among the lower pteropsids (club mosses, horsetails, and ferns), the principal agent of dispersal is the haploid spore and not, as in gymnosperms and angiosperms, the seed, the ripened ovule containing a dormant embryo. Since the embryo of lower pteropsids is not involved in dispersal, it does not usually undergo any marked period of dormancy after the differentiation of the primary organs. Development instead proceeds continuously through dependence upon the gametophyte until the young sporophyte is established as a physiologically independent plant. The embryos of gymnosperms and angiosperms pass into a state of dormancy soon after the differentiation of the primary organs and the sporophyte is dispersed in a seed.

In the period leading up to dormancy, several changes occur in the embryo. The accumulation of reserves in the cotyledons or elsewhere ceases, respiratory rate declines rapidly, and cell division, with associated protein and nucleic-acid synthesis, stops. Correlated with these events are cellular changes typical of tissues with low metabolic activity. Especially obvious is the general dehydration of the cells that constitute the seed and the thickening of the cell walls of the ovule to form the seed coat (testa). The product is a structure in which the embryo is protected from temperature extremes by its state of desication and is often guarded from further drying and from mechanical or biological degradation by the seed coats. The seed coat often contributes to the maintenance of dormancy by physically impeding the passage of water and gases to and from the embryo, by chemically inhibiting germination, and by mechanically restricting the growth of the embryo.

Role of the seed coat in dormancy

GERMINATION AND EARLY GROWTH

Dormancy is brief for some seeds, for example those of certain short-lived annual plants. After dispersal and under appropriate environmental conditions, such as suitable temperature and access to water and oxygen, the seed germinates, and the embryo resumes growth.

The "breaking" of dormancy. The seeds of many species do not germinate immediately after exposure to conditions generally favourable for plant growth but require a "breaking" of dormancy, which may be associated with change in the seed coats or with the state of the embryo itself. Commonly the embryo has no innate dormancy and will develop after the seed coat is removed or sufficiently damaged to allow water to enter. Germination in such cases depends upon rotting or abrasion of the seed coat in the soil. Inhibitors of germination must be either leached away by water or the tissues containing them destroyed before germination can occur. Mechanical restriction of the growth of the embryo is common only in species that have thick, tough seed coats. Germination then depends upon weakening of the coat by abrasion or decomposition.

In many seeds the embryo cannot germinate even under suitable conditions until a certain period of time has lapsed. The time may be required for continued embryonic development in the seed or for some necessary finishing process—"after ripening"—the nature of which remains obscure.

Low temperatures and seed germination

The seeds of many plants that endure cold winters will not germinate unless they experience a period of low temperature, usually somewhat above freezing. Otherwise germination fails or is much delayed, with the early growth of the seedling often abnormal. (This response of seeds to chilling has a parallel in the temperature control of dormancy in buds.) In some species, germination is promoted by exposure to light of appropriate wavelengths; in others, light inhibits germination. For the seeds of certain plants, germination is promoted by red light and inhibited by light of longer wavelength, in the "far red" range of the spectrum. The precise significance of this response is as yet unknown, but it may be a means of adjusting germination time to the season of the year, or of detecting the depth of the seed in the soil. Light sensitivity and temperature requirements often interact, the light requirement being entirely lost at certain temperatures.

In the process of germination, water is absorbed by the embryo, which results in the rehydration and expansion of the cells. Shortly after the beginning of water uptake, or imbibition, the rate of respiration increases, and various metabolic processes, suspended or much reduced during dormancy, resume. These events are associated with structural changes in the organelles (membranous bodies concerned with metabolism), in the cells of the embryo.

The emergence of the seedling. Active growth in the embryo, other than swelling resulting from imbibition, usually begins with the emergence of the primary root from the seed, although in some species (*e.g.*, the coconut) the shoot emerges first. Early growth is dependent mainly upon cell expansion, but, within a short time, cell division begins in the radicle and young shoot; thereafter, growth and further organ formation (organogenesis) are based upon the usual combination of increase in cell number and enlargement of individual cells.

Until it becomes nutritionally self-supporting, the seedling depends upon reserves provided by the parent sporophyte. In angiosperms these reserves are found in the endosperm, residual tissues of the ovule, or in the body of the embryo, usually in the cotyledons. In gymnosperms, food materials are contained mainly in the female gametophyte. Since reserve materials are partly in insoluble form—as starch grains, protein granules, lipid droplets, and the like—much of the early metabolism of the seedling is concerned with mobilizing these materials and delivering, or translocating, the products to active areas. Reserves outside the embryo are digested by enzymes secreted by the embryo and, in some instances, also by special cells of the endosperm.

In some seeds (*e.g.*, castor beans) absorption of nutrients

from reserves is through the cotyledons, which later expand in the light to become the first organs active in photosynthesis. When the reserves are stored in the cotyledons themselves, these organs may shrink after germination and die or develop chlorophyll and become photosynthetic.

Orientation of the seedling

Environmental factors play an important part not only in determining the orientation of the seedling during its establishment as a rooted plant but also in controlling some aspects of its development. The response of the seedling to gravity is important. The radicle, which normally grows downward into the soil, is said to be positively geotropic. The young shoot, or plumule, is said to be negatively geotropic, because it moves away from the soil; it rises by the extension of either the hypocotyl, the region between the radicle and the cotyledons, or the epicotyl, the segment above the level of the cotyledons. If the hypocotyl is extended, the cotyledons are carried out of the soil, but, if the epicotyl elongates, the cotyledons remain in the soil.

Light affects both the orientation of the seedling and its form. When a seed germinates below the soil surface, the plumule may emerge bent over, thus protecting its delicate tip, only to straighten out when exposed to light (the curvature is retained if the shoot emerges into darkness). Correspondingly, the young leaves of the plumule in such plants as the bean do not expand and become green except after exposure to light. These adaptative responses are known to be governed by reactions in which the light-sensitive pigment phytochrome plays a part. In most seedlings, the shoot shows a strong attraction to light, or a positive phototropism, which is most evident when the source of light is from one direction. Combined with the response to gravity, this positive phototropism maximizes the likelihood that the aerial parts of the plant will reach the environment most favourable for photosynthesis.

III. Later development: the sporophyte plant body

CONTINUATION OF ORGAN FORMATION

Although it is convenient to refer to the early development of the plant sporophyte from the fertilized egg as embryogenesis, the process is never actually concluded as it is in the higher animals. In vascular plants, organ formation (organogenesis) is not confined to early life, and the processes of shoot, root, and leaf formation that occur first in the embryo are repeated, albeit in modified form, throughout the life of the plant. The life-span may be short and determinate, as in annual plants such as the cereals, or long, lasting for many years—indeed potentially indefinitely, except for limitations imposed by the environment and accidents—as in trees. The protracted growth of perennials, or plants that resume growth each growing season, tends to lead to increase in size, but bulk is not necessarily directly correlated with age, because individual leaves, flowers, and even whole limbs continuously die and are shed. Some long-lived plants, however, do reach a point at which losses of body mass balance the increase resulting from continued growth and organ formation.

The activity of meristems. Characteristically, vascular plants grow and develop through the activity of organ-forming regions, the growing points. The mechanical support and additional conductive pathways needed by increased bulk are provided by the enlargement of the older parts of the shoot and root axes. New cells are added through the activity of special tissues called meristems, the cells of which are small, intensely active metabolically, and densely packed with organelles and membranes, but usually lacking the fluid-filled sacs called vacuoles. Meristems may be classified according to their location in the plant and their special functions. One important distinction is between persistent meristems, typified by those of the growing points, and meristems with a limited life, those associated with organs, such as the leaf, of determinate growth. The regions of rapid cell division at the tips (apices) of the stem and the root are terminal meristems. In the stem apex, the uppermost part is the promeristem, below which is a zone of transversely oriented early cell walls, the file, or rib, meristem. The procambium is a meristematic tissue concerned with providing the primary tissues of the vascular system; the cambium proper is the continuous cylinder of meristematic cells responsible for producing the new vascular tissues in mature stems and roots. The cork cambium, or phellogen, produces the protective outer layers of the bark.

Classification of meristems

Among meristems of limited existence is the marginal, or plate, meristem responsible for the increase in surface area of a leaf; it contributes new cells mainly in one plane. Another type of meristem of limited life is called intercalary; it is responsible for the extension of some stems (as in the grasses) by the addition of new tissues remote from the growing points.

The number of dividing cells in persistent meristems remains roughly constant, with one of the daughter cells of each division remaining meristematic and the other differentiating as a component of a developing organ. The geometrical arrangements in the particular organ determine the way in which this occurs, but in general the consequence is that the meristem is continuously moving away from the maturing tissue as growth continues. It remains, therefore, a localized zone of specialized tissue, never becoming diluted by the interposition of expanding or differentiating cells. In organs such as leaves, flowers, and fruits, in which the growth is determinate, the divisions of meristematic cells become more widely scattered, and the frequency progressively falls as the proportion of the daughter cells that differentiate increases. Ultimately, at maturity, no localized meristem remains.

The contribution of cells and tissues. The two major factors determining the forms of plant tissues and organs are the orientation of the planes of cell division and the shapes assumed by the cells as they enlarge. Clearly, if the division planes in a cell mass are randomly oriented and individual cells expand uniformly, the tissue will enlarge as a sphere. On the other hand, if cell division planes are oriented regularly or the expansion of individual cells is directional, the tissue can assume any of a number of shapes. In a stem, for example, the cell division planes of the promeristem are oriented at various angles to the stem axis, so that new cells produced contribute to both width and length. Below this region, in the rib meristem, the proportion of divisions with the cell plate at right angles to the axis increases, so that the cells tend to be oriented in files. The cells in these files expand vertically more than they do horizontally, and, accordingly, the stem develops as a cylinder.

The factors that control the orientation of cell division planes in meristems are largely unknown. Cell interactions, however, are presumed to coordinate the distribution and orientation of the divisions. In each cell microtubules in the cytoplasm help to orient the nucleus before it divides. Then, at the time of the division, other microtubules arranged in a spindle-shaped figure (the mitotic spindle) are involved in separating the daughter chromosomes and moving them to opposite ends of the parent cell. Thereafter, the residual part of the spindle helps to locate the plate that separates the two daughter cells. Microtubules are also concerned in determining the direction of growth in expanding cells, since they appear to influence the construction of the cell wall by controlling the way cellulose is laid down in it.

Influence of the mitotic spindle

Although change in shape is a form of cell differentiation, the term in the more general sense refers to a change in function, usually accompanied by specialization and the loss of the capacity for further division. Biochemical differentiation often involves a change in the character of the cell organelles—as when a generalized potential pigment body (proplastid) matures as a chloroplast, a chlorophyll-containing plastid. But it may also involve structural changes at a subcellular level, as when organelles change their character in cells engaged in intense metabolic activity.

The differentiation of plant cells for the movement of materials and the provision of mechanical support or protection invariably depends upon modification of the walls. This usually entails the accretion of new kinds of

wall materials, such as lignin in woody tissue and cutin and suberin in epidermal tissues and cork. The accompanying structural changes must be controlled, for the wall materials are not applied at random but according to a pattern appropriate to the particular cell or tissue. The development of patterns during cell-wall growth depends not only on the cytoplasmic microtubules, as in the construction of the cells that will give rise to the water-conducting vessels (xylem elements), but also on cytoplasmic membranes, as in the formation of sievelike end walls (sieve plates) in the cells that will give rise to food-conducting vessels (phloem elements).

The differentiation of xylem culminates in the death of the participating cells, and the vessels are formed of chains of empty walls. This is an example of "programmed death," not an uncommon phenomenon in plant and animal development.

THE SHOOT SYSTEM AND ITS DERIVATIVES

The shoot tip. The gametophytes of mosses and liverworts and the sporophytes of many higher plants have a shoot, or early stem, with a single cell at its tip, or apex, from which all the tissues of the stem arise. This apical cell is usually four-sided (tetrahedral), with three faces directed downward, and the fourth capping the apex. Daughter cells are continually cut off sequentially from the three inner faces, the apical cell preserving its tetrahedral shape. In cell lineages derived from the daughter cells, the division planes may remain oriented in a more or less regular manner, so that, for some distance below the apex, the three sectors can be recognized in the stem. This basic pattern occurs in the arrangement of the "leaves" of some mosses, which lie in three ranks. In many plants, however, division planes in the lower part of the apex show no particular correlation with the planes of cleavage of the apical cell, and the lateral appendages do not reflect any three-part arrangement.

The generative role of promeristem tissue

Gymnosperm and angiosperm apices do not possess apical cells. The generative role is discharged by an ill-defined zone of tissue called the promeristem. Regularities may appear in the distribution of division planes only in the extreme tip region. Over the outer part of the apex, the cells often appear to lie in one to three layers, which constitute the tunica (Figure 4). Enclosed by the tunica lies a core of cells that exhibit no distinct layering; this zone is the corpus. The layers of the tunica normally contribute to the surface layers of the plant, and the corpus provides the deeper lying tissues.

The tunica–corpus analysis emphasizes the orientation of division planes, but apices can be examined from other points of view—the sizes of cells, the degree of vacuolation, and the concentration of various cell constituents, especially ribonucleic acid (RNA), vary through the apex and this sometimes results in more or less distinctive zones. Both gymnosperm and angiosperm apices have been classified on the bases of such zonal patterns, but the validity of this approach, as well as its usefulness for understanding the function of the apex as a morphogenetic centre, has been questioned.

Since 1950, a theory of angiosperm apical zonation developed by French and Belgian botanists has been gaining support. This theory proposes that the central region of the apical dome constitutes a mass of cells with relatively low division rates, the *méristème d'attente*, or "waiting meristem." Surrounding this region is an annular zone of cells with higher division rates, the *anneau initial*, or "initiating ring." Features other than division rates characterize these zones: RNA and protein content are lower in the *méristème d'attente* than in the *anneau initial*, and the nucleoli are smaller. In longitudinal section, the differences contribute to the patterns distinguishable in apices, some of which have been used as bases for structural classification. The main contention of the Franco-Belgian school, however, is that the zonation represents a functional difference. The *méristème d'attente* is regarded as a region mainly concerned with controlling the geometry of the apex. The cells have a restricted metabolism concerned primarily in maintaining a low rate of increase in cell number, and they themselves, as well as their immediate derivatives, take no part in organogenesis or associated differentiation. The *anneau initial*, by contrast, is that part of the apex that produces the beginnings, or primordia, of lateral organs. Not only is the division rate higher, but the tissue as a whole is involved in metabolic syntheses that precede morphogenesis.

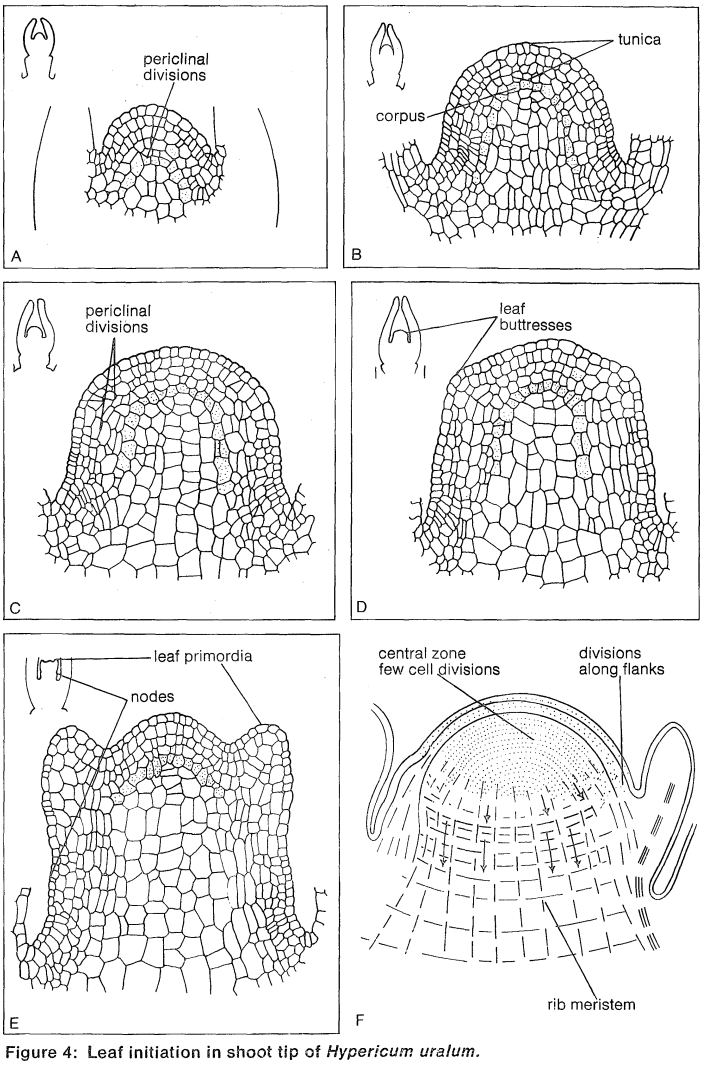


Figure 4: Leaf initiation in shoot tip of *Hypericum uralum*.
From (A-E) Zimmerman in K. Esau, *Plant Anatomy* (1965); John Wiley & Sons, Inc.; (F) F.C. Steward and M. Ram, *Advances in Morphogenesis*, 1: 245 Academic Press, Inc.

One difficulty in investigating the stem apex arises from the uncertainty about which aspects are important for the overall function: division planes, division frequency, metabolic patterns, or some combination of these. Still another complication results because the apex is in a state of constant change during the growth of the plant. A long-term developmental trend begins after the definition of the growing point in early embryogenesis and continues thereafter through juvenility and the period of vegetative growth into the reproductive phase. Superimposed on this trend is a cyclical change reflecting the periodic generation of the primordia of leaves and lateral shoots in the region immediately under the apex.

The production of leaves. Leaves originate on the flanks of the shoot apex. A local concentration of cell divisions marks the very beginning of a leaf; these cells then enlarge so as to form a nipple-shaped structure called the leaf buttress. The cells of the leaf buttress may be derived from the tunica alone or from both the tunica and the corpus.

In the early growth of the leaf primordium, new cells are contributed mainly by meristematic activity at the pole directed away from the stem, so that the buttress extends in length. The subsequent distribution of growth varies among the different groups of vascular plants ac-

cording to the shape of the mature leaf. In considering the angiosperms, a broad-leaved dicotyledon (tobacco) and a narrow-leaved monocotyledon (maize [corn]) will serve as examples (Figure 5).

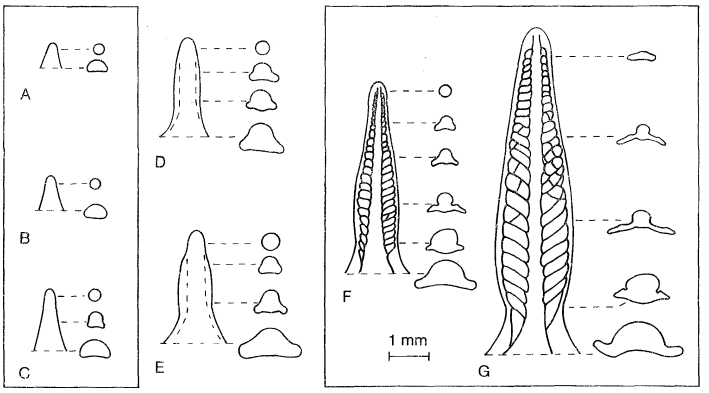


Figure 5: *Longitudinal and transverse sections of tobacco foliage leaf primordia in successive stages of development.* (A,B) The primordium in early development. (C) The small lateral ridges appearing about midway on the primordium are forerunners of the lamina. (D) Development of the lamina started. (E) Further expansion of the lamina and the beginning of development of the main lateral veins. (F) Primordium five millimetres in length showing development, toward the base, of the network of provascular strands. (G) Primordium with the development of the lamina and system of venation well started. Dotted lines in C–E represent external boundaries of midrib and lateral veins and not the provascular portion.

The tobacco leaf as an example

Apical growth dominates in the tobacco-leaf primordium until a height of about 0.5 millimetre (0.02 inch) is reached. Thereafter, the buttress becomes more and more flattened in the transverse plane by laterally oriented cell divisions and further expansion growth on either side. The dividing zones are the marginal meristems, through the activity of which the leaf gains its laminate form. In each meristem the outer file of cells, or marginal initials, contributes the epidermal layers by continued division. The cells below, the submarginal initials, provide the tissue of the inner part of the leaf. Usually a certain number of cell layers is defined in the mesophyll. Cell division is not limited to the region of the marginal meristems but continues throughout the leaf in each of the layers, always in the same plane, until the final cell number is approached. The rate then declines, ceasing in the different layers at different times. Divisions usually end first in the epidermis, then in the lower mesophyll layers of a leaf such as that of tobacco, and last in the main photosynthetic tissue, the palisade layer, just beneath the upper epidermis.

The vascular pattern in a tobacco leaf is determined early in the development of the vessel primordium. A procambial strand is formed by the elongation of narrow axial cells, and this extends both toward the base and toward the apex, eventually linking with the procambium of the stem. When the marginal meristems become active, the lateral veins of the leaf are initiated first, followed by the third and later order branchings that give the characteristic network of veins in the mature leaf.

Although the differentiation of the cells of the vascular system begins at the base, the epidermal and mesophyll cells mature from the tip inward toward the stem. The palisade cells elongate in a plane at right angles to the epidermis; those of the lower mesophyll expand irregularly to give lobed forms. The cells of the epidermis, shaped like irregular paving stones, continue to expand in the plane of the leaf after growth ceases in the mesophyll, so that the cells of the internal tissues are pulled apart to form the system of air spaces found in the mature leaf.

Dicotyledonous leaves are folded in various ways in the bud, the patterns being determined by differential growth in the tissues of the upper and lower surfaces (laminae) of the young leaves. Differential growth may cause the lamina either to roll or fold toward the leaf midrib or to fold near lateral veins, thus pleating the lamina. The folds in the bud are, of course, eliminated during the final phase of leaf expansion.

In the development of the maize leaf, the primordium arises first as a prominence some distance below the apical dome. The zone of division and growth extends laterally around the apex so that a complete collar forms; then the margins overlap. Meanwhile the original tip zone continues to elongate, eventually surpassing the stem apex. Tip growth declines thereafter, and further increase in cell number results from meristematic activity at the base. The early development of the vascular system is unlike that in dicotyledons, for several parallel procambial strands, rather than a single midrib, are initiated. The first of these grow toward the apex, but, as tip growth ceases, procambial strands form above and extend toward the base, passing through the node, or point of insertion of the leaf primordium, and into the stem below. As the leaf extends in length, the tissues begin to mature first at the tip, and a wave of differentiation passes down toward the base, where cell division and extension growth may continue long after the tip of the leaf is mature. Protection for this immature and succulent tissue of the leaf base is afforded by the sheaths of older leaves surrounding it.

Variations in leaves

These examples illustrate the principles involved in leaf development, but there are many deviations associated with variations in leaf form. Lobing and toothing result from the persistence of cell division and growth in particular stretches of the margin after growth ceases in between. Carried to the extreme, this localized growth gives the feather-like pinnate leaf. Many monocotyledons form cylindrical leaves as a result of a fusion of the margins of the primordium after it has encircled the stem.

Branching of the shoot. The shoots of most vascular plants branch according to a consistent plan, with each new axis arising in the angle between a leaf and a stem—that is, in a leaf axil. In some plants, buds may also form from the older parts of shoot or root remote from the main apices; these buds, termed adventitious, do not conform to the general plan.

A lateral shoot apex is initiated on the flanks of the main apex but at some distance below the point of emergence of the youngest leaf primordium. As in the origin of a leaf, generally the outer cell layers contribute to the surface tissues of the new apex by maintaining a consistent pattern of divisions. In some species a tunica of more than one cell layer quickly forms, so that the new apex appears as a miniature version of the main one; alternatively, the differentiation may not become apparent until the new primordium has attained considerable bulk. In all cases, the new apex must reach a minimal volume before it in turn can begin to form its own lateral primordia and to organize true axillary buds. As this volume is attained, *méristème d'attente–anneau initial* zonation appears. As in the main apex, the formation of new primordia is associated with the annular zone.

From this point on, the development of the lateral shoot is the same as that of the main shoot, except that growth may not be as rapid because the main apex, or leading bud, dominates and absorbs much of the available nutrient. The early growth of the axillary bud proceeds quite vigorously until a certain number of leaf primordia has been formed; then apical activity slows. Cell division gradually stops, and with it the associated syntheses; thus there is no increase in the DNA of the nuclei of the meristem after the last division. The bud, in effect, passes into a state of dormancy, even though the external conditions for growth are propitious. This phenomenon is known as correlative bud inhibition, since it is determined by the activity of the leading bud of the shoot. If the leading bud is removed, the inhibited lateral buds resume growth, and with it the associated syntheses.

Vascular development. Cell division planes in the zone just below the apex of the shoot tend to be oriented so that vertical files of cells are formed. This is more evident in the central core than in the surrounding cortical region, for the pattern is not disturbed by the insertion of

lateral members. The first signs of the differentiation of the vascular system appear some distance below the apex, in a zone of tissue distinguishable by the smaller cross-sectional area of individual cells. These cells, forming the procambium zone, arise by divisions oriented at right angles to the axis and may form a complete cylinder; generally, however, interruptions occur, the segments being related to the uppermost leaf primordia. In a dicotyledon such as tobacco, the cylinder at its highest level consists of strands running upward toward the points of insertion of the primordia. Thus, as the site of each primordium is determined, a strand forming in the adjacent region of the stem will contribute to the cylinder, but at a higher level than the preceding strand. The link with the earlier formed procambium is not simple, however. The strand passing upward toward a leaf primordium usually is composed of branches arising from strands that enter the two nearest older leaves below it. Because it in turn will contribute a branch for the next leaf, the cylinder is really a hollow network, the "gaps," or leaf traces, marking the points of departure of the leaf veins.

The "cylinder" of vessels

During subsequent development, the strands, or vascular bundles, increase in thickness by further cell divisions, and connections form with the vascular systems of axillary buds. The cells differentiate to give the characteristic tissues of the vascular system: phloem vessels (conducting tissue), phloem parenchyma (packing tissue), and phloem fibres (supporting tissue) toward the outside and xylem vessels (woody conducting tissue) toward the inside. The differentiation occurs in an upward direction, so that the maturation of the vascular tissues follows at a more or less constant distance behind the apex.

Although details differ, the above account of the origin of the primary vascular system is broadly applicable to gymnosperms and many ferns. Vascular development differs somewhat in certain flowering plants. In many monocotyledons, such as maize, the several vascular strands that pass down from each leaf primordium into the stem do not contribute to a single cylinder but are scattered in the ground tissue, or parenchyma, of the stem. Lateral interconnections form principally at the nodes.

Increase in stem diameter is accomplished in the older stems of dicotyledons by the activity of the cambium, which produces secondary vascular tissue. This meristem, a relic of the procambium, is composed of thin-walled cells, is flattened in the radial plane, and persists between primary vascular tissue, the differentiated outer phloem, and the inner xylem. When secondary thickening begins, the parenchymatous cells between the vascular bundles also resume division, ultimately forming a cambium cylinder. The cells of the cambium divide, producing initial phloem cells toward the outside and initial xylem cells within. Files of cells are also cut off among the initial phloem and xylem cells that remain parenchymatous and are called phloem and xylem rays.

As in the apical meristem, the number of dividing cells in the cambium remains constant, except that more cells occasionally are added by divisions in the radial plane so that the girth of the cambial cylinder expands in pace with the growth of the xylem within. The addition of new phloem toward the outside compresses the primary phloem and the cortical tissues in a radial direction while stretching them tangentially. These primary tissues do not persist, however. As the girth of the stem increases, the epidermis is disrupted, and the outer layers of the cortex become meristematic, giving rise to the cork cambium, which generates cork cells on the outside. The cork layer, or bark, then takes over the protective function of the epidermis.

THE ROOT SYSTEM AND ITS DERIVATIVES

The root tip. Plants that have a single apical cell in the shoot also have a single apical cell in the root. The cell is again tetrahedral, but sometimes daughter cells are cut off from all four faces, with the face directed away from the axis producing the cells of the root cap. The cells derived from the other faces continue to divide mostly by forming transverse walls, but occasionally also in the longitudinal plane. In this way vertical columns of cells form—tending, because of their mode of origin, to be disposed in three sectors.

In the roots of gymnosperms, angiosperms, and some lower plants, there is no single apical cell. Again, as with the shoot, such root apices can be analyzed in different ways. Perhaps the most useful approach is based upon tracing the sources of the main tissues in the apical region. Such an analysis has led to the histogen theory, which proposes that the three principal tissues of the root —vascular cylinder, cortex, and epidermis—originate from three groups of initial cells, or histogens, in the apical meristem—plerome, periblem, and dermatogen respectively. A fourth histogen, the calyptrogen, produces the root cap. The histogens have been thought to lie in linear order in the apex, with the initial cells of the vascular system toward the older part of the root, and those of the cap toward the tip.

The histogen theory of root-tissue origin

The histogen theory is difficult to apply to some types of roots, and there has been uncertainty about the numbers of histogens. The discovery of the "quiescent centre" in the root apex has clarified many features, however. The quiescent centre is a group of cells, up to 1,000 in number, in the form of a hemisphere, with the flat face toward the root tip; it lies at the centre of the meristem, in much the same position, in fact, as the tetrahedral apical cell in certain lower plants (Figure 6). The

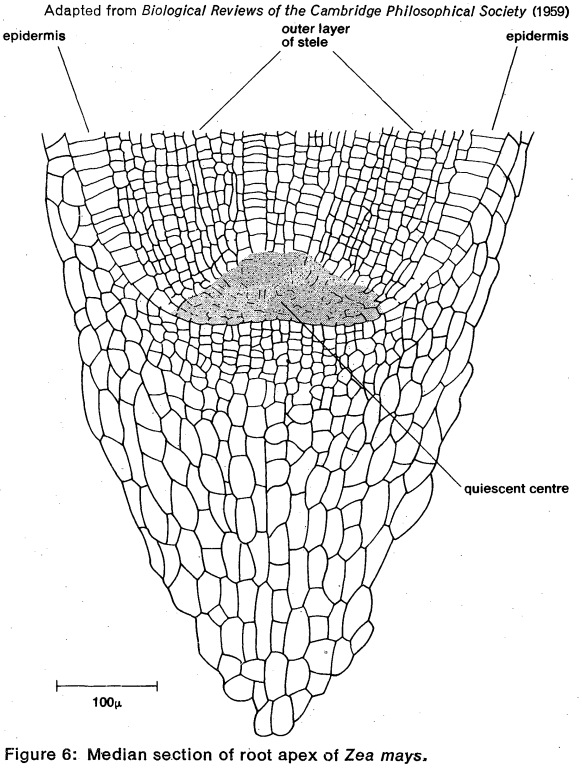


Figure 6: Median section of root apex of *Zea mays*.

cells of the quiescent centre are unusual in that their division rate is lower than that in the surrounding meristem. The cells of the centre have other distinctive features as well, notably a lower rate of protein synthesis than that of neighbouring cells.

The quiescent centre is surrounded by actively dividing cells of the promeristem that are the initial cells of the various tissues of the root. Those abutting the flat, tip-directed face contribute to the root cap; those above the quiescent centre are distributed in a cup shape. The cells in the centre of the cup produce the procambium and so, ultimately, give rise to the vascular cylinder. The annular zone of cells surrounding this central group forms the initials of the cortex; surrounding this, in turn, a ring of initial cells forms the protoderm, the layer corresponding to the epidermis at this level of the root.

The quiescent centre, a constant feature of the root tip, is apparently generally present in angiosperms and prob-

ably also in gymnosperms. The quiescent centre probably plays a role comparable with that of the apical cell in some lower plant roots, maintaining the geometry of the system. It has also been suggested that it may be concerned with the synthesis of growth hormones, although no direct evidence exists. When roots are damaged mechanically or by radiation, the cells of the centre can resume a rapid division rate, and they then participate in regeneration.

The zone of cell division extends some distance along the length of the root above the tip region. Although the girth may increase by longitudinal divisions and the widening of the daughter cells, most divisions occur in the transverse plane resulting in the formation of longitudinal files of cells.

In longitudinal section, the tissue zones become progressively better defined away from the tip. An internal protective band, the endodermis, becomes conspicuous as a single sheath of cells, surrounding the procambium. The phloem procambium, recognizable by its narrow cells, begins to differentiate in the lower part of the region of elongation. The xylem also becomes distinct, the thickenings appearing first in the upper part of the extension zone. Differentiation keeps pace with the advance of the root tip as new cells are added in the promeristem. When xylem occupies the core, there is no pith as in the shoot, but the cells of the outermost layers of the vascular cylinder remain undifferentiated, forming the pericycle, a tissue important in the formation of lateral roots. Within the bounds of the pericycle, the xylem is star-shaped in section, with the first-formed xylem elements (protoxylem) occupying the ridges. The phloem lies in the intervening grooves. Outside of the endodermis, the cortical cells elongate but remain thin-walled. Above the level of the root-cap sheath, the epidermis forms the outer layer of the root, and, beyond the extension zone, its cells begin to develop root hairs.

A more complete account can be given for the mechanics of development of the root apex than for that of the stem, mainly because of its greater simplicity. An important difference lies in the absence of a mechanism for the cyclical production of lateral organs at the apex itself.

Generation of lateral roots

Branching of the root. The branching of the root takes place in the older parts and does not directly involve the apical meristem. The tissues concerned are the endodermis and the layer immediately beneath it, the pericycle. The endodermis participates in root branching in certain lower plants with apical cells. A cell of this layer enlarges and forms a tetrahedral cell, which becomes the new apical cell; by further divisions a hemispherical volume of tissue forms around it—the whole constituting a new apex.

In many other plants, including gymnosperms and angiosperms, the lateral roots develop from the pericycle. Cells in this layer enlarge and begin to divide until a dome of tissue develops. Called the incipient apex, the dome pushes out the surrounding endodermis, which may itself resume divisions, its daughter cells enlarging to create a sheath around the new root tip. During further growth, the dome assumes an organization like that of the primary root apex. At first, all cells are meristematic; then, while the primordium is still small, cells in the central zone cease DNA synthesis, and this zone becomes the new quiescent centre. Beyond it, the root cap is produced, and, at the base, initial cells begin to develop the cell files that become the vascular cylinder, cortex, and epidermis. The vascular tissues differentiate from the base outward, and link eventually with xylem and phloem of the parent root. All this development occurs before the tip of the new root emerges from the tissues of the parent root. The growth of the new tip into the cortex first pushes out the endodermal sheath, if one is present, and then bursts it. The cortical cells are themselves crushed and probably resorbed as the root grows on, until finally the tip breaks through the epidermis.

In most roots, new laterals are initiated in the pericycle opposite to the protoxylem ridges. They tend accordingly to form vertical ranks along the length of the root, reflecting the number of bands of protoxylem. Although lateral roots arise in quite a different way from leaves and axillary shoots at the stem apex, there are certain common features. Pericyclic cells about to produce a root primordium synthesize ribonucleic acid, in anticipation of the period of growth and morphogenesis that will result in a new apex. The same behavior is seen in the cells of the annular zone, from which leaf primordia arise at the stem apex, and also in the axillary zones at a slightly lower level, from which new stem apices develop.

Later growth. In the secondary growth of the root, cell division in the primary xylem produces a cambium, which abuts the pericycle over the protoxylem ridges and passes between the phloem strands and the xylem in the grooves. Activity of the root cambium is comparable with that of the stem cambium; phloem elements are cut off outward, and xylem elements are cut off within. With continued growth in thickness, the star-shaped figure of the primary xylem is lost, and the cambium eventually forms a cylindrical sheath. Again, as in the stem, the protective function of the epidermis is ultimately taken over by cork layers produced by a cork cambium in the outer cortex.

IV. Correlations in plant development

COORDINATION OF SHOOT AND ROOT DEVELOPMENT

Although the structural organization of the vascular plant is comparatively loose, development of the various parts is well coordinated. Control is dependent upon the movement of chemical substances, including both nutrients and hormones.

Meeting demands for food and support

An example of correlation is the growth of shoot and root. The enlargement of aerial parts is accompanied by increased demands for water, minerals, and mechanical support that are met by coordinated growth of the root system. Several factors apparently are concerned with control, because shoot and root affect each other reciprocally. The root depends on the shoot for organic nutrients, just as the shoot depends on the root for water and inorganic nutrients and the flow of ordinary nutrients must, therefore, play some part. More specific control, however, may be provided by the supply of nutrients required in very small amounts. The root depends on the shoot for certain vitamins, and variation in the supply, reflecting the metabolic state of the aerial parts, may also influence root growth. In addition, hormonal factors affecting cell division pass upward from the root into the stem; although the exact role of the hormones has not yet been established with certainty, they may provide one way by which the root system can influence the activity of the shoot apex.

The control of secondary thickening is another important example of growth correlation. As the size of the shoot system increases, the need for both greater mechanical support and increased transport of water, minerals, and manufactured food is met by an increase in stem girth through the activity of the vascular cambium. Generally, the cambium of trees in temperate zones is most active in the spring, when buds open and shoots extend, creating a demand for nutrients. Cell division begins near the bud in each shoot and then spreads away from it. The terminal bud stimulates the cambium to divide rapidly through the action of two groups of hormones: auxins and gibberellins.

The inhibition of lateral buds, another example of correlated growth response, illustrates a reaction opposite to that occurring in the control of cambial activity. Lateral buds are inhibited in general because axillary shoots grow more slowly or not at all, while the terminal bud is active. This so-called apical dominance is responsible for the characteristic single trunk growth seen in many conifers and in herbaceous plants such as the hollyhock. Weaker dominance results in a bushy growth form with repeated branching. The fact that lateral, or axillary, buds become more active when the terminal bud is removed suggests that hormonal control is involved.

The flow of auxin from the shoot tip is, in part, responsible for inhibiting axillary buds. The nutritional status of the plant also plays a role, apical dominance being

strongest when mineral supply and light are inadequate. Because axillary buds are released from inhibition when treated with cell-division promoting substances (cytokinins), it has been suggested that these substances are also concerned in regulating axillary-bud activity.

DETERMINATION OF MATURE FORM

The duration of a juvenile phase

After its establishment as an independent plant, the sporophyte passes through a juvenile period before reaching maturity and becoming reproductive. Juvenility may be brief or, as in the case of trees, may extend over several years. The duration is determined partly by internal factors and partly by environmental controls related to the seasons.

In some ways juvenility is a continuation of developmental trends initiated in the embryo. In many plants, new organs are produced sequentially through early life, each of progressively more mature form. The first leaf of the young fern sporophyte, for example, is small and relatively simple, and the vascular system consists of a few forked strands. As growth proceeds, succeeding new leaves are of increasing complexity, and the shape begins to resemble that typical of the reproductive frond; in addition, vasculation shifts to the mature pattern, often one with a network of veins. Comparable trends occur in flowering plants, in which leaves at successive levels of plant maturity often show a progressive increase in the complexity of lobing or toothing.

Some of the changes associated with the juvenile period can be attributed to the gradual enlargement of the growing point, necessarily small in the embryo; its volume increases progressively with development. This increase in cell number is usually associated with the emergence of a "mature" zonation pattern. The typical internal structure of the shoot apex does not develop until a specific number of leaves form.

Gradual structural change in the growing point, however, does not adequately account for all aspects of juvenility. Sometimes, the transition from juvenile to adult leaf form is not graded but sudden. The juvenile leaves of species of the gymnosperm *Chamaecyparis*, for example, are needlelike and spreading; the adult leaves are scalelike and lie close to the stem. Among flowering plants, various species of *Eucalyptus* have juvenile leaves that are ovate and mature leaves that are sickle-shaped.

Phase change

Such sudden transitions from juvenile to adult form, referred to as phase change, seem to depend not on slow shifts in the apex but on some determinative event or correlated group of events. The two forms are relatively stable and tend to resist change; for example, cultured tissues taken from the juvenile (ivy-leaved) parts of ivy plants maintain a higher rate of cell division, and portions, or cuttings, taken from these parts tend to form roots more readily than those from the adult (simple-leaved) parts.

The establishment of these relatively stable but not wholly irreversible states is comparable with the determination of shoot and root poles during embryogenesis and, indeed, with the alternation of generations itself. The transmission of differentiated states through cell lineages presumably reflects the action of "switching" devices controlling the expression of different parts of the genetic complement. In this sense, phase change and related phenomena do not differ essentially from those of differentiation and organogenesis in general.

The transition in plants to the reproductive state is an example of a developmental event with some of the characteristics of phase change. Among seed plants, the reproductive structures are transformed shoots—strobili (including cones) of various kinds in the gymnosperms and flowers in angiosperms.

From a developmental point of view, the flower can be regarded as a shoot axis of determinate growth, with the lateral members occupying the sites of leaves differentiating as floral organs—sepals, petals, stamens, and pistils. In the transition to flowering, the stem apex undergoes distinctive changes, the most conspicuous of which is in the shape of the apical region, which is related to the kind of structure to be formed, whether a single flower, as in the tulip, or a cluster of flowers (an inflorescence), as in the lilac. The region of cell division extends over the entire apex, and the ribonucleic acid content of terminal cells increases. When a single flower forms, lateral primordia emerge at higher and higher levels on the flanks of the apical dome, and the entire apex is absorbed in the process, after which apical growth ceases. When an inflorescence forms, early changes are generally comparable to that for the single flower with one major difference—axillary primordia emerge that either become floral meristems or develop as secondary inflorescence branches. These primordia appear closer to the apex than do those of axillary buds on a vegetative shoot. In grasses, the activation of axillary meristems is the most notable early indication of the passage into flowering.

The rate of maturation and the timing of the transition to the reproductive phase are sometimes governed by internal controls and thus are relatively insensitive to the environment, provided conditions are generally favourable for growth. Frequently, however, the developmental rate is affected profoundly by recurring cycles in the environment, particularly those of temperature and of day length. In effect, these cycles provide a timetable for the plant, thus adjusting flowering, fruiting, and seed dispersal to the season and increasing the chances for successful propagation.

The control of the developmental rate by temperature is especially evident in many herbaceous plants of temperate climates. These plants, as indicated earlier, often must experience cold, either as seeds or as young plants, before they can begin flower production; otherwise they undergo an excessively long period of leafy, or vegetative, growth. After the cold experience, which can be given artificially, the plant is said to have been vernalized, or brought to the spring condition. Again the response is akin to a determination, because the condition attained is transmitted through subsequent cell divisions. Furthermore, there are indications that vernalization induces a persistent modification in the metabolism of apical cells and their derivatives. Ingenious theoretical schemes, offered to explain the apparent paradox that low temperature should actually accelerate a developmental process, are based mostly upon the proposition that a special vernalization hormone (vernalin) is involved. Although little direct evidence for the existence of vernalin exists, a class of hormones found in certain plant species, the gibberellins, does participate. The cold requirements of some species, such as the carrot, can be eliminated by the application of gibberellin, although the amounts needed are substantial.

Vernalization of the seed

The annual cycle of changing day length obviously provides the best of all "clocks" for the regulation of plant development. The effect of day length (or rather length of continuous darkness) on the transition to flowering is part of the general phenomenon of photoperiodism. Certain plants, called short-day plants, grow vegetatively when the nights are shorter than a critical minimum period (days long); exposure to longer nights (days short), however, accelerates development and brings on early flowering. Conversely, long-day plants develop very slowly toward flowering during daily cycles with longer than a minimum of darkness (days short), and are accelerated by exposure to short nights (days long). Other plants either require days of intermediate length for flowering or respond to a sequence of different photoperiods.

Photoperiod as a trigger for flowering

The leaf, rather than the stem apex, is the light-receiving organ in the photoperiodic reaction, although it is at the apex that subsequent developmental changes occur. One commonly accepted view is that, as a consequence of the photoperiodic experience, a specific flower-inducing hormone (as yet not isolated but referred to as "florigen") is synthesized in the leaf and translocated to the apex. As in the case of vernalization, photoperiod undoubtedly affects the metabolism of the known plant hormones, and so influences many other developmental responses apart from flowering. The effect of the duration of illumination on the carbohydrate balance of the plant may also be important. Nutritional effects on flowering

are well-known in many species—certain fruit trees, for example.

Whether or not environmental factors influence the passage into a reproductive state of a plant, the transition must be looked upon as part of the general developmental trend from juvenility to maturity: in this sense, flowering does not represent a radical alternative to vegetative growth but merely its culmination. Yet, entirely new organ types are produced at the flowering apex, presumably under the influence of genes that are not active during vegetative growth.

SEASONAL ADAPTATIONS

Certain plants are perennial and survive from year to year by matching their growth to the progression of the seasons or by suspending growth altogether during unfavourable times, such as winter or a dry season.

Response to an approaching unfavourable season

In the temperate zone, some time before winter begins, growth ceases in the shoots of woody plants, resting buds are formed, and deciduous trees lose their leaves. The resting bud consists of short axis, with the stem apex surrounded by modified unexpanded leaves, which protect the stem, especially from drying. The cells show marked frost resistance, similar to that of the embryo of the seed. Corresponding changes occur in herbaceous plants, in which the preparation for winter may involve the dying back of aerial parts altogether, leaving protected organs at or below the soil surface.

Growth sometimes ceases, even under favourable conditions, as a result of internal changes in the plant. This is true for some trees, which cease growth in midsummer. The passage into winter dormancy, however, is often controlled by the shortening of day length at the end of the growing season; in some plants decreasing night temperature also plays a part. Most temperate zone trees cease growth and form resting buds when the day length falls below a critical minimum.

Photoperiodic control seems to involve the formation of inhibitor compounds. In birches, for example, the leaf perceives the day length "signal" and transmits inhibitory materials to the apex, thus bringing growth to a stop and inducing the formation of a resting bud. The dormancy hormone, abscisic acid, may be concerned in this response and also in leaf abscission.

Budbreak in certain trees is controlled by photoperiod, growth resuming in the lengthening days of spring; light-perceptive organs are probably the young leaves inside the bud scales. Sometimes budbreak depends only on temperature increases that occur in spring, as in certain plants of Mediterranean climates.

The resumption of development in buds may result from a change in the balance of growth-inhibiting substances, such as abscisic acid, and growth promoters, notably the gibberellins. Buds can be caused to open prematurely by gibberellin treatment, which, as in the case of vernalization, can sometimes replace a cold experience; moreover, the gibberellin content in the buds of certain woody plants increases during chilling. Other hormones are probably also involved, however, for budbreak in plants such as the grapevine can be promoted by cytokinins, the plant cell-division factors.

An important general feature of adaptive periodicities is that the developmental changes anticipate the conditions for which they will ultimately provide the appropriate physiological or morphological adjustment. The ability of plants to utilize environmental indicators such as temperature and day-length changes is vital for the survival of plants. The production of such adaptive devices is made possible by the state of continuous embryogeny, already stressed as one of the most important characteristics of plant growth.

SENESCENCE AND DEATH

The growth of the vascular plant depends upon the activity of meristems, which are, in a sense, always embryonic. Continued indefinitely, this mode of growth could mean immortality; indeed, the longest lived individual organisms ever to have existed on earth have been certain species of trees. Plants and plant parts, however, do die, and death is often not the consequence of accident or environmental stress but of physiological decline—aging, or senescence.

Various kinds of physiological senescence and death occur and may affect particular cells, tissues, organs, or the whole plant. In the formation of the vessels of the xylem, cells conclude their differentiation by dying and contribute their empty walls to the conducting tissue. Individual organs such as leaves usually have a limited life-span. Entire shoot systems may gradually die back in the aerial parts of perennial plants, which overwinter underground. And, finally, the whole plant may die after a limited period of growth and the completion of reproduction. This behaviour is found in many annual plants, which complete their life cycle in a single growing season. The life-span may extend to two years, as in biennial plants, or longer, as in banana and certain bamboos, which die after flowering and fruiting.

Programmed death of plant cells

In the examples cited above, the death of cells, organs, or individual plants appears to be "programmed" and, in some sense, adaptive. This is clearly so with the death of individual cells during differentiation, when residual products contribute to the effective function of the entire plant body. The death of leaves and of shoot systems is part of the plant's adaptation to the cycle of the seasons. In annual species, the death of the whole individual may be viewed in a similar way. The succession of generations in this case is carried on by seeds; the sacrifice of the parent plant may, in fact, contribute to the success of the seedling by making available to the seed a pool of reserves derived from the breakdown of parent tissues.

Certain features characterize the onset of senescence. The cells show degenerative changes often associated with the accumulation of breakdown products. Metabolic changes accompany the degeneration. Respiration may increase for a period, but the rate ultimately declines as the cellular apparatus degenerates. Synthesis of proteins and nucleic acids ceases, and, in some instances, disintegration of cells has been associated with the release of enzymes through the disruption of membrane-bounded bodies called lysosomes.

The death of individual cells in tissues such as the xylem appears to be governed by internal factors, but senescence often depends upon interaction of tissues and organs. The presence of young developing leaves often accelerates the aging of older leaves; removal of the younger leaves retards the senescence of the older ones, suggesting control by competition for nutrients. A similar effect is seen in annual plants, in which the development of fruits and seeds is associated with the senescence and, ultimately, the death of the rest of the plant; the removal of reproductive structures slows the rate of aging. In these instances competition obviously has some effect, but it does not sufficiently explain why older, mature organs suffer in competition with those still in active development. The link may lie partly in the capacity of developing organs to draw nutrients to themselves, even from older parts of the plant. Developing organs thus provide "sinks" toward which nutrients tend to move. The senescence of organs drained in this way could result from the progressive loss of certain key constituents; should leaf protein, for example, turn over by breakdown of proteins to their amino-acid constituents and then be resynthesized, a steady drain of amino acids from the leaf would progressively deplete the proteins in the leaf.

Nutritional "sinks"

"Sinks" can be only part of the explanation, however, for in detached leaves of plants such as tobacco, protein synthesis decreases, and protein content falls, while the amino-acid content actually rises. Senescence in such instances can hardly depend on the withdrawal of nutrients. Furthermore, leaf senescence can be retarded locally by the application of cytokinins, hormones that stimulate plant cell division. Parallel effects have been demonstrated with growth substances of the auxin type in other plant systems. In the same way that active buds and fruits form sinks for nutrients from elsewhere in the plant, a cytokinin-treated area of a leaf attracts nutrients from other parts of the leaf. Although the metabolism of isolated leaves may differ in many respects from

that of attached leaves, leaf senescence probably does not result only from nutrient drainage but also from the synthetic activity of leaf tissues, which may be under hormonal control from other parts of the plant. The root may be important, for roots are known to export cytokinins to the shoot.

Daily length of darkness

Environmental factors, primarily photoperiod (daily length of darkness) and temperature, play important parts in governing senescence and death in plants. In annual plants, death is the natural conclusion of development; thus, conditions accelerating development automatically advance senescence. This is readily seen in short-day plants, in which precocious reproduction upon exposure to long dark periods is followed by early death. Senescence may be retarded in these cases, however, by hormonal treatments of the kind known to delay degeneration and death in detached leaves. Competition for nutrients between vegetative and reproductive structures cannot be the primary cause of death, for, in species such as hemp, the male plants—which do not produce seeds—die earlier than the females under short-day (long-night) conditions.

In perennial plants, leaf fall is associated with approaching winter dormancy. In many trees leaf senescence is brought about by declining day length and falling temperature toward the end of the growing season. Chlorophyll, the green pigment in plants, is lost; yellow and orange pigments called carotenoids become more conspicuous; and, in some species, anthocyanin pigments accumulate. These changes are responsible for the autumn colours of leaves. There are some indications that day length may control leaf senescence in deciduous trees through its effect on hormone metabolism, for both gibberellins and auxins have been shown to retard leaf fall and to preserve the greenness of leaves under the short-day conditions of autumn.

From the foregoing it may be seen that senescence and death are important in the general economy of plants. The paradox that death contributes to survival is resolved when it is understood that the death of the part contributes to the better adaptation of the whole—whether organ, individual, or species. Viewed in this way, death is no more than another—albeit the ultimate—manifestation of development.

BIBLIOGRAPHY. Developmental topics were taken up in many older botanical works, and special mention may be made of KARL GOEBEL, *Organographie der Pflanzen*, 2 pt. (1898–1901; Eng. trans., *Organography of Plants*, 1900–05). Several modern texts on plant anatomy and morphology deal descriptively with development, including A.S. FOSTER and E.M. GIFFORD, *Comparative Morphology of Vascular Plants* (1959); K. ESAU, *Anatomy of Seed Plants* (1960) and *Plant Anatomy*, 2nd ed. (1965); and ABRAHAM FAHN, *Plant Anatomy* (1967; orig. pub. in Hebrew, 1962). For cryptogams, general surveys of development appear in G.M. SMITH, *Cryptogamic Botany*, 2nd ed., vol. 1, *Algae and Fungi*, and vol 2, *Bryophytes and Pteridophytes* (1955).

Morphogenesis in plant development is treated in E.W. SINNOTT, *Plant Morphogenesis* (1960); and C.W. WARDLAW, *Morphogenesis in Plants* (1968). Many recent books are devoted to physiological and biochemical aspects. Introductory treatments are provided by A.C. LEOPOLD, *Plant Growth and Development* (1964); J.G. TORREY, *Development in Flowering Plants* (1967); and P.F. WAREING and I.D.J. PHILLIPS, *The Control of Growth and Differentiation in Plants* (1970). More advanced coverage of certain parts of the subject may be found in F.C. STEWARD, *Growth and Organization in Plants* (1968); W.M. LAETSCH and R.E. CLELAND (eds.), *Papers on Plant Growth and Development* (1967); and M.B. WILKINS (ed.), *The Physiology of Plant Growth and Development* (1969).

Many important topics in plant development are best approached through specialized monographs. The following are recommended: PANCHANAN MAHESHWARI, *An Introduction to the Embryology of Angiosperms* (1950); L.J. AUDUS, *Plant Growth Substances*, 2nd ed. (1959); F.A.L. CLOWES, *Apical Meristems* (1961); and W.S. HILLMAN, *The Physiology of Flowering* (1962). Sections are devoted to development and differentiation in each volume of the *Annual Review of Plant Physiology* (1950–). The articles are advanced and require background knowledge for proper appreciation.

(J.H.-H.)

Devonian Period

The Devonian Period is a time interval of the Paleozoic Era during which rocks of the Devonian System were formed (345,000,000 to 395,000,000 years ago). Rocks of Devonian age are known on all continents. The rocks show evidence of the earliest widespread continental and desert conditions, as in the Old Red Sandstone areas of Europe, for example, but marine deposits typically are more widespread. Sediments of this period contain the earliest abundant remains of vascular plants and of fishes of bizarre types, and four-footed vertebrates or tetrapods are contained in Late Devonian rocks. Insects are first known in the Devonian, and other invertebrates are abundant and varied. Devonian rocks crop out over wide areas and are locally of considerable economic importance in the production of building stone, cement rock, refractory and building brick, roof slate, glass sand, abrasives, and rock salt.

Sedimentary iron ore, particularly in Germany, formerly was of great importance. Lodes of tin, zinc, and copper occur in Devonian rocks. These deposits formerly were of considerable economic importance in Devon and Cornwall and were responsible for bringing trade in the Bronze Age. These deposits also later attracted the Phoenicians and Romans. Oil and gas have been produced for many years from Devonian rocks in New York and Pennsylvania.

Recently, more spectacular discoveries have been made. In 1944 oil was found in Middle and Upper Devonian sandstones in the Ural-Volga region of the U.S.S.R., and this area is the most important oil-producing area in the U.S.S.R. at present. In 1947 oil was discovered in an Upper Devonian reef at Leduc, Alberta, and this was followed by vigorous exploration. Ten years later Alberta accounted for over 95 percent of the Canadian output of crude petroleum.

Another spectacular discovery was that of the concealed Middle Devonian evaporites, anhydrite and halite, located in southern Saskatchewan. These were estimated to contain potash reserves that would last for eight centuries at current rates of consumption. Extraction there is of growing importance.

This article treats the rocks, life, and environment of the Devonian Period. For information on conditions before and after Devonian time, see SILURIAN PERIOD and CARBONIFEROUS PERIOD, LOWER. See also PALEOZOIC ERA, UPPER for an overview and STRATIGRAPHIC BOUNDARIES and FOSSIL RECORD for general, informative discussions of topics that are relevant to Devonian stratigraphy and paleontology.

DEVONIAN STANDARD SECTION

The Devonian Period was first named in 1839 by Adam Sedgwick and Sir Roderick Impey Murchison following studies in Devon and Cornwall in southwestern England. In Wales and in Scotland, deposits called the Old Red Sandstone composed of land and lake sediments had long been known between the Silurian System below and the Lower Carboniferous (Mississippian) System above. As a result of studies of fossil corals from Devon, William Lonsdale first suggested that the Devon and Cornish rocks might be the marine equivalents of the Old Red Sandstone to the north, and it was largely on this evidence that Sedgwick and Murchison proposed the Devonian System. Although recently much better understood, the rocks of Devon were for many years much more poorly known than the extremely fossiliferous and little disturbed Devonian rocks in the Ardennes in Belgium and the Eifel and Rheinisches Schiefergebirge areas in Germany.

In practice, these areas have become the world standard and have provided the names of the stages that form the subdivisions of the period. James Hall, in 1843, published a monumental account of the Devonian equivalents in New York State, which form the North American standard. The standard series of stages for the Devonian System is as follows:

Upper Devonian	Famennian (named from Famennes, Belgium) Frasnian (named from Frasnes, Belgium)
Middle Devonian	Givetian (named from Givet, N. France) Eifelian (named from Eifel, Germany)
Lower Devonian	Emsian (named from Ems, Germany) Siegenian (named from Siegen, Germany) Gedinnian (named from Gedinne, Belgium)

Among French-speaking geologists the stage name Couvinian (from Couvin in Belgium) is often used in place of Eifelian, although the lower boundary is within the upper Emsian of the German usage. The term "Coblenzian" was in use formerly and comprised the Siegenian and Emsian stages combined. As is usual, this type sequence is completely in marine rocks that consist mostly of limestones, sandstones, and shales.

Stratigraphic boundaries within the Devonian System are defined and correlated using various fossil groups. Within the Devonian the ammonoids, conodonts, brachiopods, and corals are particularly useful, and in nonmarine deposits fish and spores are used. There has been great difficulty in correlating the Silurian-Devonian boundary and until recently substantial errors were made, related to the erroneous supposition that graptolites became extinct at the boundary; they are now known to range well up into the Devonian. The graptolite zone of *Monograptus uniformis* is by international agreement taken as the base. The top of the system, by international agreement, is taken at the base of the ammonoid cephalopod *Gattendorfia* zone.

Absolute age of rocks

Radiometric evidence suggests this time interval began about 400,000,000 years ago and ended about 350,000,000 years ago. Convincing evidence for assigning precise dates has not yet been established, but two recent estimates give the beginning as 413,000,000 ± 5,000,000 years and the close at 362,000,000 ± 6,000,000 years. In practice, the evolution of fossil groups is used to subdivide and correlate rocks of the Devonian System; although this scale gives relative time only, precision in correlation is to an accuracy of about 1,000,000 years in some parts of the period.

DEVONIAN ROCKS

It generally is considered that Europe and North America were much closer together during the Devonian Period and probably were united approximately along the present continental slope margins (see CONTINENTAL DRIFT). At the close of the Silurian, and continuing in the Lower Devonian, considerable igneous activity occurred in the belt including New England, Nova Scotia, Newfoundland, Scotland, Scandinavia, and eastern Greenland. With North America and Europe united as described, the belt thus indicated formed a mountain tract of active uplift. This is the Caledonian mountain belt resulting from the Caledonian orogeny. The deposits of the Old Red Sandstones appear to be the detritus resulting from erosion of these mountain areas. The marine Devonian rocks of western Canada and those in a belt from Montana to New York in North America, in Europe from Devon to the Holy Cross Mountains of Poland, on the Russian Platform and Novaya Zemlya, and, again, in the Arctic islands of Canada appear to provide evidence that marine waters encircled the Old Red Sandstone continent.

To the south, the supposed union of the southern continents in the Devonian formed the area called Gondwanaland; this provided another land area, comprising parts of present continental South America, Africa, the Antarctic, India, and Australia. The evidence regarding the distribution of terrestrial rocks here is far less clear than in the Northern Hemisphere and, apart from the Early Devonian, the marine record is also poor. Other significant land areas were in central Siberia and northern China.

The accompanying world map (Figure 1) shows the distribution of most of the major outcrops of Devonian rocks. In many areas the Devonian rocks have been much disturbed tectonically by subsequent deformation. These fold belts may be distinguished from cratonic areas where sediments remain much as they were when formed. The main fold belts in North America are the western Cordillera and Rocky Mountains and eastern Appalachian belts. In contrast, the Devonian of the Midwest and adjoining areas is flat lying. In South America, the main fold belt is the Andes and sub-Andes, and east of this line the Devonian rocks are little disturbed. In Australia the main fold belt is in the east from Queensland to Tasmania. In Europe the Armorican fold belt stretches eastward from Cornwall and Brittany. To the south of this line from the Pyrenees to Malaysia, Devonian rocks are caught up in the Alpine-Himalayan fold belt. Similarly, the Devonian of the Urals of the U.S.S.R. is disturbed, whereas to the west, on the Russian Platform, and to the east there is less deformation. In all these cases the folding occurred well after the Devonian, but there

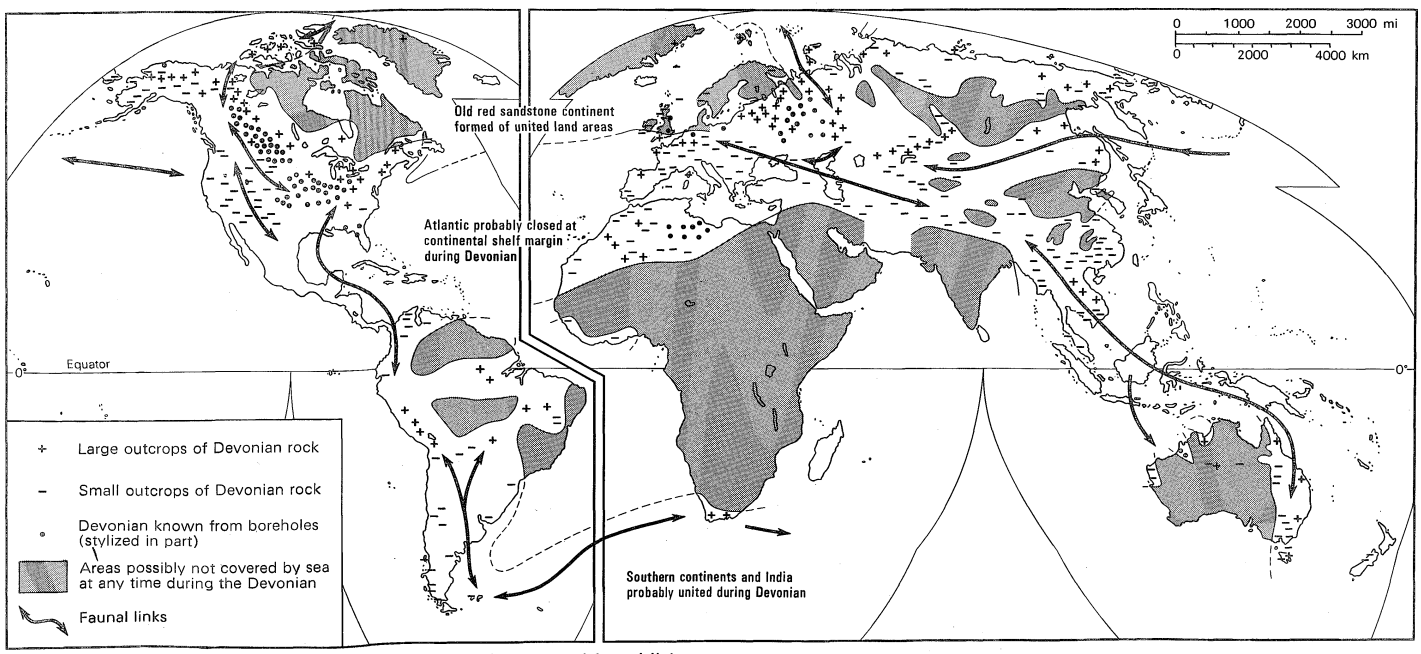


Figure 1: Devonian outcrops, inferred land areas, and faunal links.

is evidence that Devonian sedimentation contributed to the geosynclinal belts that were sites of the later mountain building.

In the regions that have suffered severe deformation, the Devonian sediments are frequently metamorphosed into slates and schists and often lose all the characters by which they may be dated. In areas where little change has taken place, all rock lithologies occur, from those characteristic of continental and desert conditions to the varied lithologies associated with shelf and deep sea accumulation. Contemporary igneous activity is widespread, both in the form of extrusive lavas, submarine pillow lavas, tuffs, agglomerates, and bentonites and also igneous intrusion. Extrusive activity is found in both continental and marine environments, whereas plutonic intrusions are usually linked with areas of uplift such as the Caledonian and Acadian belts of Europe and eastern North America.

Particular types of sediment characterize various areas of continental and marine environment and these give distinctive rock types. Considering continental areas first, the fluviatile sediments are those deposited usually by water, either in flash floods in desert areas or by rivers and streams. To these may be added the aeolian sand dunes and breccias resulting from arid erosion. All are represented in rocks of the Old Red Sandstone type. Lacustrine sedimentation, in freshwater to supersaline lakes is another type that is well represented by some Scottish sediments.

The Devonian marine lithologies are not dissimilar to those of other periods, but certain features are worthy of comment. Thick, fine-grained, dark-gray, or black shales and slates have characterized sedimentation to a significant extent. The Late Devonian Antrim, New Albany, and Chattanooga Shales in the Midwest of North America are an example; the German Wissenbacherschiefer is another. By contrast, very fine limestone reefs are known, and those of the Jasper National Park in Alberta and in western Australia must be among the best documented of any era.

For convenience in description this account will commence with a brief review of the European and North African sequences and then pass eastward to the Soviet Union, China, and Malaysia. Treatment of the southern continents from New Zealand to South America will follow, and North America will be considered last. Occurrences of various units are presented in Tables 1 and 2.

Europe

A line passing from the Bristol Channel eastward to northern Belgium and Germany roughly demarcates the Devonian marine area south of the Old Red Sandstone continental deposits, which characteristically are red stained with iron oxide. The continental deposits extend also to Greenland, Spitsbergen, Bear Island, and Norway. Robert Jameson coined the term "Old Red Sandstone" in 1808, mistakingly thinking it to be Werner's "Aelter Rother Sandstein," now known to be of Permian age. The rocks of this wide area have a remarkable affinity both in fauna and rock type and usually are considered to have united in Devonian times. The relations with the underlying Silurian System are seen in the classic Welsh Borderlands, where the Ludlow Bone Bed was taken as the boundary until international agreement placed it somewhat higher. In Wales, southern Ireland, and in the Scottish Lowlands, thicknesses of detritus, chiefly sandstones, accumulated to as much as 20,000 feet in places; and widespread volcanics occur in Scotland. These sediments are rich in fish and plants, as are the east Greenland and Norwegian deposits.

Devonian rocks in Devon and Cornwall are mostly marine, but there are intercalations of terrestrial deposits from the north. In north Devon at least 12,000 feet of shales, thin limestones, sandstones, and conglomerates occur, the latter two lithologies typical of the Hangman Grits and Pickwell Down Sandstones, which are the main terrestrial intercalations. But in south Devon reef limestones are in the Middle Devonian, and the Late Devonian locally shows very thin sequences formed on submarine rises and contemporary pillow lavas in basinal areas. In north Cornwall both the Middle and Upper Devonian are primarily in slate facies. Fossils found in these rocks have permitted detailed correlations with the Belgian and German sequences.

Devonian of mixed terrestrial and marine type is known from boreholes under London, and these form a link with the Pas de Calais outcrops and to the classic areas of the Ardennes. There, between the Dinant Basin and Namur Basin to the north is evidence of a northward land mass, as in Devon. Both the Lower and Upper Devonian consist of nearshore and terrigenous sediments that reach thicknesses of 9,000 feet and 1,500 feet, respectively. But the Middle Devonian and early Upper Devonian (that is, the Couvinian, Givetian, and Frasnian stages whose type sections are here) are mainly limestones and shales and reach at least 5,000 feet in the south. Reefs are especially well developed in the Frasnian and occur as isolated masses, usually less than one-half mile in length, separated by shales. Equivalents to the north show red and green silts and shales of marginal continental marine type. Because the Belgian Devonian rocks are well exposed along a north-south line, their changes in thickness, lithology, and fauna have been well documented.

The Eifel forms a natural eastern extension of the Ardennes, and a somewhat similar succession is known. The Early Devonian is nonmarine and the Middle Devonian and Frasnian have a poor reef development, but the calcareous shales and limestones carry a rich and famous fauna. The uppermost Devonian is not preserved.

The Rhine Valley along with the Rheinisches Schiefergebirge to the east, has been, since the early days of geology, the subject of immense study by the numerous German universities that surround it. Again a northern sediment source generally is indicated, but a borehole near Münster, well to the north, has encountered Middle and early Upper Devonian marine limestones. To the south also, approaching the Hunsruck-Taunus mountains, there is evidence of a land mass. Between these areas a rich Devonian sequence is exposed in folded terrain. The maximum thickness is 30,000 feet. The Early Devonian consists of slates and sandstones. The slate has been much worked to clad houses and castles. A ledge of Emsian sandstone in the Rhine Gorge is the setting for the Lorelei legend. Limestones are common in the Givetian and are termed Massenkalk. Middle and Upper Devonian areas of thin sedimentation, as in Devon, are interpreted as deposits on submarine ridges. These are commonly nodular limestones rich in cephalopods that occur between thick shale sequences. Evidence of volcanic activity is common, and this has been invoked to explain the concentrations of sedimentary hematite iron ores in the Givetian and Frasnian. The Harz Mountains show a more calcareous Early Devonian section. Here copper, lead, and zinc are exploited from lodes in the famous Wissenbach Slate.

A calcareous Lower Devonian succession, the Bohemian facies, occurs in the Prague Basin of Czechoslovakia. A continuous marine succession formed from the Silurian into the Devonian, and the boundary is drawn at the top of the Pridoli Formation with *Scyphocrinites*. The overlying Lochkovian and Pragian formations include the Koneprusy Limestone with substantial reefs. The Upper Devonian is not preserved. In Moravia complete successions of calcareous and basinal volcanic sediments occur.

Devonian rocks of a type analogous to southern England and the Ardennes crop out in Brittany. Farther south in Europe outcrops occur in Spain and Portugal. The successions of the Pyrenees, Montagne Noire, and Carnic Alps all include deepwater limestones; and marine deposits are known in Yugoslavia, Macedonia, and Romania. Devonian rocks are widely distributed in North Africa and have been recorded at depth in boreholes in the Sahara during oil exploration. The southern Polish outcrops of the Holy Cross Mountains are especially famous and include a lower marine and continental series with a calcareous Middle Devonian and an Upper

Devonian of reefs and shales rich in ammonoids and trilobites.

U.S.S.R. and Asia

In Podolia, along the Dneister River, are fine marine sections going well up into the Lower Devonian and overlain by the Dneister Series of the Old Red Sandstone type. During the whole Devonian the Ural Mountains formed a geosynclinal trough linked northward to Novaya Zemlya and southward to the Crimea-Caucasian geosyncline that, with the southern European outcrops already mentioned, formed part of the original Tethyan sediments of the Alpine-Himalayan fold system of the present day. In European Russia, Old Red Sandstone conditions were general, but marine tongues stretched westward from the Urals to reach Moscow in the Middle Devonian and Leningrad in the early Upper Devonian. A remarkable series of boreholes revealed this great detail (Figure 2, top), and there is widespread evidence for salt lakes. Apart from the Leningrad outcrop and those along the Don south of Moscow, these are known from subsurface data only. Of economic importance here are the Timan-Pechora oil and gas field and the oil and potash of the Pripyat Marshes. Central Siberia was a land area, but Devonian rocks crop out around it, including rich Lower Devonian calcareous sequences in Salair and the full marine sequence in Altai. The Altai-Sayan area contains a wealth of Old Red Sandstone fish and plants. In northern and northeastern Siberia, in basins around the central platform, Devonian sequences are now well known as a result of recent work. These show a rich and varied fauna grading from marine to continental facies, with associated rich salt deposits. Devonian rocks occur on many of the Soviet Arctic islands.

The Great Khingan range recently has provided good evidence for marine Middle and Upper Devonian, the latter with ammonoids. Many scattered outcrops of Devonian are known in Mongolia and in many parts of China, but the lowest Devonian is poorly represented and often is in continental facies. The plant *Leptophloeum* has been widely traced. The western outcrops are thought to link southward with the Malaysian outcrops. These rocks and those of India, Afghanistan, and Iran form part of the primitive Tethyan or Alpine-Himalayan geosyncline. Devonian rocks with Middle Devonian corals occur in Korea; those in Japan contain *Leptophloeum*.

Southern Hemisphere

In New Zealand the Lower Devonian is known in the Reefton and Baton River areas. The brachiopods in the fauna include European elements and have few typical austral types.

Devonian rocks are known in eastern Australia in a belt from Queensland to Tasmania as part of the Tasman geosyncline. Fluviatile sediments are found to the west. Thicknesses of 20,000 feet are known. *Leptophloeum* is found in the Late Devonian portion. Devonian rocks occur in central Australia in the Amadeus Basin and along the western coast in the Carnarvon, Canning, and Bonaparte Gulf basins. Complex facies changes are known, and the Canning Basin reef complexes show every detail of forereef, reef, and backreef structures exposed by modern erosion.

In the Antarctic both marine and continental Devonian occur, the latter rich in fish remains of European genera. The marine Early Devonian shows some affinity with the Bokkeveld in South Africa, which, in turn, has strong links with South America. No Devonian is known in Africa between Bokkeveld and the sections in Ghana and northwest Africa.

Early Devonian marine rocks are well developed in South America, but the Late Devonian is poorly documented. In the western geosyncline along the Andes and sub-Andes, remnants of Devonian are preserved from southern Chile north to Peru, Ecuador, Venezuela, and Colombia. The Devonian rocks of Uruguay, Argentina, and Brazil are thought to represent marine transgression from the west. Both continental and marine faunas are known. The fauna of the Falkland Islands and the Paraná and Parnaíbá basins has many genera of brachiopods and trilobites that are common within the circum-Antarctic region but unknown in the Northern Hemisphere. In Venezuela and Colombia, however, faunas of Appalachian type dominate, although austral elements, such as *Australospirifer*, linger.

North America

Belts of sediments down the Cordilleran region of western Canada and the eastern U.S. were linked by shallow shelf seas across central North America. The more rapidly subsiding Appalachian area received substantial clastic detritus from mountain ranges farther to the east and northeast. Evidence of their influence in forming the Catskill Delta is well seen in New York state. Thick sandstones and conglomerates that contain fish and plants thin westward along the Hudson River and pass into fully marine sequences (Figure 3). The rich faunas of the Devonian rocks of this area were published in a series of works by James Hall in the 19th century. In the Ontario, Michigan, and Indiana areas, early thin calcareous sequences give way to deeper water New Albany Shale, known farther south as the Chattanooga Shale.

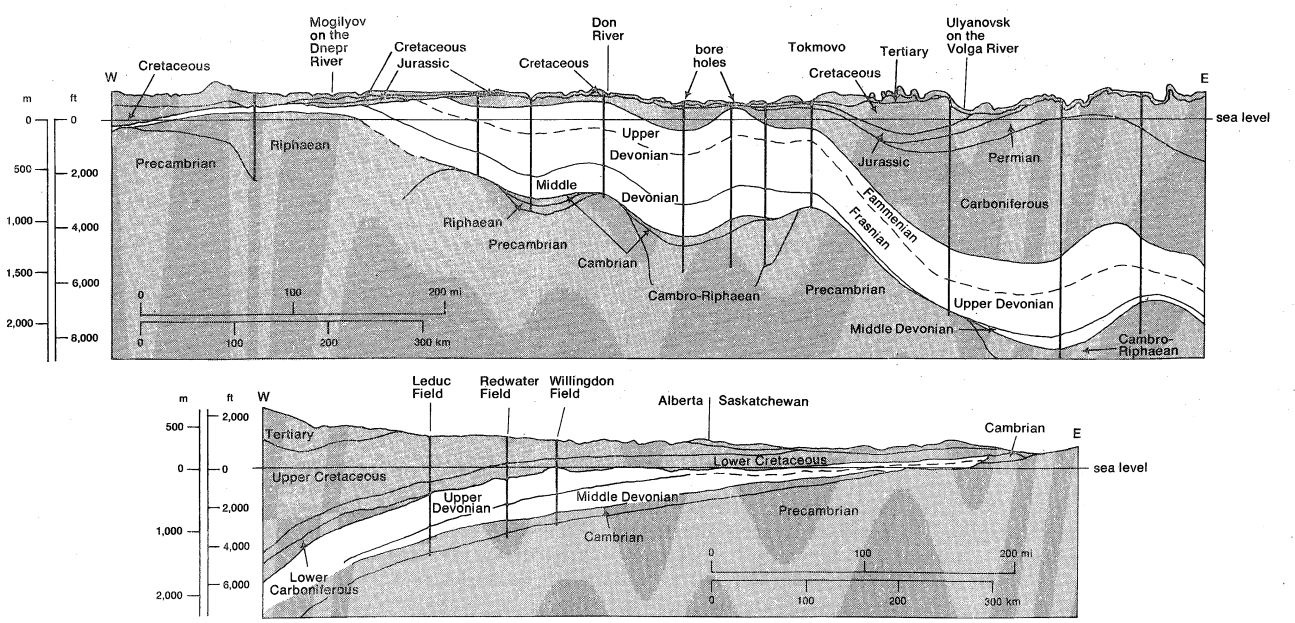


Figure 2: *Geological cross sections of Devonian rocks.*
(Top) Near the Polish frontier eastward to the foothills of the Ural Mountains. (Bottom) From the foothills of the Rocky Mountains eastward across the Interior plains of Canada.

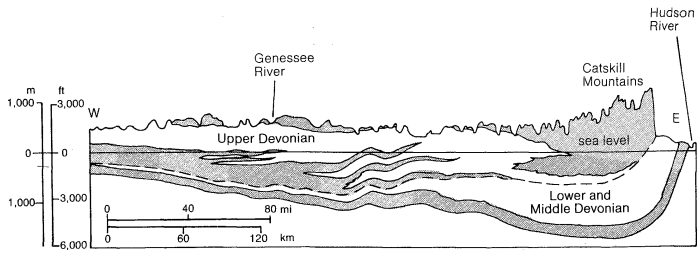


Figure 3: Cross section along the New York–Pennsylvania border showing changes in Devonian rocks. Light gray areas represent red and green shales, sandstones, and conglomerates deposited in fresh or brackish water. Darker gray area represents dark gray or black shales and limestones with marine fossils. White areas indicate gray shales and siltstones with shallow-water marine organisms.
From M. House, *Continental Drift and the Devonian System*

Farther west, in Wyoming, Utah, and the Grand Canyon, terrestrial fish-bearing rocks occur locally. The Canadian Devonian has been vigorously studied since the discovery of oil in Alberta (see Figure 2, bottom), and considerable exploration of Alaska and the Canadian Arctic has been undertaken. The faunas of eastern North America show many links with Europe.

DEVONIAN LIFE

A highly varied invertebrate fauna derived from that of the Silurian continued in the Devonian, and most ecological niches of shallow and deep marine water were exploited. The remarkable proliferation of primitive fish, which has given the period the name the "Age of Fishes," occurred both in fresh and marine waters. Derivation of carnivorous fish from mud-eating forms occurred early in the period, and the four-footed tetrapods were derived from the fish near the close of the period. Also remarkable is the rise to dominance of the vascular plants. By the mid-Devonian the first tree forests are known in place, but rich groves must have occurred earlier to provide the widespread plant debris.

Invertebrates

The Devonian invertebrate faunas, in contrast to the record of the plants and vertebrates, are essentially of the type established by the Ordovician. In nearshore sandy and silty environments bivalves, burrowing organisms, brachiopods, and simple corals abound. In offshore reefs, free from land detritus, biostromes and bioherms flourished, rich in corals, stromatoporoids, crinoids, brachiopods, trilobites, gastropods, and other forms. In deeper waters the cephalopod goniatites were abundant, one of the few new groups to appear; and there is evidence that the surface levels of these deep waters were occupied by small planktonic orthocones and by ostracods (arthropods) later in the period. Elements of these marine faunas are known on all continents of the world and have been much studied. They form the basis of the subdivision of the system and the most accurate means of correlating its members, and they enable paleoecological and paleogeographical reconstructions to be drawn with some assurance.

Both Foraminifera and Radiolaria among the Protozoa are well known, and sponges were locally abundant; the famous dictyospongids of New York are an example.

The corals and stromatoporoids among the coelenterate Hydrozoa were extremely important in the reef facies. Elsewhere, only simple corals are frequently found. The limestone reef and forereef facies and biostromal limestones are known in many areas of the world. The corals include tabulate corals, such as *Favosites* and *Alveolites*, but especially rugose corals, which have been used to establish correlations. *Amphipora* is a common rock-building type in the mid-Devonian of the Northern Hemisphere, and its twiglike form produces a "spaghetti" or "vermicelli" rock.

Bryozoa were especially common in shallow shelf seas of the period, and rich faunas are known from North America. Both stony (trepostomatous) and netted forms occurred; but the latter, the fenestellids, grow in importance during the period.

The brachiopods of the Devonian show great diversity. The spire-bearing spiriferoids were perhaps the most common and have been used for zonation. Two groups of importance began here: the loop-bearing terebratuloids and the spiny, mud-dwelling productoids. A number of groups become extinct in the period, including a number of orthids and the pentamerids.

Molluscan groups were well represented. The marine clams (Bivalvia) increased greatly during the period, especially in the nearshore environments. The earliest freshwater bivalves appeared in the Upper Devonian. The gastropods were well diversified, especially in calcareous environments, but much less than in later periods. The Scaphopoda first appeared here. A significant Devonian event was the origin of the ammonoids from their continuing nautiloid ancestors. In the chambered shell of the ammonoids, the siphuncle is ventral or outermost in position (except in Late Devonian clymenids), and the septa commence the elaborate folded patterns that culminate in the ammonites of the Mesozoic. From their appearance, probably in the Siegenian, the evolution of the goniatites and later ammonites allows a detailed zonal subdivision to be established through the end of the Cretaceous. Devonian goniatites have been found on all continents except South America.

Among the Arthropoda the giant Eurypterida are found in the Old Red Sandstone facies. Some were predacious carnivores and probably lived on fish. The first insect, a supposed collembolan, has been recorded from the Devonian of the U.S.S.R. Ostracods were locally very abundant; benthonic forms occur in shelf-sea deposits and planktonic forms in the Late Devonian, where their remains form the widespread ostracod-slate facies or cypridinenschiefer. The trilobites were well developed, in terms of size (some up to two feet long), variety, and distribution. Nearly all have clearly established Silurian ancestors. Commonest were the phacopids, which show a curious trend to blindness in the Late Devonian. Almost all of the diverse Lower Paleozoic trilobite stocks that entered the period were extinct before the close, and only the proetaceans survived into the Mississippian.

Among the Echinodermata, holothureans, asteroids, and ophiuroids are known, but they are rare. Crinoids were abundant, including free-living types with grapnel-shaped anchors. The blastoids diversified considerably, but the cystoids did not survive the period.

Conodonts, small toothlike structures usually less than one millimetre in size, had perhaps their greatest diversification during the Late Devonian and have proved of major importance for correlation.

Vertebrates

Many groups of Devonian fish were heavily armoured, and this has led to their good representation in the fossil record. Fish remains are widespread in the Old Red Sandstone rocks of Europe, especially in the Welsh Borderland and Scottish areas; these are mostly associated with freshwater or estuarine deposits. In other areas marine fish are known and some of these, such as *Dunkleosteus* (*Dinichthys*) from the Late Devonian of Ohio may have reached 30 feet in length.

The earliest groups, comprising the Agnatha, were without jaws and presumably were mud eaters and scavengers. These are usually termed the ostracoderms, and they are thought to be the ancestors of the living lampreys and hagfish. Some, such as the osteostracan cephalaspids, had broad, platelike armour of varied form; and the brain and nerve structures in some of these are well known. The anaspids also were covered with armour in the form of scales. The heterostracans, which include the oldest known fish, have an anterior armour basically of upper (dorsal) and lower (ventral) plates; *Pteraspis* is an example. The Early Devonian saw the entry of jawed forms or gnathostomes, and the armoured forms of these, the Placodermi, characterize the period. The arthrodires with a hinged frontal armour in two portions and the grotesque antiarchs belong here. The close of the Devonian saw the diminution and extinction of most of these groups, but several other groups continued and have significant later history. Sharklike fish, the Chondrichthyes, have been found in the Middle Devonian. The bony fish,

Table 1: Devonian Period Stages of the World

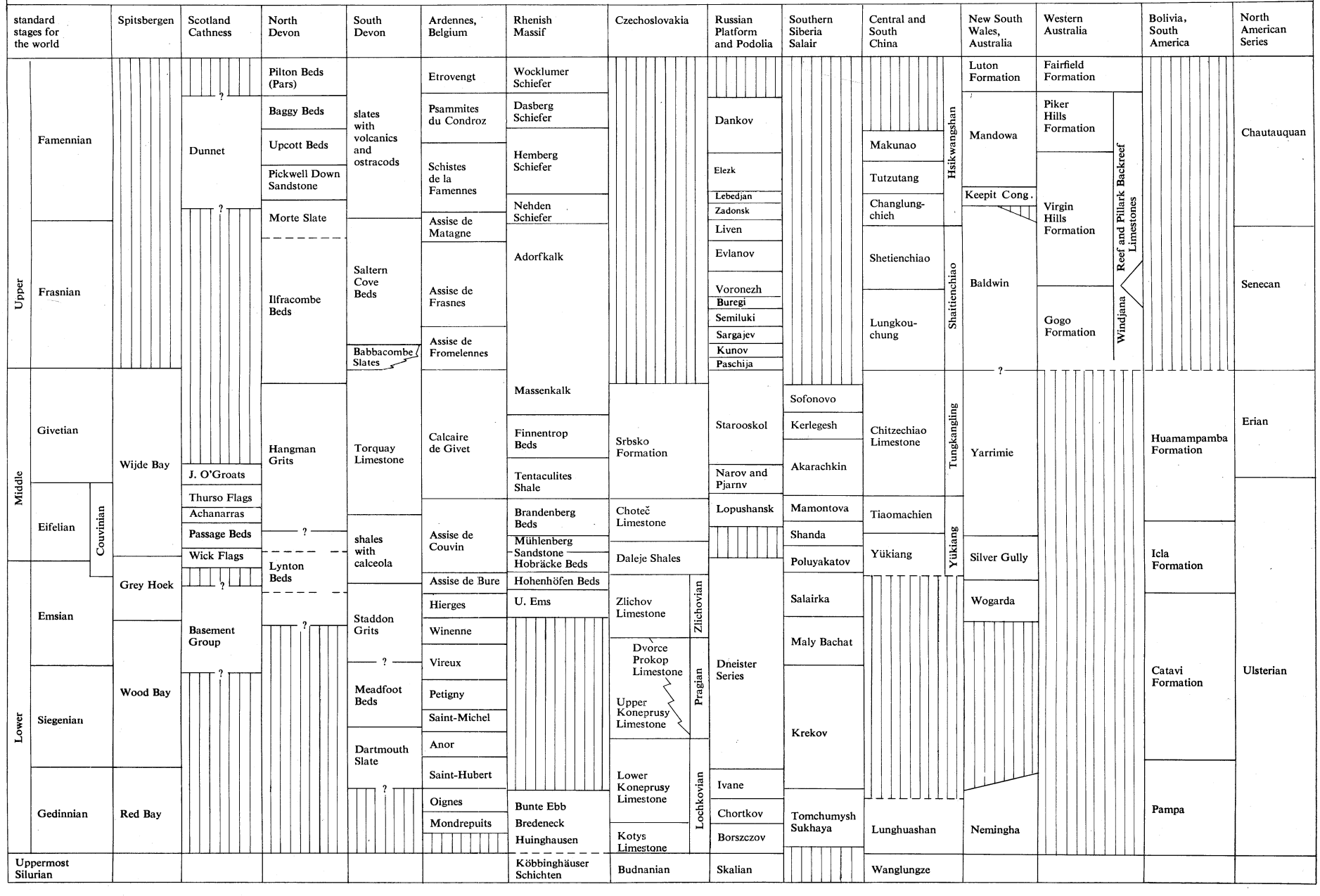
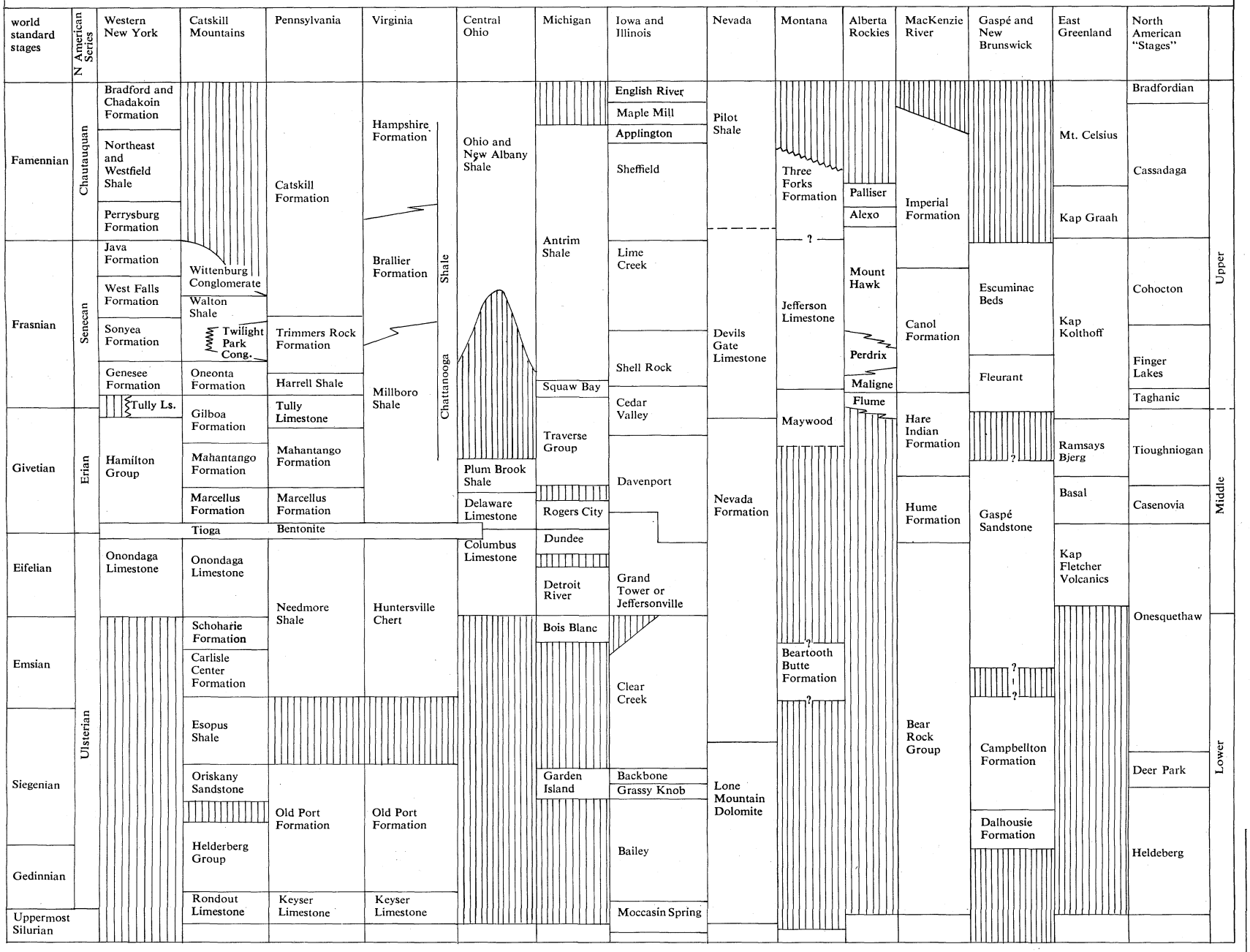


Table 2: Devonian Period Stages of North America



or Osteichthyes of current classification, include the climatioid acanthodeans, which had appeared before the period began; but the lungfish (Dipnoi), the coelacanths, and the rhiphidistid fish made their first appearance. The last group is thought to have given rise to the four-footed amphibians and all other higher groups of vertebrates.

These fossil fish have proved to be of great value to the geologist in correlating areas of Old Red Sandstone sedimentation and, more recently, in correlating such areas with marine deposits into which the fish often were washed.

Plants

In the history of vascular plants Devonian evidence is of fundamental importance because there was a remarkable initiation of vascular plants of diverse type. Their colonization formed the first forests such as the rich Gilboa forest of New York, of Middle-Late Devonian age. Much new information on spores is being provided by palynologists, and this may soon enable the antecedents of the Devonian flora to be established. Evidence of algae is common in the period, Bryophyta are first known here, and Charophyta are locally common. Freshwater algae and fungi are known in the Rhynie Chert of Scotland.

Some supposedly Silurian floras, such as that at Baragwanath, Victoria, Australia, are now known to be Early Devonian. The *Cooksonia* Late Silurian record of Czechoslovakia seems to be the earliest unquestionable evidence of vascular plants. By the Early Devonian a varied flora was established.

The Psilophytopsida is the most primitive group of the Pteridophyta; they did not survive the Late Devonian. *Cooksonia, Rhynia,* and others possessing a naked stem with terminal sporangia belong here. In other members, sporangia are borne laterally, but no true leaves were developed, and the branching was often of a primitive dichotomous type. The Psilophytopsida forms a basic stock from which other groups apparently evolved. *Asteroxylon,* known with *Rhynia* in the Lower Old Red Sandstone Rhynie Chert of Scotland, forms a link with the Lycopsida by having lateral sporangia and a dense leafy stem. This group soon gave rise to treelike forms and later to the important lepidodendrids of the Carboniferous flora. Another apparent derivative, the Sphenopsida, with jointed branches, is represented by *Hyenia* and *Pseudobornia.* This group became much more important after the Devonian and has descendants today in the scouring rushes (*Equisetum*). The Pteropsida also appeared in the Devonian. Primitive gymnosperms are known and drifted trunks of *Callixylon,* up to five feet in diameter, are common in Late Devonian deposits of the eastern United States and the Donetz Basin of the U.S.S.R.

The rich record of land plants may be related to the fact that the Old Red Sandstone represents the first widespread record of continental conditions. But the primitive nature of the stocks seen and the absence of a long earlier record, even of drifted fragments of vascular plants, suggest that the colonization and exploitation of this environment was a real Devonian event. Fortuitous finds, such as the silicified flora of the Rhynie Chert and the pyritized tissue from the Upper Devonian of New York, have enabled the intimate anatomy of many of these forms to be elucidated in a detail equivalent to that of modern forms.

Guide and index fossils

Most groups contribute to the establishment of the faunal and floral chronology that enables Devonian rocks to be correlated. For the continental deposits, fish and plant spores are most important. The fish give a very precise zonation in parts of the system. The Baltic Frasnian, for example, can be divided into at least five time zones using psammosteids (Agnatha), thus probably equalling the precision possible for the better-known marine Frasnian sequences. But many problems still remain in the correlation of the continental deposits with their marine equivalents.

The faunal succession in marine strata has been established for many groups, but only those of significance for international correlations will be mentioned here. Traditionally the goniatites and clymenids (ammonoid Cephalopoda) form the standard. The succession established first in Germany by Rudolf Wedekind has in 1917 been found to hold for all continents where representatives have been found. The index genera are shown in Table 3. All of these genus zones or *Stufen* are subdivisible into species zones. Since this group possessed float chambers, their widespread dispersal was probably related to drifting after death.

Table 3: Index Genera for Devonian Faunal Succession in Marine Strata

Famennian	*Wocklumeria* *Clymenia* *Platyclymenia*	clymenid genera
Frasnian	*Cheiloceras* *Manticoceras*	
Givetian	*Maenioceras*	goniatite genera
Eifelian	*Anarcestes*	
Siegenian and Emsian	*Mimosphinctes*	

Rivalling the ammonoids in parts of the Late Devonian and useful for defining the base of the system are the conodonts. The Late Devonian is characterized by a spectacular evolutionary radiation of *Palmatolepis* and its relatives. The widespread distribution of conodonts indicates that the unknown, soft-bodied organism to which they were attached was planktonic.

The brachiopods, although more restricted, are also important, especially the spiriferids of the Early Devonian, and the entry and evolution of the cyrtospiriferid types in the Late Devonian. The rhychonellids are of great value also in the subdivision of the Late Devonian. Some brachiopods, however, show diverse distribution patterns. *Stringocephalus,* a well-known Middle Devonian guide fossil in western U.S., Canada, Europe, and Asia, is entirely absent from the rich New York succession; yet *Tropidoleptus,* elsewhere confined to the Lower and Middle Devonian, ranges high in the Devonian of New York. Corals also have been used for correlation, but further work suggests they were particularly sensitive to changing local environments and thus are poor time indicators.

Faunal realms and migrations

There is a marked similarity in the faunas and floras of the continental facies the world over. Recent records from such deposits in China containing Early Devonian *Cephalaspis* and *Pterichthys* or the widespread Australian records of *Bothriolepis* link closely with the Old Red Sandstone faunas of Europe. Yet when studied in more detail, specific differences become apparent. It has been suggested that the Baltic fish succession is so rich that it must have formed a migration centre. This may be so, but the wide distribution of supposed estuarine and freshwater fish raises many problems. Many of these can be resolved if the continents were closer together during the Devonian than at present.

The marine faunas of the Devonian give little evidence of faunal provinces. It is true that in the Lower Devonian the brachiopod *Australocoelia* has been recognized only in the Antarctic, the Falkland Islands, South America, South Africa, and Tasmania and that *Australospirifer, Scaphiocoelia,* and *Pleurothyrella* share parts of this distribution. These are not known in the marine Lower Devonian of northern continents, and this seems to establish an "Austral" fauna of limited circum-Antarctic distribution at this time (if the southern continents were then united as Gondwanaland). Elements of this fauna are often called "Malvinokaffric" after the Falkland (Malvinas) Islands and the South African Bokkeveld Beds. But at other levels in the Devonian provincial distinctions are not apparent, with the exception of local coral provinces that are distinguishable in the U.S.S.R.

Evidence is less ambiguous with regard to migration routes (Figure 1). The circum-Antarctic Lower Devonian fauna is linked perhaps to the north by *Leptocoelia* common in southern and eastern North American Devonian. *Leptocoelia* has not been recognized in Europe or Asia apart from Kazakhstan and farther east in the U.S.S.R. The faunas of the western U.S. and western

Canada are firmly linked with the U.S.S.R. and Europe in the mid-Devonian and Late Devonian. There seems to be a circum-Arctic route around a united Old Red Sandstone continent linking the Cordilleran faunas with those of the Urals and Poland, Germany, and England. Evidence of this is forthcoming in the very rich records of *Stringocephalus*, *Manticoceras*, and *Amphipora* all along this line. Faunal links between the fine New York sequence and western Europe, however, are poor. The absence of many typical goniatite genera in New York, which are present in the western United States, and also the absence of *Stringocephalus* and strange ranges of *Paraspirifer*, *Tropidoleptus*, and *Cyrtospirifer* emphasize the dissimilarity. The Caledonian mountain chain, linked with the source of Appalachian sediments, must have provided an effective barrier even if North America and Europe, in predrift positions, were close together. It has been claimed that American faunal types occur in northwest Africa, and some European-type goniatites occur in the Middle Devonian of Virginia. But the general weakness of the transatlantic faunal ties is striking.

DEVONIAN ENVIRONMENTS

Paleogeography

The evidence from the Devonian supports the belief that the present-day continents were formerly united in some way. In the Northern Hemisphere the linking of North America, Greenland, and Europe permits the postulation of a single Old Red Sandstone continent. In the Southern Hemisphere the closer union of Australia, Antarctica, South Africa, and South America would bring together the Austral elements previously mentioned, and for later periods this union is imperative to explain the distribution of glacial deposits.

There is no evidence that the orogenic phases of the Devonian were worldwide and synchronous. Mountain-building orogenies were often widespread in time. Thus, although the Caledonian orogeny was important in the north Atlantic area for initiating Old Red Sandstone conditions, both in the eastern United States (Middle Devonian-Acadian orogeny), southwestern United States (Antler orogeny), and in Greenland and Scotland, significant uplift and erosion occurred within the period. The same is true in other areas. Both volcanism and intrusion are associated with those orogenic phases. When regarded on a world scale, it is clear that volcanism in the Devonian is associated with the belts described as geosynclinal and to areas immediately adjacent where antipathetic uplift was occurring. Broad areas of cratonic and shelf sedimentation rarely exhibit volcanic deposits.

Devonian climates

The wide distribution of evaporite basins in the Northern Hemisphere, of coals in Arctic Canada and Spitsbergen, of desert conditions, and of widespread marine faunas and carbonate reefs, suggests that warm and equable climates covered large areas. But from New York, where rich forests grew, *Callixylon* trunks with annual rings typical of the seasonal growth of higher latitudes are known. Studies of growth lines on Devonian corals indicate that the number of days in the Devonian year was in the order of 400 and that the lunar cycle was about 30½ days (see DATING, RELATIVE AND ABSOLUTE). The wide distribution of salt deposits suggests high evaporation and warm climates in many areas from western Canada to the Ukraine and Siberia and, again, locally in Australia. On the basis of paleomagnetic evidence from studies of rock magnetism, an Equator passed from California to Labrador and from Scotland to the Black Sea. By analogy with the glacial tillite distribution of the subsequent periods, the southern continents must have been near a pole. There is some evidence that the Austral province suffered glaciation in the Lower Devonian. The northern pole, paleomagnetic evidence suggests, may have lain in the north Pacific.

BIBLIOGRAPHY. The following collection of papers by international authorities is now the standard work for all aspects of the world Devonian: D.H. OSWALD (ed.), *International Symposium on the Devonian System*, 2 vol. (1967).

(M.R.H.)

Dew

Dew is a deposit of waterdrops formed at night by the condensation of water vapour from the air onto the surface of objects freely exposed to the sky. It forms on clear nights when the air is calm or, preferably, when the wind is light. If the temperature of the surface is below the freezing point of water, the deposit takes the shape of hoarfrost (see FROST). Dew forms on clear nights because on such nights freely exposed surfaces lose heat to the sky by radiation. Unless this loss is offset by an efficient conduction of heat from the interior of the object, the surface will cool. Most objects, including grass blades, leaves, and petals, are much better radiators than air and, as a result, are usually colder at night than is the air. The cold surface cools the air in its vicinity, and if the air contains sufficient atmospheric humidity, it may cool below its dew point. When this happens, water vapour will condense out of the air onto the surface (see also MICROCLIMATES; HUMIDITY, ATMOSPHERIC).

The formation of dew is sustained by the diffusion of water vapour. Regarding the vertical diffusion of water vapour over soils carrying vegetation, there are two possible situations. First, there is the downward movement of water vapour from the atmosphere, which occurs when the water vapour content of the air increases with height. Second, there is the upward movement of water vapour, which occurs when the soil surface temperature is higher than that of the leaves. Accordingly, dew may be classified (1) as formed when water vapour diffuses downward in the air and (2) as formed from water vapour diffusing from the underlying soil surface. The name dewfall is proper to (1), and dew arising from (2) may be called distillation.

Dew measurement

There have been various attempts to measure dew. Among the various instruments are R. Leick's porous gypsum plates and S. Duvdevani's dew gauge, consisting of a wooden slab treated with paint. To determine the amount of dew, Leick's plates are weighed, whereas Duvdevani's gauge involves the use of an optical dew scale. Other investigators (*e.g.*, J.M. Craddock, E.G. Jennings, and J.L. Monteith) developed recording dew balances whose surface and exposure conform with the surrounding surface as far as possible. It is by means of such dew balances that one can best observe the phenomenon of distillation: on some occasions no gain in weight or even some loss in weight may be recorded despite the fact that dew had formed on the leaves. Clearly, this dew must be attributed to the diffusion of water vapour from one part of the weighed system to another; *i.e.*, from soil to leaves.

The amount of dew formed on plants is not well known. It would appear that during dew nights the amounts vary from very small quantities to about $\frac{1}{50}$ of an inch. G. Hofmann (*Die Thermodynamik der Taubildung*, Bad Kissingen, 1955) estimated that the maximum possible amount is about $\frac{3}{100}$ of an inch for a ten-hour night, but such amounts would occur only under exceptional circumstances. Total annual dew precipitation may lie between about one-half inch in cold climates and in nearly arid warm climates, to about three inches in semihumid warm climates. Because dew produced by distillation from the soil cannot be regarded as a gain of moisture, not all of the annual dew may be significant from a hydrological point of view. In some desert areas and semiarid regions the net gain may be a substantial fraction of the rainfall, however, and dew may be the principal moisture source for plants and animals. Under such conditions, it also may assume an important role in some aspects of rock weathering (*q.v.*). From the biological viewpoint, the usefulness of dew is doubtful as dew may stimulate the growth of fungi harmful to plants.

In ancient times it was believed that dewdrops fell from the sky. Aristotle suggested (in *Meteorologica*) that dew or hoarfrost was formed by the condensation of water vapour that had evaporated from the surface by day. Louis-Constant Prévost (*Recherches Physico-Mécanique sur la Chaleur*, 1792) was probably the first to connect

dew with radiation losses from the surface. The first thorough investigation of dew, however, was made by Charles Wells, physician to St. Thomas's Hospital, London. He showed (*An Essay on Dew and Several Appearances Connected With It*, London, 1814) that dew was the result of cooling of the air at night. He also pointed out that although dew tended to form on windless nights, "a slight agitation of the air . . . when the air is pregnant with moisture . . . will render greater the quantity of dew" —the first statement appreciating the effect of slight turbulence on dew. In 1885 John Aitken (*Trans. Roy. Soc. Edinb.*, no. 33) published the results of his studies, which demonstrated, among other things, that trays of turf at times lost weight when dew formed on the blades. He interpreted this as the result of vapour diffusion from the soil. Aitken also made a distinction between dewdrops proper and drops of water that are formed by guttation; *i.e.*, the exudation of water from leaves under the action of root pressure.

BIBLIOGRAPHY. N.T. ZIKEEV, "Annotated Bibliography on Dew," *Met. Abstr. Biblphy.*, 3:360–391 (1952), annotated listings from the year 1725 (a memoir by Robert Boyle) to the year 1951, encompassing 208 literature items; *Meteorological and Geoastrophysical Abstracts* (monthly), publishes annotated items on dew as they appear in the scientific literature; J.L. MONTEITH, "Dew," *Q. Jl. R. Met. Soc.*, 83:322–341 (1957), a paper that describes the results of a thorough experimental study and treats the physical-theoretical aspects of dew formation, including heat balance and turbulent transport considerations; S. DUVDEVANI, "An Optical Method of Dew Estimation," *ibid.*, 73:282–296 (1947), a description of the author's optical dew gauge, widely used for routine measurements, with an evaluation of its relative merits.

(J.N.)

Dewey, John

Almost equally famed as philosopher, psychologist, and educator, John Dewey shares with two fellow Americans (Charles Sanders Peirce and William James) the distinction of having founded the philosophical school of Pragmatism, is recognized as having helped establish the school of functional psychology, and is regarded as the outstanding thinker of the so-called progressive movement in American education and in educational reform during the first half of the 20th century. Dewey's political and social philosophy articulated and made more meaningful many of the ideals and deeper aspirations of the American people. He is generally recognized as an authentic voice of 20th-century American democracy. If there were such an office as that of national philosopher, writes Morris R. Cohen in his *American Thought*, Dewey is the person who could most properly be mentioned for it. When the University of Paris, in 1930, awarded him an honorary degree, it referred to him as "the most profound, most complete expression of American genius."

Early life. Dewey's family origins were modest. Three generations of his forebears were Vermont farmers, and both his parents were born and brought up on a Vermont farm. Archibald Dewey, John's father, when a young man broke with family tradition, however, and entered the grocery business in Burlington. Here many years later, at the age of 44, Archibald met and married Lucina Artemisa Rich, 20 years his junior. The couple had four sons, the first of whom died in infancy. John, the third child, was born on October 20, 1859, in a house that is still standing in Burlington.

The Burlington in which Dewey lived as a youth had a population of approximately 15,000, divided almost equally between native- and foreign-born; the latter came chiefly from Ireland and Quebec. The native-born included the "old Americans," descendants of middle class, Anglo-Saxon, Protestant families long established in Vermont or some other part of New England, and it was in the traditions of this group that Dewey was brought up.

Dewey attended the public schools of Burlington and there entered the University of Vermont. Particularly significant to him were the studies of the fourth year, designed to introduce the student to fundamental political, economic, philosophical, and religious theories and intended to be the capstone of the student's undergraduate academic experience. Dewey found these courses interesting and provocative, and his thinking along broad intellectual and philosophical lines may be said to have started during his senior year in college.



Dewey.
EB Inc.

Dewey's high school and college days marked the time when his intellectual interests were broadening and when he sought the company of books. He was, as he himself has written, a "voracious reader" at this time and found in books a source of deep satisfaction. His parents encouraged him in his reading, and Dewey was free to peruse almost any book which caught his fancy.

Interest in books, however, did not at this time crowd out other interests, as it threatened to do in Dewey's graduate-student years. He welcomed opportunities to swim and fish in Lake Champlain, only three blocks away from where he lived as a boy, and to hike, climb, and camp overnight in the Green Mountains. During the summer he was expected to work at a job, and this he did in the extensive lumber yards that at the time lined the shores of Burlington Bay.

Though he was allowed considerable freedom in his studies, reading, and recreational activities, Dewey was carefully restricted in other ways. His mother was deeply pious and required that her sons go to Sunday school, attend church services, and in other ways properly observe the sabbath. As they grew older she cautioned them against the temptations lurking in neighbourhoods bordering Burlington's industrial area on the lakefront, and when they had been out for an evening, upon their return, she would very likely question them as to where they had been and what they had done. In later years Dewey recalled with distaste this phase of his upbringing and recalled also the sense of guilt that this questioning evoked, even when there was nothing to be guilty about.

After finishing college, Dewey taught three years in high school and, in the fall of 1882, entered Johns Hopkins University, in Baltimore, for advanced study in philosophy. Here he came under the influence of George Sylvester Morris, visiting professor of philosophy from the University of Michigan and a leading exponent of Neo-Hegelianism, a revival of the thought of the 19th century German philosopher Hegel. Dewey found in this philosophy, with its emphasis on the spiritual and organic nature of the universe, what he had been vaguely groping for, and he eagerly embraced it.

Influence of Hegel's philosophy

While at Johns Hopkins, Dewey devoted himself almost exclusively to his studies and tended to neglect other activities. This preoccupation with study was noticed by the president of the university, who strongly advised him to study less and socialize more.

Upon being awarded the Ph.D. degree by Johns Hopkins University, in 1884, Dewey, in the fall of that year, went to the University of Michigan, where, at the urging of Morris, he had been appointed an instructor in philosophy and psychology. With the exception of the aca-

demic year 1888–89, when he served as professor of philosophy at the University of Minnesota, Dewey spent the next ten years at Michigan. During this time his philosophical endeavours were devoted mainly to an intensive study of Hegel and the British Neo-Hegelians and to the new experimental, physiological psychology then being advanced in the United States by G. Stanley Hall and William James.

Dewey's interest in education began during his years at Michigan. His readings and observations revealed that most schools were proceeding along lines set by early traditions and failed to adjust to the latest findings of child psychology and to the needs of a changing democratic social order. The search for a philosophy of education that would remedy these defects became a major concern for Dewey and added a new dimension to his thinking.

At the end of his second year at Michigan, Dewey married Harriet Alice Chipman of Fenton, Michigan. She had been a student in a number of Dewey's classes, and the friendship culminated in their marriage on July 28, 1886. Mrs. Dewey's academic interests centred mostly on social and educational problems, and it was in large measure because of her that Dewey's interests widened to include these. Alice had an outgoing personality, liked people, and helped her husband overcome the shyness with strangers that had troubled him since early adolescence. Both Alice and Dewey were fond of children and were pleased as their family grew to include three sons and three daughters. Two of the sons died in early childhood; and, to help soften their loss, the Deweys, while travelling in Italy, adopted an Italian boy near the age of one of the sons who had died. The adoption proved a happy one.

Career at the University of Chicago. Dewey left Michigan in 1894 to become professor of philosophy and chairman of the department of philosophy, psychology, and pedagogy at the University of Chicago. One feature of the new position that appealed to him was that the department was responsible for instruction not only in philosophy but also in psychology and pedagogy. Dewey saw in this arrangement an opportunity to unite these three disciplines and, more particularly, to bring pedagogy into closer relations with psychology and philosophy.

Dewey's achievements at the University of Chicago brought him national fame. The increasing dominance of evolutionary biology and psychology in his thinking led him to abandon the Hegelian theory of ideas, which views them as somehow mirroring the rational order of the universe, and to accept instead an instrumentalist theory of knowledge, which conceives ideas as tools or instruments in the solution of problems encountered in the environment. These same disciplines contributed somewhat later to his rejection of the Hegelian notion of an Absolute Mind manifesting itself as a rationally structured, material universe and as realizing its goals through a dialectic of ideas. Dewey found more acceptable a theory of reality that holds that nature, as encountered in scientific and ordinary experience, is the ultimate reality and that regards man as a product of nature who finds his meaning and goals in life here and now.

Since these doctrines, which were to remain at the centre of all of Dewey's future philosophizing, also furnished the framework in which Dewey's colleagues in the department carried on their research, a distinct school of philosophy was in operation. This was recognized by William James in 1903, when a collection of essays written by Dewey and seven of his associates in the department, *Studies in Logical Theory*, appeared. James hailed the book enthusiastically and declared that with its publication a new school of philosophy, the Chicago school, had made its appearance.

Advent of the Chicago school of philosophy

Dewey's contributions to psychology were also noteworthy. Many of the articles written at that time are now accepted as classics in psychological literature and assure him a secure place in the history of psychology. Most significant is the essay "The Reflex Arc Concept in Psychology," generally taken to mark the beginnings of functional psychology—*i.e.*, one that focusses on the total organism in its endeavours to adjust to the environment.

His writings on education, notably his *The School and Society* (1899) and *The Child and the Curriculum* (1902), presented and defended what were to remain the chief underlying tenets of Dewey's philosophy of education. These were that the educational process must begin with and build up on the interests of the child; that it must provide opportunity for the interplay of thinking and doing in the child's classroom experience; that the school should be organized as a "miniature community"; that the teacher should be a guide and coworker with the pupils, rather than a taskmaster assigning a fixed set of lessons and recitations; and that the goal of education is the growth of the child in all aspects of its being.

Among the results of Dewey's administrative efforts were the establishment of an independent department of pedagogy and of a laboratory school in which the educational theories and practices suggested by psychology and philosophy could be tested. The school, which began operations in January 1896, attracted wide attention and enhanced the reputation of the University of Chicago as a foremost centre of progressive educational thought.

The city of Chicago provided Dewey with an excellent opportunity to observe the operations of raw, unrestricted industrialism and to note its effects on large segments of the city's population. As a trustee of Hull House, a settlement house in a Chicago slum area, he became familiar with the exploitation of immigrant and minority groups and joined radicals and other liberals like himself in supporting legislation to aid the underprivileged, legalize labour unions, and curtail the power of monopolies. At Hull House, Dewey was introduced to Jane Addams, the distinguished founder of the organization. A warm and lasting friendship developed between them.

Career at Columbia University. The disagreement between President William Rainey Harper of the University of Chicago and Dewey concerning the administration and financing of the university's educational program led, in 1904, to Dewey's resignation of his posts and to his acceptance of a professorship of philosophy at Columbia University in New York City.

Dewey was associated with Columbia for 47 years, first as professor and then as professor emeritus of philosophy. During his 25 years of active teaching, his fame and the significance of what he had to say attracted thousands of students from at home and abroad to his classes, and he became one of the most widely known and influential teachers in America.

Dewey's scholarly output at Columbia was enormous; one bibliography devotes approximately 125 pages to listing the titles of his publications during these years. His thought covered a wide range of topics, including logic and theory of knowledge, psychology, education, social philosophy, fine arts, and religion. Major works dealing with each of these fields appeared over the years and clearly established Dewey as the foremost philosopher in America and as one of the nation's most productive scholars. His *Experience and Nature*, published in 1925, brings together in a systematic way the more important aspects of his philosophy and is generally regarded as his magnum opus.

Publication of *Experience and Nature*

His interest in current affairs prompted Dewey to contribute regularly to liberal periodicals, especially *The New Republic*. His articles focussed on domestic, foreign, and international developments and were designed to reach a wide reading public. Because of his skill in analyzing and interpreting events, he soon was rated as among the best of American commentators and social critics.

A man of action as well as of thought, Dewey gave generous amounts of time and energy to the support of organizations and causes in which he believed. After World War I, with S.O. Levinson, a Chicago attorney and long-time friend, he played an important role in the "outlawry of war" movement, which helped bring the Kellogg–Briand Pact of 1928 into existence. He was one of the founders and the first president of the American

Association of University Professors and was a charter member of the first teachers' union in New York City. He helped found the New School for Social Research in 1919 and the University-in-Exile in 1933, established for scholars being persecuted in countries under totalitarian regimes, and he held office in the American Civil Liberties Union, the League for Industrial Democracy, and the People's Lobby. He thought that the two major parties in Congress had failed to come to grips with the problems generated by the Depression of the 1930s, accused them both of being the "errand boys of big business," and took a leading part in an attempt to organize a third political party. In 1937, at age 78, he headed a commission of inquiry that went to Mexico City to hear Leon Trotsky's rebuttal of the charges made against him in the Moscow trials of 1936 and 1937.

His reputation brought Dewey invitations from a number of foreign countries, among them Japan, China, Turkey, Mexico, Russia, and South Africa. Russia, Mexico, China, and Turkey were undergoing social upheavals, which Dewey reported on in articles sent to *The New Republic*.

Great personal loss was suffered by Dewey when his wife died in 1927. The two had been unusually close, and her death left a void not easily filled. After her death Dewey shared an apartment with one or another of his children until 1946, when he married Roberta Lowitz Grant, whose parents he had known when Roberta was a child in her early teens. Dewey was 87 and Roberta 42 when they were married. Shortly after their marriage they adopted two small Belgian children, a sister and brother, made orphans by the war. The four quickly became a closely knit and affectionate family group.

Second family

Dewey was the recipient of many honours. He was awarded honorary degrees from at least 13 colleges and universities at home and abroad and was decorated with the Order of the Jade by the government of China in 1939 and the Order of Merit by the Chilean government in 1949. He was elected honorary president of the National Education Association in 1932, honorary president for life of the American Philosophical Association in 1938, and honorary vice president of the New York State Liberal Party in 1952. Meetings and dinners celebrating the anniversaries of his 70th, 80th, and 90th birthdays and attended by notables from all walks of life were held in his honour in cities throughout the nation.

During the last two decades of his life, Dewey's philosophy of education was the target of numerous and widespread attacks. Greatly agitated by the alleged failure of the schools to train pupils adequately in essential core subjects and in manners and discipline, critics blamed Dewey and his progressive ideas for these failures and in article after article made him the scapegoat of their grievances and frustrations.

Dewey enjoyed remarkably good health, and rarely during his long career was he forced to miss classes or cancel off-campus engagements because of illness. But his advancing years were exacting their toll, and during his 80s he required repeated hospital attention. He remained physically active, however, and participated in the celebration of his 90th birthday in New York City and in his native Burlington and also accepted in person the honorary degree awarded him by Yale University at its 250th commencement in June 1951.

In November 1951, Dewey broke his hip, and the failure of the bones to knit properly kept him weakened and confined to his apartment during the winter and spring of 1952. Yet he remained intellectually busy and continued to work on projects he had begun earlier, including a revised edition of his *Experience and Nature*. On May 31 he was stricken with pneumonia, from which he failed to recover. He died in the early evening of June 1, 1952.

MAJOR WORKS

EDUCATION: *The School and Society* (1899); *Democracy and Education* (1916); *Experience and Education* (1938).

PSYCHOLOGY AND PHILOSOPHY: *Psychology* (1887); *Studies in Logical Theory* (1903); *Ethics*, with James Tufts (1908); *How We Think* (1910); *Reconstruction in Philosophy* (1920); *Human Nature and Conduct* (1922); *Experience and Nature* (1925); *The Public and Its Problems* (1927); *The Quest for Certainty* (1929); *Art As Experience* (1934); *A Common Faith* (1934); *Logic, the Theory of Inquiry* (1938); *Freedom and Culture* (1939); *Knowing and the Known*, with Arthur F. Bentley (1949).

BIBLIOGRAPHY. JOHN DEWEY, "From Absolutism to Experimentalism," in GEORGE P. ADAMS and WILLIAM P. MONTAGUE (eds.), *Contemporary American Philosophy: Personal Statements*, vol. 2, pp. 13–27 (1930), a sketch in which Dewey traces the stages of his intellectual development and the main influences encountered; JANE DEWEY (ed.), "Biography of John Dewey," in PAUL ARTHUR SCHILPP (ed.), The *Philosophy of John Dewey*, pp. 3–45 (1939), an account of Dewey's training, philosophical development, teaching positions, family and friendships, and travels; MAX EASTMAN, "John Dewey," *Atlantic Monthly*, 168:671–685 (1941), an account written by a friend who gives interesting details of the life of the Dewey family; GEORGE DYKHUIZEN, "John Dewey: The Vermont Years," *Journal of the History of Ideas*, 20:515–544 (1959); "John Dewey at Johns Hopkins (1882–1884)," *ibid.*, 22:103–116 (1961); "John Dewey and the University of Michigan," *ibid.*, 23:513–544 (1962); "John Dewey: The Chicago Years," *Journal of the History of Philosophy*, 2:227–253 (1964); and "John Dewey in Chicago: Some Biographical Notes," *ibid.*, 3:217–233 (1965), detailed accounts of Dewey's life during the period indicated; IRWIN EDMAN, *Philosopher's Holiday*, pp. 138–143 (1938), a fascinating account of Dewey as a teacher by a former student; SIDNEY HOOK, "Some Memories of John Dewey," *Commentary*, 14:245–253 (1952), an essay that gives an appreciative account of Dewey as a person, written by a former student and close friend; *John Dewey: An Intellectual Portrait* (1939), a very readable and authoritative account of the several aspects of Dewey's philosophy; GEORGE R. GEIGER, *John Dewey in Perspective* (1958), an exposition of Dewey's philosophy in the light of the misunderstandings and distortions which have centered about it; RICHARD J. BERSTEIN, *John Dewey* (1966), an illuminating study of the leading ideas of Dewey's philosophy; PAUL ARTHUR SCHILPP (ed.), *The Philosophy of John Dewey* (1939), contains in addition to a biographical sketch of Dewey, a collection of 17 essays by different authors, each a critical study of some aspect of Dewey's philosophy; a concluding essay by Dewey in response to his critics; and a bibliography of Dewey's writings to October 1939; JO-ANN BOYDSTON (ed.), *A Guide to the Works of John Dewey* (1970), a collection of essays by Deweyan scholars each introducing the reader to a particular area of Dewey's thought; MILTON HALSEY THOMAS, *John Dewey: A Centennial Bibliography* (1962), the most complete listing now available of writings by and about Dewey.

(G.Dy.)

Diaghilev, Sergey

Promoter of the arts and founder of the Ballets Russes, Sergey Pavlovich Diaghilev revitalized the stereotyped form of ballet of the early 20th century. Integrating the ideals of other art forms—music, painting, drama—with those of the dance, he developed the ballet as a creative whole. As an innovator of artistic movements in Europe during the first quarter of the century and as the inspiring genius of musicians, painters, and dancers, he holds a unique place in the development of the arts in the 20th century.

Born in the province of Novgorod on March 19, 1872, Diaghilev was the son of a major general and a noblewoman, who died in childbirth. Diaghilev was remarkable for his large, dreamy eyes acquired from his mother, and perhaps he also inherited from her his craving for luxury in both the personal and artistic spheres. He repeatedly made it clear that he followed a purely hedonistic philosophy. From his stepmother, Helen Valerianova Panaieva, on the other hand, he acquired a sense of discipline and the instinct to dominate. Moreover, as a youth his artistic sensibilities were encouraged by her many musical connections. The piano lessons he took while still at school enabled him to play the Schumann concerto in public; as a child he also showed a gift for composition.

In 1890, while studying law at the University of St. Petersburg, he became associated with a group of friends interested in the social sciences, music, and painting—the first of a series of intellectual gatherings over which he presided throughout his life. Among his companions dur-



Diaghilev, *c.* 1916.
By courtesy of the Dance Collection, the New York Public Library at Lincoln Center, Aster, Lenox and Tilden Foundations

ing this period were the painters Alexander Benois and Léon Bakst, both of whom were later to contribute brilliantly to his productions. His first experience of ballet was, curiously enough, a disappointment. The work was Peter Ilich Tchaikovsky's *Sleeping Beauty*, which he saw with little enthusiasm at the Mariinsky Theatre in St. Petersburg about 1890. (The work was later to become one of Diaghilev's most successful productions.)

In 1893 he made his first journey abroad, visiting Germany, France, and Italy, where he met the distinguished French novelist Émile Zola and the opera composers Charles Gounod and Giuseppe Verdi. Even in his youth he sought out and was stimulated by the company of the great.

In 1896 Diaghilev graduated in law, but he was determined to follow a musical career. The composer Nikolay Rimsky-Korsakov, however, discouraged him from developing his talents as a composer, wisely no doubt, since a vocal work of Diaghilev that had been performed in public had left a poor impression. In Moscow he met the patron of the famed bass Fyodor Chaliapin and proposed revolutionary scenic ideas for productions of operas in which Chaliapin appeared. Although he was uncertain of his own artistic gifts, Diaghilev was convinced of his vocation: that of a great patron of the arts like the Roman Maecenas. His theatrical ventures in the sphere of opera and ballet and his literary projects, demanding huge investments, were hampered by the fact that he embarked on this career with no private income. Moreover, in 19th-century Russia, his homosexuality was a serious handicap in the development of his career. He had personal charm and audacity, however, and he used them to advantage.

Vocation as an impresario

In 1899 he realized the first of these international ventures when he founded, as editor in chief, the review *Mir Iskusstva* ("World of Art") which continued to appear until 1904. This was a counterpart of the London *Yellow Book*, reflecting the ideas of the graphic artist Aubrey Beardsley and the writer Oscar Wilde.

In 1902 Diaghilev published a monograph on the Russian portrait painter Dmitry Gregoryevich Levitsky (1735–1822), and three years later he organized a historic portrait exhibition of Russian art treasures at the Tauride Palace in St. Petersburg.

The great turning point in his life came when he left Russia for Paris in 1906. It was there that he helped to found what was later referred to as the Franco-Russian artistic alliance. He organized an exhibition of Russian art and then, in 1907, a series of historic concerts devoted to the work of the Russian nationalist composers. In 1908 Modest Mussorgsky's opera *Boris Godunov* was produced in Russian by Diaghilev at the Paris Opéra with Chaliapin in the title role.

The time had arrived for him to launch the venture that was to fulfill his ideal of a combination or interpenetration of the arts. Appointed in 1899 as assistant to Prince Sergey Volkonsky, director of the Imperial Theatre, Diaghilev had met the dancer Michel Fokine, whose creative activity was powerfully influenced by the American dancer Isadora Duncan. Influenced by the dance innovations of Isadora Duncan, the ideas of composer Richard Wagner, and the theories of the poet Charles Baudelaire, Diaghilev opened his season of Ballets Russes at the Théâtre du Châtelet in Paris in 1909. The dancers Anna Pavlova, Vaslav Nijinsky, and Michel Fokine were in his company.

First season of the Ballets Russes

Before long it became clear that conventional choreography was to have no place in his novel spectacles. Mime or action dances were the aim of the choreographers who, largely under the influence of Fokine and Léonide Massine, were creating an entirely new tradition. The composers chosen to transform the old art forms were themselves inspired by the fantasies of painters and choreographers. This was Diaghilev's lofty creation, an ideal of artistic synthesis, based on an innate sense of taste. Diaghilev's art reached its height in the three early ballets of the young Russian composer Igor Stravinsky: *The Firebird* (1910), *Petrushka* (1911), and *The Rite of Spring* (1913). In *Petrushka*, perhaps the greatest of the Diaghilev ballets, Stravinsky, at Diaghilev's insistance, transformed a conventionally conceived piano concerto (on which he had been working) into a mimed ballet, bringing into real life the fantasy dramas of puppets at a showman's fair. The incident is indicative of the extraordinary psychological influence Diaghilev was able to exert over his collaborators. In *The Rite of Spring* Stravinsky produced one of the most explosive orchestral scores of the 20th century, and the production created an uproar in the Paris theatre at its first performance. The scandalous dissonances and rhythmic brutality of the music provoked among the fashionable audience such protestations that the dancers were unable to hear the orchestra in the nearby pit. They carried on, nevertheless, encouraged by the choreographer Nijinsky, standing on a chair in the wings, shouting out and miming the rhythm.

Diaghilev left his native Russia and never returned. An aristocrat, he carried on the revolution in the arts not in the Soviet Union, but in the intellectual circles of Paris. There he collaborated with the French poet Jean Cocteau, among others. He toured with his ballet throughout Europe, in the United States, and in South America. Despite his influence, however, Diaghilev was a lonely and dissatisfied man, impecunious and personally unhappy. He was an idealist, never realizing perfection and yet sowing the seed of an exploratory spirit. Seasons of the Diaghilev ballet were given uninterruptedly from 1909 to 1929. During his later seasons he introduced the works of forward-looking composers and painters from France, Italy, Great Britain, and the United States. Through Diaghilev's intuition they entered the world scene. Among the composers represented in his repertory were Richard Strauss, Claude Debussy, Maurice Ravel, and Sergey Prokofiev.

Diaghilev had long suffered from diabetes, and by the end of his brilliant 1929 season at Covent Garden, London, his health had gravely deteriorated. He nevertheless left for a holiday in Venice, where, following an alarming rise in temperature, he sank into a coma. He died there on August 19, 1929, and was buried in the island cemetery of San Michele.

BIBLIOGRAPHY. S. LIFAR, *Serge Diaghilev* (1940); and ARNOLD L. HASKELL and W. NOUVEL, *Diaghileff: His Artistic and Private Life* (1955), are authoritative biographies. S.L. GRIGORIEV, *The Diaghilev Ballet, 1909–1929* (Eng. trans. 1953), is a standard work on the Russian ballet; and JOHN PERCIVAL, *The World of Diaghilev* (1971), a succinct, popular study.

(E.L.)

Diagnosis

The term diagnosis refers either to an active process or to the conclusion reached by that process. Modern medical diagnosis in the active sense includes the process and art of using scientific methods to elucidate the whole compass of problems that influence a sick person. It includes the collection of all necessary facts and critical evaluation of every bit of evidence obtained from any and all sources by whatever method is useful. It is a lively art based on a sound and growing science. From the facts so obtained, and in the light of a knowledge of the principles of anatomy, physiology, and pathology, concepts of the causes of the trouble, the pathological lesions, and the disordered processes that make up the patient's disease are formed. From an array of possible hypotheses, the correct diagnosis is singled out as having the best fit with the findings. The process of selection is properly called differential diagnosis.

Diagnosis is ordinarily the forerunner of treatment, and treatment must be based on an understanding of diagnosis and of prognosis; that is, the outlook of a patient without treatment and with various kinds of treatment, including the accepted risks that certain treatments necessarily entail, whether the treatment be taking a pill, receiving an injection, or undergoing a surgical operation. The information upon which the diagnosis–prognosis–treatment continuum is based falls into three classes, each requiring a separate critical analysis for accuracy, error, and clinical significance for the particular patient involved.

Types of information leading to diagnosis

These classes are (1) facts obtained from the patient's story of his life, his health, his accidents, operations, and illnesses, as a specific unique member of society—the medical history; (2) facts obtained from the physical examination; and (3) data obtained from routine laboratory examinations and special tests. The next step is the development of the medical record—the recording of data followed by a codification of the important data and summary ordering of them in terms of their significance, from which a preliminary diagnosis is derived. This is then modified as further explicit tests are obtained to correct errors, solve puzzles, and measure degrees of abnormality more precisely. An estimate is made of prognosis. Treatment is then undertaken, with or without modification of the original diagnosis.

When the term diagnosis is used as a label, it refers to the conclusion that stems from the process of diagnosis. The conclusion cannot be considered final until after an examination of tissue specimens, surgical exploration, or autopsy, and even then the diagnosis is rarely a complete and total summary.

Any explanation of medical phenomena must be in terms of contemporary knowledge and thus must be provisional. It is based on the natural history of disease as determined by man's observing man in his environment. Historically, observation was, for a time, simply looking and seeing. The next development was the actual examination of the sick person, in which history taking, or seeking the seeds of the present trouble from the story of a patient's past health and disease, was a major factor. In the early historical period, and probably for ages before that, the examination was a bedside examination. One of the first advances from a mere bedside examination of the sick person was scrutiny of his urine by a physician-priest. Today, technical machinery and laboratory tests have created a myriad of instruments and procedures. Every day there are more kinds of scopes, electrical measurements, X-rays, and tolerance tests. In such a way, the physical examination carries the fluid secretions and excretions, groups of cells or bits of tissue, from the patient to the laboratory. Tumours and organs may be removed for examination.

Hippocrates knew the succussion splash; *i.e.*, the sound made when anyone (from a sick senator to a slave) with air and fluid free in the chest cavity was shaken. He heard and felt the sounds of the creaking of joints, intestinal rumblings, belching, and the expulsion of flatus. These were commonplace and treated by not being treated. But no man had real understanding of what went on inside until the anatomists had paved the way for an Austrian physician, Leopold Auenbrugger, who developed the use of percussion (see below) in diagnosis and a French physician, René Laënnec, who developed the stethoscope. William Harvey had added experimentation to biological observation and thus reduced the amount of time and energy wasted on speculation detached from data.

The introduction of active examination into a process that had been merely observation helped set modern medicine on its way. The success of the physical examination steadily improved when it became possible to correct findings by autopsy.

In the distant past, medicine was a matter of opinions supported by other opinions. It dealt in terms of myth that was ultimately modified, but neither organized nor stabilized, by astrology and alchemy.

Development of anatomy and pathology

Medical science could not exist until the body was first systematically dissected. Thus, among the sciences, anatomy came first. Perceptive men realized that there was a complementary relationship between structure and function. Ideas of the physiology of living functions began to develop. Next, pathology began to evolve on a sound anatomical basis, passing far beyond the primitive pathology of war wounds, injury, and the visible diseases of skin, bone, deformity, or ruin. A first order of understanding came with the disclosure in autopsy of such diseases as the fatal effect of spontaneous massive internal bleeding, or the mechanical ruin of various fabrics of the body by invasion of a destructive tumour. Partial or total mechanical obstruction of the body's pipes and tubes, a first comprehension of many of the mistakes of nature, deformities and inborn errors that were not immediately lethal, began to recall insights forgotten since the time of Alexandrian medicine.

With William Harvey's work on the circulation of the blood, physiology suddenly grew up. Experimental medicine, a twin of the revelation of the circulation, was born. With medicine's first understanding of the mechanics and hydraulics of the movement of blood, the function of the lungs could be seen as something other than a sophisticated cooling mechanism for the heart. Further advance in the workings of the lungs became possible only when the chemistry of gases had been understood.

Then came a series of advances unexpected and unorganized. The development of the microscope and the staining of tissues led scientists much deeper into the microcosm. Electrophysiology added yet another dimension. Observations that could not be clearly understood at the bedside were taken to the laboratory. Clues were brought back. They were tested. Some failed. A few worked. The process continued.

The natural history of disease has always provided the chemist with some of his most challenging opportunities. The American chemist Linus Pauling, who uncovered the nature of the physical flaw of the hemoglobin in sickle-cell anemia, for example, could not have done so if someone else had not first described and classified the anemia.

The changing concepts of disease

Just as certain elements that were typical of each stage and type of medical diagnosis are still used today—the mere observation, the careful examination coupled with the sick person's account of his present and past circumstances and ills, the laboratory and other tests—so too, traces of man's changing ideas of disease linger on: that disease is a magic spell or curse, a spirit that invades man's body; that disease is the experience of a particular person who can be observed, interrogated, and examined (the teachings of the Greek physician Hippocrates and his school); that there is no disease, just as there is no real external world, the Stoic concept; that disease is retribution and punishment; that the physical body and the cell are the seat of disease, the anatomic concept; that structure and function are both involved, the physiologic concept; that each disease has its own particular cause, the etiologic concept; that man is the source of his own ills, the psychologic concept; that disease is maladaptation to environment, the socioeconomic concept.

Finally, there emerges an ecologic concept, the understanding of a composite in which man and life are made up of many forces and entities interacting. Man's struc-

ture and function, as determined by his genetic legacy, become for each man a unique medley of nearly numberless possibilities. The accretions of changes from the physical events and memory traces of his own past and the race's past are continuously impinged upon by stimuli, great and small, sure and unsure, heard and unheard, inside and out. Life then becomes a process. It operates in time; goes on in a finite place. Man both reacts to and influences his environment. Both change. Disease, too, is a process. The body's functions fail or decay, the machinery rattles when adjustments are inadequate or the stimuli too great or too many. Disease results when adaptations, reactions, repair, and renewal diminish.

Today, disease is seen through a series of concepts, and vigilance must be exercised to keep separate the sometimes interchanged meaning of word, idea, and thing.

This article is divided into the following sections:

I. Direct evidence from the patient
The medical history
Physical examination
II. Biological, chemical, and radiological tests
Laboratory diagnosis
Multiphasic health screening
Use of radiations in diagnosis
III. Other important considerations
Surgical diagnosis
The medical record
Sources of error in medical diagnosis

I. Direct evidence from the patient

THE MEDICAL HISTORY

Factors affecting quality of medical history

The medical history, the information gleaned from the patient concerning everything that could have a direct or remote bearing on the present illness, must be as accurate and as exact as conditions permit. The excellence of the history depends on the patient's memory, his judgment, his present and past health, and his articulateness, and accuracy. The history must include vital statistics, dates, habits, occupation, and many points that may not bear immediately upon the problem at hand. Even when working with a well-informed and most articulate patient, the record may be poor if the physician who prepares it is hurried, careless, writes illegibly, fails to include pertinent material, or fails to become skilled in the use of the searching, carefully phrased question.

Most people are interested in talking about themselves. An unguided discourse is likely to be rambling, full of irrelevancies, and difficult to use or to learn from. It may blur the focus and distract the physician. Since the physician must take the patient's story apart and look at his dissection of it in terms of pathological physiology as indicative of departures from normal form and function, the skillful physician translates the ordinary patient's common speechways into what he understands as departures from normal function, on the basis of all the evidence that he can accumulate. He then translates these findings into a program of further diagnosis and therapy. The success of a physician, particularly in chronic illness, often will depend on a careful discussion and interpretation, in simple words that the patient can understand, of the nature of what is wrong, the necessity for therapy, and the nature of the therapeutic program.

Present illness. In obtaining information on the current state of the symptom complex called an illness, the physician guides the patient away from making diagnoses—she has seen a television program on cancer, perhaps; or a friend has died of it. At the same time, however, precise detail must be secured about the way the patient feels and the functions that seem to be or are out of kilter. Often the absence, the presence, or the degree of concomitant features will enable the interrogator to put the proper weight and significance upon each symptom.

Since the patient is likely to be frightened, worried, concerned, in pain, or uneasy, skilled physicians try to calm him by being calm, show interest by being interested, and exhibit genuine sympathy upon which not only successful diagnosis but successful management depend. The doctor has empathy and sympathy; he is neither overbearing nor oversolicitous.

Always a person's reason for seeing a physician is clearly identified. In effect, the interview is opened by asking, "What may I do for you?" rather than, "What brought you here?" Few patients view their own problems unemotionally or even perceptively. After a patient has made his initial statement, the physician asks questions and keeps the focal point of the discussion on the patient's actual problems. While the physician must possess and exhibit genuine concern and care, he must avoid all moral judgment, favourable or unfavourable, upon the patient.

Reliability of patient's story

The value of a patient's story depends upon the assumption that it is truthful. Errors may occur from forgetfulness or a failure to understand the intent of a question. Occasionally there may be willful deceit out of fear or for ulterior motives such as the thought of compensation claims.

From the patient's history, clues for diagnosis will evolve. They may be only vaguely suggestive, or tantamount to a precise and complete diagnosis. Thus, as the physician accumulates information from which he derives medical data that are continually evaluated in the light of additional details of time, place, degree, and intensity, he begins to form a judgment of the credibility of the patient.

Toward an analysis of pain. As an example of the diagnostic query, the problem of pain looms large in medical experience. The physician usually wants to know several things about any particular pain. In order to experience pain, a person requires: (1) a stimulus; (2) a mechanism for transmission of the impulse; (3) an area in the higher levels of the central nervous system where perception occurs; (4) a state of conscious awareness. Diagnosis depends in part on a person's description of what he feels, and the description will be conditioned by the patient's intelligence, imagination, clarity of mental processes, earlier experiences with pain, and many other factors.

Localization and extension of pain

The first thing that the diagnostician must learn about pain is its location, first the general area in which it is felt and then the more precise localization in depth; *i.e.*, whether it is confined to an area that might be pointed to by a finger, the finger tips, the hand, both hands, or hands rubbed over the entire abdomen, and whether it is superficial or deep, or both. Next, the doctor is concerned with the extension of the pain by spread or irradiation, whether another pain, called referred pain, is felt simultaneously at a distance. Diagnostically it is important whether the pain extends from the point of origin over a continuous territory or is, for example, in the so-called ulnar portion of the left arm (that is, in the forearm on the side away from the thumb) and occurs simultaneously with separate pain under the breastbone, as occurs in angina pectoris, the suffering associated with the heart muscle's reception of insufficient oxygen.

Diagnosis is also concerned with the characteristics of the pain; the quantity and the quality of it, its severity, its character. All that can be learned of the character of a pain is learned from what the patient says by way of analogy with familiar experience. Thus, pain may be said to be burning, crushing, stinging, throbbing, or cramplike. Such information needs to be considered with a clear understanding of whether the patient is ordinarily sensitive to pain, rather insensitive, or hypersensitive. His reaction to pain may be a hypochondriacal exaggeration or a stoical denial or downgrading of its severity.

The next series of diagnostic factors about pain are concerned with time, how long it lasts, whether it is a flash of lightning that appears for an instant and is gone, a slow but steadily intensifying sensation that builds up to a peak and then fades out, whether it is first established at an agonizing plateau of anguish and misery, is constant and enduring for hours, or brief—a crisis or a nuisance.

The next factor is the frequency with which pain occurs, if it is a repeated experience, whether it is habitual and inveterate, comes every day or many times a day, or only occasionally and in unforeseen circumstances, is irregular or has happened only once.

The exact circumstances under which pain occurs are important and generally must be considered in terms of special influences on pain: whether or not these are fac-

tors that induce or aggravate it; whether, for example, a specific degree of exercise invariably induces it, its relation to meals, whether it is provoked by motion of a particular joint or limb or by general exercise, whether it comes only in circumstances charged with emotion or on the basis of season or weather.

The circumstances that aggravate a pain when it exists may be different from those that induce it, though they tend to be similar or identical.

Diagnosis is concerned with such pain-attending manifestations as shortness of breath; faintness; eructation; belching; passing of flatus; abdominal rumblings; flushing; a sense of dying, *angor animi*.

While a patient's story is the first and most important source of clues about pain, it must be related to what is learned in the examination of the patient. Much referred pain is interpreted by the brain as coming from sensitive structures innervated by the same segment of the spinal cord that is connected to nerves through the autonomic nervous system. Referred pain may be intense when an organ in the body proper is disturbed. The supporting structures of the skeleton must be considered—*i.e.*, bones with the membrane, the periosteum, that encloses them and tendons and ligaments with their attachments. The muscle mass, which, with all of the muscles taken together, is the largest organ in the body, may be the site of referred pain or may have pain associated with insufficient blood supply, spasm, injury, cuts, bumps, bruises, or encroachment or invasion by bleeding, tumour, or other abnormal material.

Referred pain

Next, the sensory nerves themselves may have lesions. It is diagnostically critical that a disease affecting the peripheral nerves of infectious, toxic, or metabolic origin may itself be responsible for sometimes bewildering pains.

Pain in the chest, in addition to coming from the supporting structures of skeleton, muscle, ligament, and skin, may come from a bewildering variety of abnormalities that affect the internal structures. The diaphragm (the muscular partition between the chest and the abdomen), for example, may be injured or irritated by disturbances of the upper portion of the abdomen by such phenomena as hiccups, spasms, or the stitch in the side that occurs when running. In diaphragmatic irritation, pain may be felt at the base of the neck because the central diaphragmatic muscle mass migrated down from the region of the neck in embryonic life after the nerves had formed their attachments, taking the nerves with them.

The examining physician must be aware that pain in the chest may come from a variety of disorders of the alimentary canal, which include diverticula (abnormal pouches) of the esophagus, spasm, failure of the sphincters to open properly, ulceration of the esophagus, and hiatal hernia—protrusion of the stomach through the esophageal opening in the diaphragm; and that pain in the lower chest may be provoked by peptic ulcer (ulceration of stomach or intestinal lining), inflammation of the pancreas, gallbladder disease, and even insufficient blood supply to the intestine—ischemia. Nearly all the organs and structures in the chest must be considered.

Evaluation of pain leads the examiner to questioning based on knowledge of the anatomical structures that may themselves give rise to pain or may be important in conveying the nerve impulses correlated with pain. The heart is not only symbolically but actually the site of disorders that commonly give rise to chest pains. The coronary arteries, the arteries that supply the heart, are the seat of blood vessel disease that is often responsible.

In addition, one may have pain caused by general body states such as thyrotoxicosis, or overactivity of the thyroid; polycythemia, or too many red cells; and anemia, or too few red cells. Finally, pains occur in tense, nervous people for reasons that are obscure.

Formulation and critique. The skillful physician must extract from the patient a description of symptoms that explains what the underlying facts are when the patient says, "I feel sick," "I'm breathless," "I'm weak," "I have a pain," or "I have an ache," or "I'm dizzy," statements so vague in themselves as to provide little help.

Thus, a complaint of weakness may suggest disease of the central nervous system, inflammation of the peripheral nerves, chronic intoxications, undernutrition, vitamin deficiency, fatigue from overwork, or almost any chronic severe debilitating disease. Difficult breathing or shortness of breath needs to be investigated along lines similar to those used in the investigation of pain. In the complaint of swelling of the ankles, one needs to know whether it is firm, red, hot, tender, or painful; whether it leaves an indentation on pressure; tends to diminish or disappear overnight only to reaccumulate during the subsequent day's activity.

In confirmed alcoholics, the quantity and kind of drinking the patient has done is hard to determine. His own statement may vary substantially from that of his cronies or his wife. A sense of guilt or shame may diminish the apparent extent of drinking; or, in an occasional patient, a sense of bravado may lead to exaggeration.

The actual colour of the urine, the colour of feces, the actual amount of blood coughed or vomited is diagnostically important. In judging the severity of hemorrhage, the patient's impressions may be exaggerated by terror or by objects, clothing or bedclothes, that have spread a little blood a long way. The occurrence of yawning, dizziness, tendency to black out, actual fainting, sensation of the heart racing, palpitation, and drenching sweat may be more helpful in arriving at an estimate of the degree of blood loss.

Judging the chronology of disease

The physician will be helped in working out the chronology of a disease if the patient can tie in an episode or feature of an illness with a specific event, a happening, a journey on a specific date. Especially in long-drawn-out diseases, reference to old hospital records or physician's notes, or a diary, or letters may highlight a specially important event with enhanced accuracy.

In evaluating the impact of the illness on the patient's life, a physician will need to know whether the illness confined the patient to bed, kept him from work, or merely distracted him while he went through the day's usual routines. Thus a patient who complains of fearful pain but continues the routine of daily activity has presented conflicting testimony that needs to be judged. Again, the physician is aided in estimating from his character or nature. He knows, or should know, whether the patient is easily upset, exaggerates complaints, or is stoical; whether an illness or symptom can be used consciously or unconsciously by the patient to gain some objective—release from work, getting attention from an inattentive spouse, love and care from family, or sympathy from neighbours.

When a patient presents a long, complex medical history, actual records and reports may be essential to a proper understanding of the problem. The physician will need to known how many medicines are currently being used, what ones had been used in the past, what idiosyncrasies or allergies a patient has. It is essential to know whether some regimen such as a program of rest, of exercise, of weight reduction, of shock therapy, or diabetic management has important bearings on a current condition.

Past history. Although the history of a patient's past health or medical problems is more likely to be of value in understanding his problem in illnesses that are chronic, an understanding of a current acute disturbance may occasionally be possible only with the knowledge of the past. An unexplained fever may move a step toward clarification when it is learned that the patient had rheumatic fever some years previously. A pain in the abdomen may be understood only in the light of knowledge of earlier operations.

The general health of a person is not likely to be learned from a diary. Growth curves and charts of weight rarely are available, though any precise information helps. Whether or not a change in body weight is associated with the present sickness may have to be understood in the light of past averages and extremes of body weight.

There is need to know where the patient has lived, whether overseas travel or military duty may have exposed him to exotic diseases, drugs, or injury; whether he

has ever been a patient in a hospital for some disease, injury, or operation; has had examinations for military service, for work, or for insurance, and what the results of these examinations were.

Review of systems

Special diagnostic procedures and specialty problems. An important part of the general history of a patient consists of the review of systems. Here, functional rather than structural characteristics of the body may reveal a pattern. In regard to many such systems, medical specialties have arisen, each with its own separate constitution and bylaws, its strong proprietary feelings, and its tendency to isolate itself from the central core of general medicine. The following list of systems is ordered alphabetically.

Allergy and immune disorders. An array of changes, many of them not belonging to a single medical domain, are grouped as allergic and immunological disorders. Sensitivity causes lesions in the skin, edema, and disorders of mucous membranes such as hay fever. Some doctors put migraine into this group. Pollens, danders, and easily discovered sensitizing agents are well known offenders.

Bones. The growth of bones, their shape, size, deformities such as the acquired ruin of rickets; perfect healing compared with the limping awkwardness of a misset fracture; congenital anomalies; the acquired overgrowth of hypertrophic osteoarthropathy are significant in diagnosis. The troubles in and around inflamed joints, the vast array of bone weakness from inadequate minerals or protein deficiency in osteoporosis and osteomalacia, bone cysts, tumours in bone and pathologic fractures add to the list of important metabolic and orthopedic problems.

Cardiovascular system. Since every cell in the body depends on the cardiovascular system for its successful birth, function, and survival, a complete understanding of it would require, and be equal to, the comprehension of all of the functions of physical life in health and disease. Since the time of Harvey, but not before, physicians have looked upon the heart as the essential element of this system, adding scientific validation about the heart to what folk speech had already kneaded into the language. For the diagnostician, a grossly abnormal functioning of the heart may call attention to its difficulties through pain in the chest, shortness of breath, palpitation, a quick beat or a sense of the heart's racing, or difficulty in breathing at night that makes one sit up, gasp for breath, and cough. Swelling of the ankles and the abdomen may occur next, as may faintness or fainting. The legs may signal circulatory deficiency in an intermittent limp and pain if there are large arterial impediments. Raynaud's phenomenon may blanch the fingers. The veins may become clogged and inflamed, or obstructed with the secret clot of phlebothrombosis, recognized only after detachment of the clot causes death of lung tissue.

The diagnostic story of cardiovascular trouble may include heart attacks, coronary artery trouble, angina pectoris, high blood pressure, rheumatic fever, congenital heart disease, traumatic lesions of the heart as well as systemic diseases such as diphtheria, syphilis, or muscle disorders.

Endocrine system. The hormones made by the body temper and modulate the sharper controlling actions of both the voluntary and the autonomic nervous systems. Growth, reproduction, body type, weight, size, hair, skin pigment, muscle strength; libido and potency or impotence; texture, moisture or dryness of skin and hair; tolerance to heat and cold; control of thirst; gluttony, sugar in the urine, loss of appetite; extraordinary ambiguities of physical sex characteristics; the diminuendoes of aging and senility—all form a heterogeneous endocrine-governed mixture. Errors of the system produce different forms of dwarfism and gigantism. Precocious senility is contrasted with the bewildering occurrence of puberty in those scarcely out of diapers.

Gastrointestinal system. Just as physicians who specialize in ailments of the nose and throat are said to be convinced that many, if not all, patients have trouble with the upper and lower respiratory tracts, so doctors dealing with the alimentary canal may be convinced that this is, perhaps, the most fertile terrain for imperfect function and symptoms. Study of the alimentary canal includes an estimate of appetite, weight, the mechanics of eating, difficulty in swallowing, nausea, belching, flatulence, abdominal aches, pains or colic, vomiting, jaundice, and vomiting of blood. With jaundice, there may be fever, pain, or change in the colour of feces or urine.

A preoccupation of many people with alimentary complaints is called stool watching. Patients reporting their bowel movements to an examiner can give the most minute and detailed description of frequency, consistency, odour, gas, the effect and noneffect of cathartics, almost everything, including weighing the feces. The doctor must know about sphincter control, hemorrhoids, fistulas, bleeding, diarrhea, constipation, and bowel habits.

Genetics. A large group of physical and emotional difficulties may be determined by obvious chromosomal disasters or by mutations, which are getting increasing attention. They need to be known in order to be dealt with appropriately even in the afterthought of counselling. There is need to know whether any genetically determined disease is transmitted by a dominant or recessive mechanism, is a de novo sport or part of an ancient heritage.

The hematopoietic system. The bone marrow and sometimes, by a borrowed skill, the spleen and liver manufacture cellular and other constituents of circulating blood. "Anemia" is a general reference to numerous disorders with reduced hemoglobin and red cell constituents. Anemias fall into three large varieties with some crisscrossing and overlapping. A hypochromic microcytic anemia occurs when iron is not available in adequate amounts to manufacture hemoglobin, and the red cells are small and deficient in this protein, which is essential for the nearly instant transfer of oxygen and carbon dioxide. At the opposite extreme there are large cells full of hemoglobin but far too few of them. This form, called macrocytic hyperchromic anemia, results from inadequate function or supply of vitamin B_{12} and folic acid. Pernicious anemia is the classical example of this disorder. Tropical or nontropical sprue with malabsorption cause the same morphologic change. Normochromic and normocytic anemias reflect a situation in which size and shape of red cells, and perhaps the function, are adequate, but there are not enough of them. This may occur shortly after a significant hemorrhage, in situations in which red cells are destroyed faster than they should be; and in some states that inhibit bone marrow and slow down the production of red cells. Too few cells are made, but the cells are not significantly abnormal.

Types of anemia

Lymphatic tissue. The spleen and the liver may be enlarged in various tumours of lymphatic tissue (lymphomas) and leukemias. Peripheral lymph nodes become enlarged at the point at which infection invades the body. Tumours may develop. Infected lymph nodes may become swollen and painful or mat together. They may suppurate, forming abscesses and abnormal passages (fistulas). Lymph nodes play an important role in a wide range of disorders.

Muscles. The enlargement of the forearm of the traditional blacksmith, the rippling muscles of the gymnast and the weight lifter, have their opposite in various atrophies and muscle disorders that may be inborn or acquired from damage and destruction to motor nerve cells. The athlete's bruise, the stiffness of the sprain, and the harmless but miserable, sometimes wretchedly, painful nocturnal cramp of dowager, mountain climber, or sprint star may all be part of a diagnostic picture.

Nervous system. Diseases of the nervous system include those that belong to psychiatry and those that occur in the field of general neurology. Special and supplemental procedures and diagnostic procedures are required. The past history that is of concern to the neurologist includes the functions of speech and language, disorders of head and neck, the integrity and health of the cranial nerves, the whole motor system, coordination, the state of reflexes, the state of the sensory system, the spinal cord, as well as the gait, stance, and station of the

Neurological symptoms

patient. Special studies may be required for a comatose person, for children, for the very old, for those suspected of malingering, and for those who may be motivated by the hope of compensation and insurance payments. In the history of a patient with suspected neurological disease the physician needs to know about trouble with smelling, seeing, hearing, chewing, taste, equilibrium, speech, swallowing. The functions of urination, defecation, reactions to heat and cold, sweating, redness, pallor, and cyanosis are important, too.

Obstetrics and gynecology. For women such things as growth, age at menarche (the first menstrual period), regularity, duration, and frequency of menstrual periods, amount of flow, degree of comfort or discomfort, the exact date of the last normal period, occurrence of leukorrhea (whitish discharge from the vagina), unusual bleeding; premenopausal and postmenopausal events, and the menopause are all important. The whole medical and psychological story is necessary too.

The diagnostician will need to know of details of any communicable diseases, particularly German measles (rubella), the immunization record, inheritable diseases, kidney disease, diabetes, cardiac problems, and previous blood transfusions. He must know also of previous pregnancies, abortions, dates of delivery, method and ease of delivery, complications, weight of the child, condition at birth, and how the child developed.

When dealing with infertility problems, it is necessary to ascertain whether the uterine tubes are open; to make X-rays of the uterus; to examine a specimen of the uterine lining, just before the menstrual period, and to keep temperature graphs to see if ovulation can be detected; estrogen and other hormone levels must be investigated if any abnormal conditions are found.

The skin. Besides the colour of the skin, or the lack of it, the examiner needs to learn about changes in appearance that may range from the black of the black man to the noncolour of the albino; the acquired pigmentation of tanning or of Addison's disease (failure of the outer substance of the adrenal gland); tendency for the skin to be moist or dry, hot or cold, itching or comfortable, scaling or bleeding, easily bruised or weeping, wrinkled or fresh, distended by edema or shrivelled by malnutrition. Of the appendages of the skin, the most important are the hair and the nails. Also important are studies of the pattern of ridges and valleys that go to make up the fingerprint. Presence or absence of hair, its rate of growth, texture, the places it grows, and its colour must be noted. Whether it has disappeared or grows overabundantly, and where, may be critical. Generally neglected, the nails may give an intimate view of health, disease, growth, and aging. The spectacular changes of clubbing, the opposite alteration in spoon nails, pitting, colour changes, and the deformity produced by fungus infection or psoriasis make a gaudy and sometimes painful spectacle.

Urinary system. The urinary system, more likely to trouble old men than old women, may change the colour of the urine, increase excretion or reduce output, or manifest excessive urination at night and painful urination. The occurrence of pus or blood in the urine, frequency, colic from stones in the ureters or stones in the bladder, chancre, bubo (the inflammation of lymph nodes in the armpit or groin), and the urethral discharge of gonorrhea present a long trail of recurring disorders.

PHYSICAL EXAMINATION

Elements of the physical examination

The physical examination is a continuation of the exercise in diagnostic detective work. Any sensory function may be used by the physician as a legitimate means of getting information. The standard formal divisions of the physical examination are inspection, palpation, percussion, and auscultation (see below). Occasionally, smelling is of help in recognizing diabetic acidosis, uremic breath, fetor hepaticus, or lung abscess. Most physicians today doubt that old clinicians really could smell typhoid.

The physician working toward a diagnosis uses the same senses that everyone has. Because of his training, his perceptions have special meaning in terms of normal and abnormal structure and function. A working diagnosis evolves on the basis of gathering and then interpreting the accumulated data. Even people who are tone deaf can become excellent at auscultation. It takes not so much talent as long practice to get accurate information from palpation and percussion.

Inspection. As for inspection, vision is so conditioned by training that most people see and perceive what they look *for* much better than they do what they look *at*. Inspection is the sheet anchor of physical diagnosis, and the hardest procedure to do well. Even so, correct diagnosis is probably made more often by inspection than by any other part of the physical examination. One sees what he can form a mental as well as a visual picture of—in short, what he understands. Looking, everyone can do; but seeing must be learned. The yield from inspection depends on the understanding and training of the observer. Within surprisingly broad limits, it depends little upon the acuity of vision as ordinarily measured, though good eyes and a good light always help.

Inspection in a sense goes on systematically during the entire time the patient is with the physician. It begins with greeting the patient and continues through taking the history. During the actual examination, it is concentrated. Everyone recognizes and identifies persons at a distance by their walk, their size, their shape, or their clothes. Their voice and manner may give further clues. Thus, one "diagnoses" the presence of a friend, an acquaintance, a relative, or a stranger. A layman seeing a person with heart failure might or might not know the significance of the bluish skin, the puffy ankles, the shortness of breath, though he would recognize that the person was in trouble.

In very gross abnormalities a quick diagnosis may be correct. Usually such a diagnosis results in fact from the compression in time of a multitude of steps that the diagnostician has taken unconsciously. There is no magic or sixth sense.

The importance of the examiner's concentrating his whole attention in diagnosis, both in controlling and in directing his thoughts, is illustrated by the experience almost everyone has had of looking at a watch to satisfy some particular need to know the time, only to realize on being asked what time it is that he has to look again because he was thinking not so much of actual time but rather in terms of some engagement, appointment, or plan.

Examination of the body as a whole

Inspection of the body as a whole will reveal at once size, shape, skin colour, approximate age, gross deformity, sex, spontaneous motor activity, body build, and state of nutrition. Something is learned of manner, mood, posture, speech, and the subtle kaleidoscopic composite that helps tell whether the patient is sick with organs disabled or is anxious, worried, in pain, or in trouble, and, if so, to what degree. The symptoms may be just as real and as severe when the function of healthy organs is disordered as when they are physically diseased.

A region of the body may be inspected minutely or generally, for the body gives rise to hundreds of normal or abnormal physical signs. The initial and sometimes the only necessary diagnosis in skin diseases may be made by inspection. The tables of contents listing the illustrations in dermatological and in surgical textbooks of diagnosis give an indication of the importance of seeing to the understanding of such manifestations. Diagnostic inspection goes beneath the skin. Endoscopes, devices for viewing internal structures, may disclose a whole realm of internal structures from one end of the body to the other. Specialized instruments are available for viewing the interior of the eye, the nose, the ear, and so on. Beyond the eyes' unaided power, the microscope enables one to see tissues, cells, crystals, and so on. The X-ray discloses function and structure with fluoroscopic screen and films. The X-ray film may be in the form of a motion picture, a series taken in rapid sequence, or as the representation of a single moment.

Palpation. Palpation is investigation by means of tactile senses. Such senses may provide all kinds of information including that of touch, temperature, vibration,

shock, crepitation (grating) of a joint, vibration of moving fluid in an irregular conduit, and sense of position, location, or place. For the doctor, practice and experience add a keenness to the native sensory capacity, as he acquires a special skill.

The tips of the fingers are the most sensitive parts of the body for touch, as would be expected because of their constant use in exploration of the reachable environment. In growing children palpation appears better than vision in orientation, at least early in life. It includes not only examining and feeling objects but putting them in the mouth. Temperature is perceived most readily by the backs of the fingers and the back or the side of the hand where the skin is thin. Vibrations and thrills (tangible vibrations or pulsations) are felt best in the region of the palmar calluses where the fingers join the hand.

The tactile sense is used to some extent in testing strength of grip. The ideas imparted about the shape, consistency, and the mobility of an abdominal organ or mass are perceived by the simultaneous fusion of several aspects of touch.

Importance of the sense of touch

Although palpation may be thought of mainly as the art of acquiring information about the external structures of the body, many internal organs can be felt indirectly, both in health and in disease. The sense of touch may be vital in finding abnormalities in the mouth, tongue, region of the tonsils, thyroid gland, rectal and pelvic structures, as well as in the external genitalia, skin, and hair. Abnormal masses or enlarged organs may be hard, soft, moveable, firm, or quiet; may pulsate, or crackle, as in the crepitus of joints. The crinkly, squashy feel of tissues into which air has escaped under skin or along fascial planes (sheets of fibrous tissue) where it does not belong is unpleasant. Thrills over turbulent flowing blood in aneurysms (local dilations of arteries with thinning of the vessel wall), arteries, or in the heart challenge the skill of the palpating hand. Useful information is imparted by the creaking, rubbing, leathery sensation over a point at which there is friction between two surfaces of the pleura, the membrane that encloses the lungs and lines the chest.

Surgeons have a special opportunity to become skilled in palpation. First they feel organs and masses through guarding muscle or intervening tissue; then at operation they feel the structures directly. This instructing, and at times correcting, experience gives the conscientious surgeon an intimate knowledge of lesions, tissues, and organs that other physicians may touch only at autopsy.

Percussion. Percussion is used in physical diagnosis to obtain information about what lies below the surface of the body, by striking it, setting up vibrations, and listening to the resulting sounds. It is a primitive form of echolocation. In addition, the practiced clinician obtains information in percussion from the sense of touch.

While, except when examining internal passages and structures, inspection is in the class of nonintervention, the manipulations of palpation and percussion represent an intrusion or an intervention, minor to be sure, in which the passive observational aspect of the diagnostic quest is replaced by a simple, gentle experiment. Auscultation (see below), as it is practiced with the stethoscope, also has an element of intervention as the physician puts the stethoscope in appropriate places and may instruct the patient about breathing or holding the breath or changing position.

Direct and indirect percussion

Percussion may be direct or indirect. When it is direct, the tips of the fingers strike the chest. When a finger is interposed, the percussion is indirect and is called bimanual or mediate percussion. The finger that is placed upon the chest or abdomen is called the pleximeter. Only the terminal phalanx is applied, since the other fingers may damp out vibrations and sounds. The striking finger is called the plexor. Another form of percussion is done with the middle finger while the other fingers and thumb of the same hand are pressed firmly on the surface.

When percussion is desired, the examination begins with the patient lying on his back. After the front of the chest has been percussed in this position, the patient sits up and the borders of the lung, particularly the bases, are estimated and the outline of cardiac dullness is obtained. Very ill patients remain in the recumbent position, and the front and back are examined with the patient lying on one side or the other.

Auscultation. Auscultation means listening and hearing. The body generates many noises, most of which come from air movement. A patient's breathing, voice, and manner of speaking, as well as what he has to say, are important. Coughing, sneezing, belching, the expulsion of flatus, the embarrassing sounds of the movement of air through the bowel within the abdomen are familiar. The cry, the snore, the groan, the shriek, the snort, the stertorous breathing (which may occur in heart failure or in brain disorders), regular–irregular stop and go breathing patterns all give diagnostic clues. But for practical purposes auscultation concerns what may be heard by the appropriate use of the stethoscope.

Development of the stethoscope

The stethoscope, beginning as a tightly rolled cylinder of paper first used by René-Théophile-Hyacinthe Laënnec, next became a solid cylinder, then a hollow cylinder, and finally a device with tips that fit snugly into the two ears. The stethoscope has even been modified in the symballophone, in which two terminal cups, or bells, simultaneously capture sounds from two different points, thus permitting an approach to three-dimensional localization of sounds.

Traditionally, the chief terrain of auscultation is the chest, to hear sounds generated by the heart and lungs. A broader understanding of disease and a great new variety of technique have broadened the uses and emphasized the value of auscultation.

The use of sound in diagnosis is complicated by the differences in sounds from person to person. The capacity of the listener to hear sounds is related in part to his perspective capacity, though, significant deafness aside, it is rare to find a person who cannot master the difficult art and science of auscultation.

The stethoscope employed today consists of two ear pieces, which fit snugly into the external ear, rubber tubing, and a chest piece. The tubing connects the ear pieces to the chest piece, which ordinarily includes a hollow bell and a wider flat cup. The cone of the bell, which transmits all sounds but is best for sounds of low pitch, has a rim of hard rubber or plastic around it. The wider flat cup is covered with a thin semirigid diaphragm of plastic to exclude low-pitched sounds so that isolated high-pitched sounds may be heard. Different forms of stethoscopes offer various advantages and disadvantages. The snugness of the fit in the ear is important for concentrating the sounds the doctor wishes to hear and excluding others.

Bakelite, or other plastics with proper acoustical properties, are necessary for the cup. When the diaphragm is used, it is pressed snugly against the surface. The bell is applied lightly but with an all-around contact to keep out extraneous sounds. If it is pressed lightly, the stretched skin acts as a diaphragm and blocks out low-pitched sounds.

Examination of the chest. When air flows at various rates through tubes of varying calibre, shape, or angulation, it is thrown into turbulence, setting up vibrations. The purpose of auscultation is to separate normal from abnormal sounds of breathing, after coughing, and by the transmission of the spoken or whispered voice through the lungs to the surface of the chest. When sounds are not heard, the physician assumes that part of the tube is blocked, that air is not reaching its proper destination, or that something is interposed between the air-containing lung and the surface at which the stethoscope is applied. This could be fluid, a solid mass of tissue, or air free in the chest cavity.

Types of breathing

Ordinary breathing produces breath sounds, called vesicular breath sounds, which are longer in the inspiratory phase than in the expiratory phase. They can be heard over the lungs except where structures between the two lungs are in the way. If the stethoscope is placed over the trachea, bronchial breathing, breathing with a longer, louder expiratory phase and a short inspiratory phase is heard. When this breathing is heard over other parts of the lung, it means that sounds generated in the trachea

are being transmitted through consolidated (solidified) or compressed pulmonary tissue, as in pneumonia. Intermediate between bronchial and vesicular breathing is bronchovesicular breath, which usually indicates pulmonary consolidation of less extent than when bronchial breathing is heard.

The sound and time relationships may be changed completely in the breath sounds heard during an attack of bronchial asthma. The expiratory phase may be many times longer than the short gasping inspiratory phase. The pitch is high, sometimes amounting almost to a squeak. The sounds of this sort of breathing, called amphoric, resemble the noise produced by blowing over the mouth of an empty bottle.

Ordinary breathing does not produce loud sounds. Much louder ones are produced when a person whispers or speaks. Whispered sounds usually give better discrimination because they are not too loud. Consolidation increases the loudness of the sounds from the spoken or whispered voice. Egophony, the sound resembling the bleating of a goat, may be heard over areas of pulmonary compression, collapse, or consolidation.

Rales

In addition to sounds produced by breathing or phonation the examiner may hear sounds called rales, or little rattles, which come from the turbulence set up by air going through constricted or distorted parts of the airways or over fluid or thicker secretions that do not completely plug the tube. The classification of rales is chaotic. The major designations concern whether the rales are moist or dry, the dry ones actually being produced by moist but viscid secretions. Large gurgles or rhonchi occur when the noise is generated in the trachea or larger bronchi. They occur especially in persons in coma. Rales occurring when death is approaching have been called the "death rattle."

Less impressive rales, called medium or crepitant, arise from the disturbance in the flow of air and vibration set up in smaller branches of the bronchial tree. They sound either like clicks or bubbles. Fine moist rales may be also called crackling or subcrepitant. They probably arise from air moving over fluid in the very small bronchioles. Dry rales have a musical or a sonorous quality, depending on whether they are high pitched or low pitched.

Friction rubs

The diagnostician may hear friction rubs in pleurisy, injury, pulmonary infarction, or pneumonia. They are produced because the fibrinous surface of the pleura—membrane—covering the chest wall and that covering the lung are dry or sticky, and the movement between them, ordinarily gliding, is interrupted. The noise is said to resemble that made by rubbing two pieces of leather together. With free air and fluid in the pleural space, movement may cause a succussion splash, a very unpleasant sound.

A great variety of sounds may be heard over the heart in pericarditis, inflammation of the fibrous sac that encloses the heart. Even more astonishing ones are heard in the crunchy, crackly loud noises generated when air escapes from its normal tubular pathways in the lungs into the lung's free supporting tissues and fascial planes.

Information gained by auscultation must be considered in conjunction with what may be found by palpation and percussion. Percussion may disclose the hyperresonance of a pneumothorax, air free in the chest cavity, or the flat note of consolidation, fluid, or interposed tissue. Inspection may reveal bulges or an asymmetry in the respiratory movements of the two sides of the chest.

Variations in the amount of subcutaneous fat, in the thickness and muscularity of the chest wall, in the capacity of the patient to follow instructions in breathing, in the state of vigour and health of the patient, all introduce differences in sound produced that can be appreciated only with painstaking and prolonged practice and experience.

Auscultation of the heart. Normal sounds in the cardiac cycle are as follows: first, sudden contraction of the ventricles (the lower chambers of the heart) closes the atrio-ventricular valves, between the lower and the upper chambers, and opens the valves into the aorta and the pulmonary artery, transmitting vibrations to the chest wall. This first heart sound is brief. The second heart sound occurs when the aortic and pulmonic valves snap closed at the beginning of diastole, the interval between the pumpings of blood out into the general circulation and to the lungs. If the right and left sides of the heart do not beat simultaneously, the cause may be abnormally high pressure in the pulmonary artery, narrowing of the valve into the pulmonary artery, or blocked conduction in the right bundle branch, a collection of modified heart muscle fibres that normally plays a part in conduction of impulses from the pacemaker of the heart, the sinoatrial node (see CARDIOVASCULAR SYSTEM, HUMAN).

A third and a fourth heart sound may be perceived. The third represents the ordinarily quiet opening of the aortic and pulmonary valves as blood leaves the ventricles, the lower chambers of the heart. The fourth, or atrial, heart sound is generally pathologic and represents overactivity of the atrial contraction or abnormality in the ventricle at the time it receives blood.

The use of the stethoscope enables the physician to make a tolerably accurate judgment about cardiac rhythm and irregularities, as well as an accurate count of the rate. In atrial fibrillation, the heart beats fast and irregularly and there may not be a pulse for each heartbeat. If the pause between beats has been short the amount of blood ejected by the heart is too small to propagate a pulse wave.

Counting the heart beat

The ordinary heartbeat is characterized by the two sounds that repeat themselves continually. With irregularity, it is necessary to count the heart rate rather than feel the pulse. The difference between beats recorded over the heart and at the wrist is called the pulse deficit.

After the rate per minute has been determined, the rhythm is studied. It may be entirely regular. There may be a slight slowing and speeding related to phases of respiration, an irregularity called sinus arrhythmia.

To the diagnostician anything that is regular at extremely rapid rates suggests a disorder of rhythm called atrial flutter. Regular impulses are generated in the pacemaker in the atrial wall at approximately 230–300 beats a minute, but protracted ventricular response at the rate of 200–300 beats a minute is incompatible with life. There may be a 2:1, 3:1, or 4:1 block, which means that only every other, every third, or every fourth atrial beat evolves a ventricular response.

The heart rate may be slow. When it is exceptionally slow, and regular, there may be a 2:1, 3:1, or 4:1 block. Very fast rates may occur with atrial fibrillation, atrial flutter, a simple rapid heartbeat (sinus tachycardia), or paroxysmal atrial flutter. Ventricular rates between 150 and 230 seem regular. In ventricular tachycardia, rates may vary from 150 to more than 250 per minute. Irregularity of rhythm causes changes in intensity of sounds because of variation in filling of the ventricles. Irregularities of rhythm may be erratic as with premature atrial or ventricular beats. Atrial fibrillation causes complete irregularity of ventricular activity.

Gallop rhythms are so named because of their resemblance to the rhythmic hoof beats of a galloping horse. The gallop of clinical significance heard early in diastole suggests an overfilled ventricle that is failing in its function as a pump.

Cardiac murmurs

All sounds generated from the heart result from vibrations being set in action. When blood flow is smooth, turbulence does not occur, and vibrations are not imparted to heart or vessel wall. Vibrations heard as murmurs are increased if blood is of low viscosity as in anemia, when the rate of the flow is rapid, or when the calibre of the vessel undergoes an abrupt change. The doctor can identify the time at which the murmurs occur with respect to the action of the heart, whether they are early, intermediate, or late with respect to the process of pumping blood into the system and the lungs (systole), or in the periods between pumping (diastole).

Murmurs, once identified in time, can be divided into those produced by organic lesions and those produced by abnormal function as in anemia or when the heart is dilated. Short systolic murmurs confined to early, middle,

or late phase systole are likely to be benign. All diastolic murmurs have pathologic significance.

Murmurs have pitch and quality, or timbre. This the physician recognizes as a musician identifies the same note produced by bass viol, violin, drum, or flute. Murmurs produced when the pressure is low are low-pitched; those coming from high pressure and narrow orifices are high-pitched. A cardiac murmur may be loud or faint, and loudness is not necessarily significant.

Whenever he hears an extraneous sound, the examiner must decide whether it is a murmur or not, whether it is generated in the heart or is outside it, as in a pericardial or pleural-pericardial friction rub. Ordinary heart sounds begin and end abruptly. Murmurs are less likely to do so. The bell and the diaphragm must be used to hear high- and low-pitched murmurs. The stethoscope must be moved to locate the region of greatest intensity that suggests that a particular valve is involved. The timing of the murmur, whether in early or late systole or diastole, must be determined.

Loudness is usually recorded on a scale of six. One is of small intensity and six is audible without a stethoscope away from the chest. A murmur may be loudest at the beginning, at the end, in the middle. The type of murmur depends on whether it is low-pitched and heard best with the diaphragm. Very low pitched noises may come from friction rubs and have puzzled even experienced clinicians.

Effect of position on heart murmur

Heart sounds and murmurs may be influenced by position—whether the patient is lying on his back, on his left side, is sitting upright, or is leaning forward. The phase of respiration is important, too, since the base of the heart is brought closer to the chest wall with full and held expiration. The presystolic murmur of mitral stenosis, the narrowing of the opening between the left atrium and ventricle (upper and lower chamber), may be heard in a small area while the patient is lying on his left side. Leaning forward may increase the loudness of the murmur of aortic stenosis, narrowing of the opening into the aorta. Exercise with faster flow of blood may bring out a murmur otherwise silent or increase the noise of a murmur.

Smell. While smell seems the least dignified of all the senses used in physical diagnosis and is rarely of critical value, its cultivation may be helpful in obscure problems and may solve a few.

The smell of the highly perfumed, pomaded, and pomandered male patient; the outdoors or barnyard smell of an unbathed worker in the fields; the rancid and fetid personal stalactites and stalagmites of the unkempt vagrant—hobo, hippie, or prisoner of war—are intrusive. In the more strictly medical application of the sense of smell there is much to learn after passing through the atmosphere of body odours. On the breath can be detected such things as the characteristic odour of the diabetic in coma; of someone with long-standing untreated kidney disease; the marshy, musty, swampy odour associated with liver disease; the fragrant, rancid, or vomity, alcoholic odour of the acutely ill or nearly dead drunkard.

A lung abscess may produce a powerful, sometimes nauseating odour, which is particularly distressing to the victim of the disease as well as to all within smelling range. The sputum may have a putrid odour, nauseating with its sickeningly sweet stench.

Poisons and fistulas

The breath may carry the odour of poisons, as may vomitus, which may smell of alcohol, lysol, and aromatic poisons. Fermenting or decayed food from a diverticulum or pouch in the esophagus (the passageway from throat to stomach) or a suddenly emptying pouch of intestine may have a bad odour. Though a fecal odour to vomitus may come from an abnormal opening between colon and stomach, so-called fecal vomiting may occur with peritonitis (inflammation of the membrane lining the abdomen or covering its organs) or may result from intestinal obstruction.

For thousands of years the body and excreta gave their smell to human habitation. The smell of feces is not generally diagnostic. It is particularly unpleasant in some forms of malabsorption in which fats become rancid in the alimentary canal.

The odour of urine may be ammoniacal or nitrogenous. Sometimes it smells of the product of other forms of decomposition, usually microbial in origin, with yeasts, fungi, and bacteria doing their part. Pus may have a fecal odour when produced by protein-attacking bacteria from the gut. A sickeningly sweet smell may be associated with gas formation in gas gangrene. Smell, on the whole, is not a critical or often a valuable source of help in the diagnostic examination but may give an initial clue that suggests clinical changes.

Miscellaneous examinations. *Emergency physical examination.* An emergency may fall into a wide variety of categories. No specialty in medicine is without emergencies. The degree of emergency varies quantitatively as well as in terms of the kind of difficulty. The field of trauma has always been illustrated most vividly in warfare, where the form, variety, and severity of injury depend upon the sophistication of methods of destruction.

Outside the field of physical trauma, emergencies may be classified as arising in the natural history of a disease. This may be illustrated by massive internal or external bleeding, stroke, a ruptured intestine, or the anguish produced by efforts at the passage of a stone in ureter or bile duct. Intoxications from overwhelming poisoning, assault by radioactivity from a disaster in an atomic energy plant, or the multitudinous accidental or intentional poisonings make a long list. In fields of special concern, the anaphylactic shock of a bee sting may be promptly fatal. A myocardial infarction, death of a section of heart muscle, may be lethal instantly or within moments. The end stage of chronic undernutrition may present a lethal emergency as death approaches. In the field of nose and throat disorder, impaction of food caught in the windpipe or the clot formations in veins within the skull in connection with infections of the mastoid or petrous portion of the temporal bone may be critical emergencies. In endocrinology, though emergencies are much less common, severe insulin reactions or diabetic coma are examples.

Foreign bodies in the eye

As for the eye, the acute foreign body injury, lacerations, burns, detached retina, or the tight eyeball crisis in glaucoma are serious emergencies. Besides bleeding of the alimentary canal, rupture of gullet, stomach, or other portions, peritonitis from appendicitis, and severe bleeding constitute examples. Some infections, stones, and obstructions make serious emergencies in the genitourinary system. In women, serious menstrual bleeding, pregnancies outside the uterus, and complications of labour are among the grave emergencies. In the field of infection, emergencies usually represent acute and overwhelming sepsis—poisoning from bacteria and their products—now that yellow fever, typhus, and plague are checkreined. A systemic *Meningococcus* infection may be fatal within 12 hours of the first sign or symptom. Rabies, tetanus, diphtheria, and poliomyelitis are emergencies. The nervous system has cerebral hemorrhage, clots, and meningitis (inflammation of the brain covering). As for psychiatric problems, profound depression with suicidal impulse or act, hysteria, and mania produce difficult emergency problems. The first sign of alcoholism that is noted may be a ruptured varix (dilated vein) of the esophagus, delirium tremens, or maniacal drunkenness.

Burns, shock, and poisoning

Severe burns, decompression illness, electric shock, heat stroke, and sometimes dehydration exhaustion from heat may be critical emergency problems from environmental trauma. The doctor must consider poisonings of all sorts.

The emergency patient may be encountered in combat, at the site of an automobile accident, in the home, in an ambulance, in a mobile coronary care unit, or in the emergency rooms of a hospital. The skill and equanimity of the physician in such cases are sorely tested, for he needs to make plans accurately and at once. He must know when instant action is required and when thoroughgoing treatment should be delayed.

For the severely injured patient, the diagnostician follows a sequence of priority that brings immediate attention to life-threatening injuries and leaves for later attention and definitive care less serious troubles. The

restitution of tissue continuity, the setting of fractures, the general support of cardiorespiratory activity, the appropriate use of antibiotics, and the comprehensive instant alerts of an intensive-care unit may mean the difference between survival and death.

In the cases of severe internal bleeding, it is important to know the exact site of bleeding, but transfusion and other supportive measures gain time for further studies. Though surgeons are uneasy about operating for bleeding from an unknown site in the alimentary canal, it is sometimes necessary if the bleeding goes on faster than transfusion can keep up safe blood volume.

Chronic and acute poisoning

Toxicological examination. A poisoned person may be a victim of homicidal or suicidal intent, accident, or assault from the environment or his occupation. He may be acutely overwhelmed or slowly stifled from chronic exposure. In chronic poisonings, detective work of all sorts is necessary with a thorough toxicological study of the home and work circumstances, the situation in which poisoning occurred, and the probable reasons for it. In acute poisoning, especially in a comatose person who can give no history, the doctor must proceed on the basis of what is analogous to the enlightened practice of veterinary medicine. The odour of the breath or vomitus, the identification of pills, capsules, or other recognizable material in the vomitus or stomach washings, the degree of diminution of consciousness, alterations in respiration, circulation, blood pressure, skin colour, temperature, all may be helpful. Twitchings, focal neurological changes, convulsions give important clues. When he suspects or is reasonably certain that a poison has been taken, the doctor will order the stomach to be washed out. Chemical studies of blood and body fluids, minerals, and gases may give diagnostic clues as to the nature of the intoxication. Many poisons can be tested for directly in vomitus, stool, urine, or blood. Any clues from recent history may be helpful—children climbing up to explore medicine cabinets, a telltale empty bottle—and real suicidal attempts must be distinguished from dramatic gestures in which the need and desire for attention and concern may evoke the demonstration.

Diagnostic examination in obstetrics and gynecology. The history of an obstetrical patient includes previous pregnancies and methods of delivery, complications, weight of child, condition at birth, subsequent development, abortions, miscarriages, date of last menstrual period, history of communicable diseases, especially rubella (German measles), and previous immunizations. Emphasis is given to the past history of blood transfusions, kidney disease, diabetes, and heart disease. In the examination, pelvic size and shape are recorded. Complete blood and urine studies are done, and tests for rubella antibody.

On later occasions blood pressure, the presence or absence of accumulations of fluid in the tissues, weight, and urinalysis for albumin and sugar are noted. The size of fetus, its position, and heart rate are recorded.

The main concern in the gynecological patient is menses: time of onset, interval, duration, flow, pain, last menstrual period, intermenstrual bleeding, leukorrhea (a white discharge from the vagina), recent cell smears, contraceptive techniques, previous pregnancies, abortions, operations, and children. The initial pelvic examination includes investigation of the external genitalia, the support of the outlet, and estimates, based on bimanual examinations, of the size, position, and mobility of the uterus and associated structures. The vagina and cervix are inspected.

Neurological examinations. Neurological examinations complement the general medical and psychiatric examinations. They are divided into five large headings: (1) history; (2) mental status; (3) neurological examination proper; (4) special situations; and (5) supplemental procedures as indicated. The mental status includes general appearance, behaviour, state and degree of consciousness, stream of thought, spontaneous speech, hallucinations, mood, affect, and functions of the intellect.

The neurological examination proper inquires into speech and language; functions of the head and neck; the cranial nerves; the entire motor system, coordination, reflexes; the sensory system; the spinal column; and station and gait. Examination of comatose patients, infants, children, the elderly and senile, the person suspected of malingering, and the person with compensation or insurance ambitions requires special testing. Special diagnostic procedures include complete examination of cerebrospinal fluid, X-rays of the skull, the brain, the cerebral blood vessels, and recording of the brain-wave patterns by electroencephalography. Other procedures include psychological and personality tests and examination of specimens of muscle and nerve.

Clues to mental status

Psychiatric diagnosis examination. Experienced diagnosticians get much information, formally and informally, about mental status. The look of the patient, his age, grooming, posture, facial expressions, manner, and attitude; the spontaneous stream of his talk, its rate, form, and quantity; his psychomotor activity are all noted. Affectively, the whole emotional life is judged from observations and questions: whether the patient is composed, complacent, irritable, happy, angry, elated, suspicious, boastful, self-satisfied, expansive, distant, aloof, indifferent, dissociated, perplexed, fearful, tense, tormented, desperate; what the nature of the thought content may be; whether his attention flits about or he is phlegmatic; whether there are somatic expressions such as perspiration, flushing, fast heartbeat, tears, tics, grimaces, and moist palms; and whether his association and thinking are logical or autistic—given to self-centred fantasies and day dreams.

Feelings of unreality and depersonalization must be inquired into delicately. The doctor then must study carefully phobias, obsessions, compulsions, expansive moods, illusions, and hallucinations of various sorts. The patient's mental grasp and capacity and his orientation as to time, place, and person must be evaluated. His memory may be judged by the history of remote and recent events; his retention and recall of digits; his counting and calculation, simple multiplication, addition, subtraction, or division; his level of awareness; his intelligence; his judgment, as in terms of how he manages his business and family affairs. The doctor must know whether the patient's understanding is real or merely verbal; what plans he has for the future; what his insight is; whether he senses difficulty and is aware of real trouble.

Urological examination. In practice the most important part of the urological examination is the intravenous pyelogram (X-ray of the kidney and ureter after injection of a contrast medium into a vein). It may give an estimate of kidney size and function, the renal pelvis (the main cavity of the kidney), the ureters, urinary bladder, and urethra. The examination of the male is concerned with the glans, foreskin, body of the penis, scrotum, testicles, seminal vesicles, vas deferens (the seminal vesicle and the vas deferens join to form the ejaculatory duct), and a digital examination of the prostate through the rectum. Urethral discharge, swelling, redness, and any abnormality of external genitalia may disclose orchitis, inflammation of the testes, of the epididymis (a structure back of the testes), or of the prostate (a gland encircling the urethra). Massage of the prostate and examination of prostatic secretions are important. Detailed examination of the urine may include cultures. With the instrument called an endoscope, direct inspection of the lining of the bladder is possible. Examination of semen along with study of sex hormones and thyroid function may be needed in cases of sterility and impotence.

II. Biological, chemical, and radiological tests

LABORATORY DIAGNOSIS

Tasting a diabetic's urine and the detection of its sweetness may have been the first laboratory test. Out of the fields of physiological chemistry and biochemistry diverse laboratory tests have arisen. They are an extension of the physical examination and are increasingly a part of the professional consultations, which means that certain salient features about the patient are made known to those doing the tests. Tests are done on blood, urine, cerebrospinal fluid, feces, sputum, gastric contents, se-

men, and abnormal fluids obtained from the chest, the pericardium (the sac enclosing the heart), or the peritoneal (abdominal) cavity. Hair, nails, and skin scraping may also be tested. The usefulness of a test has a clear relationship to the speed with which the results are made available.

Routine and special tests

One of the critical problems of diagnosis today is determining what should be looked upon as a routine screening laboratory test and what should be a special test. Certain studies of blood, urine, and feces have become accepted as routine. Multiphasic screening, testing for several disorders simultaneously with a number of tests, is now broadening this base substantially. After whatever screening procedures have been done, the next stage in the examination is a critical planning based on all the information obtained. Availability, cost, physical and chemical principles, methods, error, and other limitations are related to final interpretation and application of the results.

Blood. A well equipped clinical laboratory is able to examine many constituents of blood. Organic chemicals normally occurring in the blood include glucose, acetone, lactic acid, pyruvic acid, uric acid, creatine, creatinine, and nitrogen in the form of urea nitrogen, nonprotein nitrogen, amino acid nitrogen, and ammonia. Minerals are tested to find out whether the concentration is high, low, or normal. These include sodium, chloride, calcium, phosphorus, potassium, carbonates, copper, lithium, magnesium, lead, and iron. Hormones and enzymes may be measured. Blood gases that can be measured include oxygen, carbon dioxide, and carbon monoxide. In addition to the usual blood proteins, albumin and globulin, ceruloplasmin (a globulin that transports most of the plasma copper) and cryoglobulins (proteins that precipitate or gel when cooled below body temperature) may be tested.

Drugs in the blood

Important drugs, used therapeutically or found in toxic levels from accident or attempted suicide, include barbiturates, bromide, ethyl alcohol, methyl alcohol, salicylate, and sulfonamide. Blood lipids are measured. Routine studies among the vitamins may determine levels of ascorbic acid, vitamin A, and the carotenoids, the precursors of vitamin A. Protein changes in the blood related to liver disease are measured. Testing bile pigments separates bilirubin into the varieties called direct reacting and indirect reacting, a test useful in classifying jaundice when it is present. Other properties of blood that may be tested include blood volume, degree of alkalinity, tolerance tests of various sorts, and the clearance rate for normal materials such as urea, creatinine, or a dye such as sulfobromophthalein.

Urine. The urine is examined by inspection as to its colour, which may be light or very dark from jaundice, breakdown of hemoglobin, or the dispersion of red cells through it. It may contain gravel, clots, pus, micro-organisms, and crystals. Most significant are casts of the urinary tubules (long microscopic tubes forming part of the functioning unit of the kidney, the nephron). The specific gravity is important as a measure of the concentration of the urine. The volume of urine is important also, particularly when there is any impairment of kidney function or any obstruction. Among chemicals that may be tested for in the urine are the sugars, acetone (suggestive of diabetes), various nitrogen compounds, hemoglobin and myoglobin (oxygen-carrying proteins of the blood and muscle), homogentisic acid (present in the hereditary disease alkaptonuria), coproporphyrins (suggestive of lead poisoning), porphobilinogen (present in the metabolic disease porphyria), and urea and uric acid. Enzymes and hormones are tested for. A great variety of drugs can be detected in the urine; these are important in cases of overdosage, sensitivity, or suicidal attempts. Urobilinogen is the important bile derivative tested for chemically. Its high secretion suggests liver disorders; its absence, bile duct obstruction. Phenolsulfonphthalein concentration in the urine after intravenous injection is a good test of kidney function. Among the minerals tested for, calcium, phosphorus, copper, and lead are perhaps the most important.

Stool. Examination of feces includes a description of the form; the colour; whether the consistency is normal, diarrheal, or oily; and the occurrence of mucus or blood. The bulk of the feces averages 100 to 200 grams a day, of which about 30% is solid material and the rest water. The stools normally contain little protein. Up to 30% of the dry weight may consist of fat, 5% of which is fatty acid combined as soap. Free fatty acid usually runs about 5.5%, neutral fat about 40% of the total fat. Ordinarily, less than two grams of nitrogen are lost in the stool daily. The normal stool urobilinogen ranges from 40 to 280 milligrams per 24 hours.

Other tests. Gastric (stomach) juice is examined primarily for hydrochloric acid levels and the quantity of digestive enzymes (*e.g.*, pepsin).

Studies of the cerebrospinal fluid include investigation of its appearance (it should not be cloudy); the number of cells present (abnormal increase suggests infection); and its protein levels (increased in a number of diseases). The fluid is also tested for syphilis.

A wide range of tests assess the functioning of a number of organs, including the heart, the lungs, the kidneys, the liver, the pancreas, and the thyroid.

MULTIPHASIC HEALTH SCREENING

It has been an accepted tenet of medical thought over the years that early detection would be beneficial in controlling and diminishing disease in the community, as well as in helping individual sick persons. Multiphasic health screening has been suggested as one way to detect cellular and biochemical changes that have not yet advanced to the stage of producing symptoms impressive enough to cause the patient to seek medical aid. The people affected by any health scheme include the well and the ill who have no symptoms, and the well and the ill who do.

Multiphasic health screening is a health examination for a wide range of variations and given large numbers of people. It has as its purpose the sorting out of the ill and the well, with and without symptoms. Most important is the detection of established disease and the prevention of its progress by finding it early. At the present time it is customary for the tests, examinations, questionnaires, and other procedures of mass screening to be applied or administered by technically trained assistants. Then the physician reviews all the data to make plans for further diagnostic and therapeutic moves.

In the period between the two world wars, and gaining momentum after World War II, screening programs were set up by public health organizations for the detection of specific diseases. Skin tests and massive X-ray surveys were done, for instance, to detect tuberculosis, and tests of large populations were done for syphilis. Extensive examinations have been conducted for admission into the armed services. These have detected, in various categories, an appalling degree of inadequacy, asymptomatic illness, or gross dysfunction about which nothing had been done by those examined.

The next step in multiphasic screening came with the development of equipment such as vacuum tubes for obtaining blood that permitted the introduction of multiphasic health screening in various demonstration centres throughout the country. Administrative complexity and lack of experience led to the accumulations of piles of paper and miscellaneous data. The high cost of discovering simple medical problems, or anyone who was seriously sick but salvageable, was discouraging. The result was substantial reduction of support for such pilot studies.

Introduction of automation

By the late 1950s, equipment was available to automate the chemical analysis of blood. When, nearly a decade later, such apparatus began to be tied into electronic data processing, the third chapter of the story began to unfold as automated multiphasic health testing linked computer technology and sundry automated equipment with clinical tests and procedures. Though this was hailed as the new salvation, few such combinations of systems give a comprehensive overview of health or illness. The ultimate objective of automated multiphasic health testing is to establish a total health profile of the person to aid his physician in the diagnosis of disease and its management.

The selection of screening-test procedures and the way in which electronic data processing is used vary widely among the different systems now in operation. The ground covered is much the same in most currently used batteries of screening tests. A representative one would certainly include information from the medical history; tests of visual acuity, body measurements, hearing tests, measurements of air breathed in and out, and measurement of pressure within the eye; X-ray of the chest; electrocardiogram (record of the electrical impulses from the heart muscle); determination of blood pressure; study of cells from the cervix to check for cancer; blood cell count; hematocrit (determination of the proportions of the blood volume accounted for by the red cells); and a variety of blood tests including tests for glucose, cholesterol, urea, sodium, and potassium. Other tests could easily be added. The procedures are usually partially or fully automated, and the emphasis is on using less highly trained people to do the tests and collect the data. Many currently operating facilities process data electronically. The report may consist of many pages or a one-page summary.

Such systems become more efficient as they approach the mass production practices of industry. The unit cost is reduced by increasing the number of patients processed. This is true whether the persons cared for constitute a prepayment group, physician referrals, self referrals, patient referrals, employees of a specific company, members of a labour union, identifiable members of an academic community or a geographical region, or mixtures of these.

The proponents of multiphasic health screening maintain that properly trained technicians, using automated equipment, can provide an adequate substitute for the traditional periodic health examination by the physician, with expenditure of only one-third the time.

Criticisms of multiphasic screening

Fears have been expressed that the mass of irrelevant data through which the physician will have to wade to find the initial clues, and the subsequent multiple diagnostic studies, will add to, rather than subtract from, the present demands for work. There is not yet enough evidence to know whether the machine's reduction of human error really provides better quality control or whether the false-positive or false-negative results that do occur will lead to mismanagement of the patient's problems. There is an as yet unresolved conflict between those who boast of the efficiency—getting everything done at once and in one place, instead of the seemingly endless procession of patients around, to, and from hospitals, clinics, offices, and laboratories—and those who fear that any kind of conveyor-belt machinery will further depersonalize the already strained patient–physician relations. In the matter of costs, multiphasic health screening as currently available costs approximately a third to a fourth of the cost of the same care in the traditional system. The advantages or disadvantages for individual physicians have scarcely been evaluated in any objective terms. Both advocates and opponents of the system have agreed that it has not been demonstrated scientifically that multiphasic health screening improves health or saves lives. Experience with the system must be expanded for really reliable testing of its contribution. Finally, there is no satisfactory formula for generalizing the cost–result relationship in achieving the objective of health. There is general expectation that a technical staff, using automation and computers, will play a significant role in the future achievement of those functions that lead to better health. Technical and functional capacities have demonstrated that such services are feasible. The real question is how the new technology can be introduced into the existing complicated system of health care for the good of the patient, of the larger community, and of the several elements of the health profession.

USE OF RADIATIONS IN DIAGNOSIS

X-rays. The discovery of the roentgen ray (X-ray) introduced a radical new improvement into medical diagnosis by making available an immediate and detailed view of internal structures. Originally, results were recorded on photographic film. The pattern of shadows cast was related to the density of the structures penetrated or partly penetrated by the X-rays. Bones, air, fluid, and solid structures were viewed in the living body. Modern techniques have permitted many advances. Fluoroscopy, which projects X-ray images upon a fluorescent screen, gives visualization of living, moving structures. Now accurate motion pictures with image intensifiers permit detailed studies of contrast material flowing through the coronary arteries in the beating heart. Many exciting observations have been made possible by perfecting methods of increasing the contrast of various pipes, tubes, and vessels and passageways in the body. Radiopaque material such as barium may be swallowed and its progress through the alimentary canal followed in detail; or it may be introduced in an enema. The functional capacity of organs such as the kidney and the liver to concentrate soluble radiopaque material permits getting a vivid picture of the outline of bile ducts, gallbladder, the larger renal (kidney) collecting system, the renal pelvis and calices (the largest cavity in the kidney and subdivisions of it), ureter, and bladder. Direct injection of soluble radiopaque material into artery and vein can be done selectively, to concentrate on a particular portion of the vascular tree or functional unit such as individual coronary artery or a renal artery. X-rays of veins may be obtained in the same selective way. A more recent development is the injection of material into lymphatic vessels, which permits mapping out normal and diseased tributary lymphatic structures and lymph nodes.

Fluoroscopy and moving X-ray pictures

Another contrast method involves the use of air or rapidly absorbed gases such as carbon dioxide, which may be introduced into the abdomen; the chest; the fallopian tubes; or into the ventricles, or cavities, of the brain. An oily radiopaque material may be introduced into the bronchial tree to map out the shape, size, and distribution of nearly the whole of the tracheobronchial tree, the complete passageway for air except the nose and larynx.

The whole world of bone and joint defects may be demonstrated. The salivary duct system may be outlined. The vascular structures of spleen and portal system can be seen. X-rays of the spinal cord may reveal deformity or encroachment by tumour. Double contrast may be obtained by the use of a thin mixture of barium and air to outline irregularities or lesions of the mucous membrane of the intestine.

Stereoscopic views

Stereoscopic studies of the chest are helpful in getting a three-dimensional view of the location of lesions in the lungs. A different technique, planigraphy and tomography, by rotating camera and film blurs out all but the focal point or plane under study, giving, in effect, a three-dimensional idea of solid lesions whose density differs from that of surrounding structures.

Use of radioactive isotopes. Radioactive isotopes, with their system of tagging, have been used in a number of ways for diagnostic purposes. (Isotopes of any element are forms of that element that are alike chemically but differ in mass of nucleus.) Two of these are illustrated briefly. The element iodine is absorbed from the alimentary canal, taken up in the circulation, and removed selectively by the thyroid gland for incorporation in active hormones. Thus, the state of function of the thyroid gland, its adequacy, overactivity, or failure, can be judged by giving the subject a drink of a dilute solution in water of iodine containing a small amount of the radioactive form of iodine, iodine-131. The concentration is then tested by putting a Geiger counter over the thyroid and measuring the radioactivity taken up by the end of four hours and of twenty-four hours.

Pernicious anemia, which is characterized, among other things, by a failure of the stomach to produce hydrochloric acid and an enzyme (called the intrinsic factor) that helps the digestion and assimilation of vitamin B_{12}, may be studied by giving a large amount of vitamin B_{12} containing radioactive cobalt. The subsequent excretion of radioactivity in the urine reveals how much vitamin B_{12} has been absorbed, taken into the blood stream, and excreted in the urine. By measuring radioactivity of both urine and stool, one may find that all the radioactivity is

in the stool, that none has been absorbed, and that the patient has pernicious anemia or perhaps has had the whole stomach removed.

In tests that involve tagging by radioactive isotopes, it is possible to measure biologically active materials that exist in almost infinitesimally small concentrations in the body. While such studies have been more important in building an understanding of normal mechanisms in disease, they may become important in routine diagnostic work.

Ultrasound. Ultrasound (sound waves above the range audible to man) has been used routinely to measure the motion of the heart's valves and to determine the existence and degree of pericardial effusion, the collection of fluid in the space around the heart. The measurement of left ventricular wall thickness and of atrial tumours, and the study of the aortic root, have been made possible by more advanced techniques. Newer techniques have made it possible to measure the aortic arch, the right pulmonary artery, and the left atrium.

III. Other important considerations

SURGICAL DIAGNOSIS

Surgical diagnosis includes routine and special diagnosis in patients who have problems amenable to surgical procedures and the use of surgical techniques for diagnosis. At times, for example, the problem of life-endangering bleeding from the alimentary canal from an undetermined site may require surgical intervention without knowledge of the bleeding area. In some situations such as obstructive jaundice, an operation may be made to determine whether there is a stone or remediable stricture (narrowing) or whether a tumour, operable or inoperable, is causing the trouble.

A further diagnostic aid during an operation is the pathologist's report from frozen sections of material obtained by the surgeon. Gross inspection and palpation may not disclose whether tumour or inflammatory reaction is responsible. Information from the pathologist may determine whether the operation proceeds with a radical destruction of tissue, mere palliation, or doing nothing.

Biopsy

Surgical means may be employed for obtaining tissue for diagnosis. This ranges from the gentle scraping of surface cells from the skin to get material for cultures of yeast and fungi, simple biopsy (removal of a specimen) of the skin, to biopsy of lymph nodes or tissue from organs. Washings or specimens of abnormal tissue may be obtained from the bronchi. Devices may be passed along the alimentary canal to obtain small bits of mucosal tissues. Culture and other analyses may be fruitful and helpful. The so-called Papanicolaou smear of cells from the cervix to test for cancer is another example of such procedures.

Somewhat intermediate between an operation that explores the abdominal cavity and the scraping of the skin are needle biopsies to obtain material from skin, muscle, liver, lung, and kidney. Such techniques, particularly with the employment of the electron microscope, have greatly advanced diagnosis.

THE MEDICAL RECORD

A patient's medical record in a doctor's office or in the hospital is a written, typewritten, or electronic record that contains the medical story of a person's health and disease, the findings from the physical examination, the reports of laboratory tests, the results of special examinations, the findings and opinions of consultants, and then the synthesis of these matters in the diagnoses made by the responsible physician. There follows a continuum of notes on treatment, tests, medicines, operations, radiation, physical therapy, and progress. It is a stage-by-stage summary of events by physicians, nurses, and others. To be helpful it must be completely accurate. To be good it must be brief.

Uses for medical record

The medical record should assist all responsible persons in the care and treatment of the patient. It must be a teaching record, a document for clinical research, and a source of statistics. The patient's medical record documents the criteria for insurance claims. It is the source of legal proof in claims for injury, poisoning, homicide, or malpractice. A medical record must be up-to-date for the proper study and treatment of the patient. It has to be accurate for any subsequent use as a medical record.

It is essential that those who write in the record be identified: physician in charge, junior staff man, resident, intern, undergraduate student, nurse, or aide. For the record to reflect the essential moving picture, dates and times must be identified. Because the record is a privileged communication, it has to remain in the physician's locked files or in the hospital record room.

While the exact format of the recorded history may vary, the advance guard of medical teaching and care is collecting information to provide a data bank that becomes the basis for problem solving. The changing pattern of the disease and response to therapy is considered in the light of problems.

But the standard history, which is usually kept on one of a variety of forms, begins with complete identity of the patient. These are necessary vital statistics and vital signs; information about the informant if other than the patient; a statement of the main problem; and the present illness. The rest of the information in the history deals with past events. This includes general health, specific diseases, hospitalizations, operations and injuries, social history, family history, and a systematic review of the organs of the body and the identifiable functioning systems of the body. This gives an indication of their performance and functional state, past and present.

Many of the vital statistics are obtained by clerks. Some are entered on a plastic plate used to identify patient, tests, procedures, consultations, and charges.

SOURCES OF ERROR IN MEDICAL DIAGNOSIS

In analyzing the mistakes and false moves in the study of many patients whose problems ultimately came to solution, several sorts of errors are identified. Most important is failure to obtain available information because of improper history taking. Another is in the failure to understand the meaning of a patient's adequately elicited subjective complaints. Such flaws in the history produce more diagnostic error than failure to feel a mass, hear an abnormal heart sound, or obtain a useful laboratory test. Obviously, the whole of a person's past existence is likely to be more significant than his present state at an instant in time. The immediate past and recent events usually are more important than those in the distant past, but there are many exceptions.

In the communication systems required for the physical record of a patient, there may be errors in labelling samples of blood or urine, errors in recording the results of tests on laboratory reports, technical errors in the test procedure, errors in entering the results in the patient's record, and this may include mixups from similar or identical names, or a clerical error in picking up a wrong chart.

Many conditions—language barriers, coma, stupor or delirium, inability to articulate or recall—prevent the story of the past from being obtained. This makes diagnosis like trying to make a judgment of what has gone on in a play from seeing only a still photograph of a single scene.

Errors may occur when a doctor forgets to ask a question or to do an examination. Errors of omission are commoner than those of commission.

An understanding of the emotional and intellectual capacity and state of the patient is essential. It is as necessary to tone down the interminable complaints of the perennial complainer as it is to probe for information from a stoical, repressed, or reserved person.

Just as disease is an ever-changing process, so the diagnosis is the best approximation, in terms of structure, function, and malfunction, of those derangements and troubles that interfere with the patient's natural existence by causing pain, malfunction, or fear.

BIBLIOGRAPHY

Historical: H.I. BOWDITCH, *The Young Stethoscopist, or the Student's Aid to Auscultation*, 2nd ed. (1848), a good discussion of the problem of auscultation, long enough after Laën-

nec's discovery for experience to have accumulated; A. FLINT, *A Manual of Auscultation and Percussion; Embracing the Physical Diagnosis of Diseases of the Lungs and Heart, and of Thoracic Aneurism*, 2nd rev. ed. (1880), a good discussion of the state of the art nearly a hundred years ago; J.B. HERRICK, *A Handbook of Medical Diagnosis for Students* (1895), one of the early American books devoted to a study of diagnosing diseases both by kind and by symptom.

General references: R.C. AHLVIN, "Biochemical Screening—A Critique," *New Eng. J. Med.*, 283:1084–1086 (1970), a contemporary paper emphasizing the shortcomings of biochemical screening; H. BAILEY and A. CLAIN, *Demonstrations of Physical Signs in Clinical Surgery*, 12th ed. (1954), notable for excellent photographs, line drawings, and brief descriptions; W.B. BEAN, "Precordial Noises Heard at a Distance from the Chest," in *Monographs in Medicine*, 1:22–65 (1952), a critical review of a large variety of disorders in which the noises generated by the heart are exaggerated so that they may be heard at a distance from the chest, *Vascular Spiders: And Related Lesions of the Skin* (1958), a review of the characteristics and significance of vascular spiders, hereditary hemorrhagic telangiectasia, and many other diseases of the skin, extensively illustrated, and *Rare Diseases and Lesions: Their Contributions to Clinical Medicine* (1967), contains an important chapter on bleeding from the alimentary canal, discussions of rare diseases with diagnostic skin lesions, and a concluding essay on the naming of diseases; L. CLENDENING and E.H. HASHINGER, *Methods of Diagnosis* (1947), a book notable for its introductory principia diagnostica, "Logic and Diagnosis," and "The Organon of Diagnosis," in which the rules of formal logic are shown to be not only valid but essential for medical examination; E.L. and R.L. DEGOWIN, *Bedside Diagnostic Examination*, 2nd ed. (1969), the best single manual on medical diagnosis, excellently organized for an introduction to the medical examination but equally valuable as a reference book; J.A. JACQUEZ (ed.), *The Diagnostic Process* (1964), a discussion of the approaching era of the computer in medicine; A.R. FEINSTEIN, *Clinical Judgment* (1967), a thoughtful critique of traditional approaches to medical examination; S.R. GARFIELD, "Multiphasic Health Testing and Medical Care as a Right," *New Eng. J. Med.*, 283:1087–1089 (1970), contemporary views of the problems of applying modern techniques to medical care; B.B. GOLDBERG, "Suprasternal Ultrasonography," *JAMA*, 215:245–250 (1971), an example of steady progress and change with the new development of diagnostic techniques; R.H. KAMPMEIER, *Physical Examination in Health and Disease* (1950), notable for its broad coverage and many illustrations; M.A. KRUPP, M.J. CHATTON, and S. MARGEN, *Current Diagnosis and Treatment*, 10th ed. (1971), an excellent manual relating diagnosis and therapy; C.M. MACBRYDE (ed.), *Signs and Symptoms: Applied Pathologic Physiology and Clinical Interpretation*, 4th ed. (1964), a collection of excellent monographs developed around major signs and symptoms of disease; R.H. MAJOR, *Major's Physical Diagnosis*, 7th ed. by M.H. DELP and R.T. MANNING (1968), an excellent general introduction to physical diagnosis with good pictures, line drawings, and an excellent bibliography; W.L. MORGAN, JR., and G.L. ENGEL, *The Clinical Approach to the Patient* (1969), an excellent introductory book that details the blending of the art of medicine with the sciences of medicine; T.L. STEDMAN, *Stedman's Medical Dictionary*, 21st ed. (1966), an excellent medical dictionary that has kept up to date in a rapidly growing field; L.B. PAGE and P.J. CULVER (eds.), *A Syllabus of Laboratory Examinations in Clinical Diagnosis: Critical Evaluation of Laboratory Procedures in the Study of the Patient*, rev. ed. (1960), a syllabus that emphasizes laboratory errors, variability, quality control, and the necessity for critical interpretation; J.C. TODD and A.H. SANFORD, *Clinical Diagnosis by Laboratory Methods*, 14th ed. by I. DAVIDSOHN and J.B. HENRY (1969), an excellent reference book for laboratory methodology.

(W.B.B.)

Dialects

A dialect is a variety of a language. The word comes from the Ancient Greek *dialektos* "discourse, language, dialect," which is derived from *dialegesthai* "to discourse, talk." A dialect may be distinguished from other dialects of the same language by features of any part of the linguistic structure—the phonology, morphology, or syntax. In the sound system of American English, for example, certain dialects distinguish the vowel in "caught" from that in "cot," while others do not, and in some dialects "greasy" is pronounced with an *s* sound and in others with a *z* sound. In morphology (word formation), various dialects in the Atlantic states have "clim," "clum," "clome," or "cloome" instead of "climbed," and, in syntax, there are "sick to his stomach," "sick at his stomach," "sick in," "sick on," and "sick with." On the level of vocabulary, examples of dialectal differences include American English "subway," contrasting with British English "underground"; and "corn," which means "maize" in the U.S., Canada, and Australia, "wheat" in England, and "oats" in Scotland. Nevertheless, while dialects of the same language differ, they still possess a common core of features.

Frequently, the label dialect, or dialectal, is attached to substandard speech, language usage that deviates from the accepted norm; *e.g.*, the speech of many of the heroes of Mark Twain's novels. On the other hand, the standard language can be regarded as one of the dialects of a given language. In a special historical sense, the term dialect applies to a language considered as one of a group deriving from a common ancestor; *e.g.*, English, Swedish, and German are Germanic dialects.

There is often considerable difficulty in deciding whether two linguistic varieties are dialects of the same language or two separate but closely related languages; this is especially true of dialects of primitive societies, in which the difference is essentially one of degree. Many decisions regarding dialects versus languages must be arbitrary.

Normally, dialects of the same language are considered to be mutually intelligible, while different languages are not. Intelligibility between dialects is, however, almost never absolutely complete; on the other hand, speakers of closely related languages can still communicate to a certain extent when each uses his own mother tongue. Thus, the criterion of intelligibility is quite relative. In more developed societies, the distinction between dialects and related languages is easier to make because of the existence of standard languages and, in some cases, national consciousness.

Dialects, patois, idiolects

Among the synonyms for dialect, the word idiom refers to any kind of dialect, or even language, whereas patois, a term from French, denotes rural or provincial dialects, often with a deprecatory connotation. An idiolect is the dialect of one individual person at one time. This term implies an awareness that no two persons speak in exactly the same way—*i.e.*, without slight differences in vocabulary—and that each person's dialect is constantly undergoing change—*e.g.*, by the introduction of newly acquired words. Most recent investigations emphasize the versatility of each person's speech habits according to levels or styles of language usage.

Another synonym for dialect is the term vernacular; it refers to the common, everyday speech of the ordinary people of a region. The word accent has numerous meanings; in addition to denoting the pronunciation of a person or a group of people ("a foreign accent," "a British accent," "a Southern accent"), it also refers to features of pitch or stress. In contrast to accent, the term dialect is used to refer not only to the sounds of language but also to grammar and vocabulary.

VARIETIES OF DIALECTS

Geographic dialects. The most widespread type of dialectal differentiation is geographic. As a rule, the speech of one locality differs at least slightly from that of any other place. Differences between neighbouring local dialects are usually small, but, in travelling farther in the same direction, differences accumulate. Every dialectal feature has its own boundary line, called an isogloss (or sometimes heterogloss). Isoglosses of various linguistic phenomena rarely coincide completely, and by crossing and interweaving they constitute intricate patterns on dialect maps (see Figure 1). Frequently, however, several isoglosses are grouped approximately together into a bundle of isoglosses. This grouping is caused either by geographic obstacles that arrest the diffusion of a number of innovations along the same line or by historical circumstances, such as political borders of long standing, or by migrations that have brought into contact two populations whose dialects were developed in noncontiguous areas.

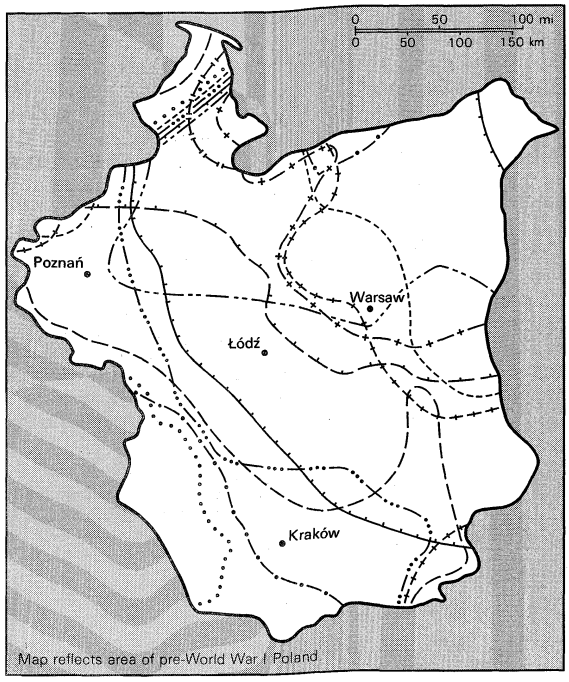


Figure 1: Isoglosses of morphological simplifications in Polish dialects.
Adapted from K. Nitsch, *Dialekty jezyka polskiego* (1957)

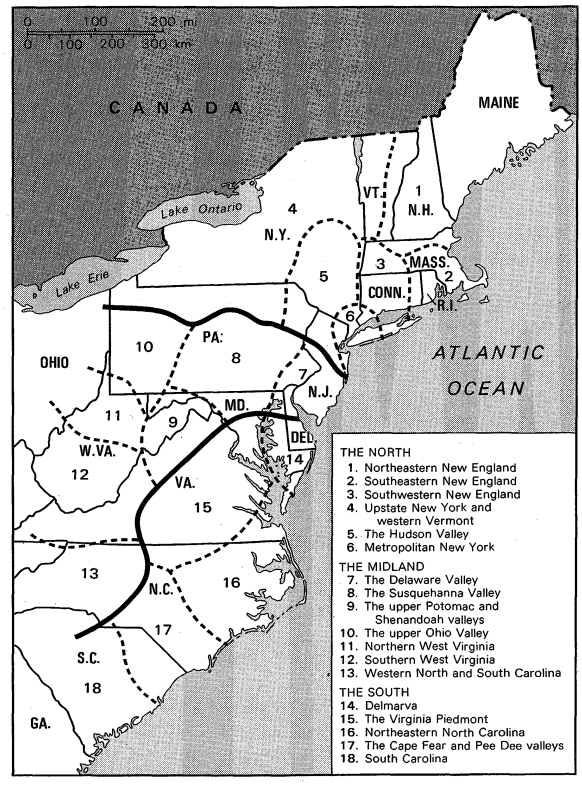


Figure 2: Dialect areas of the eastern United States.
Adapted from H. Kurath, *Word Geography of the Eastern United States;* the University of Michigan Press

Local and regional dialects

Geographic dialects include local ones (*e.g.*, the Yankee English of Cape Cod or of Boston, the Russian of Moscow or of Smolensk) or regional ones, such as Delaware Valley English, Australian English, or Tuscan Italian. Such entities are of unequal rank; South Carolina English, for instance, is included in Southern American English. Regional dialects do have some internal variation, but the differences within a regional dialect are supposedly smaller than differences between two regional dialects of the same rank. In a number of areas ("linguistic landscapes") where the dialectal differentiation is essentially even, it is hardly justified to speak of regional dialects. This uniformity has led many linguists to deny the meaningfulness of such a notion altogether; very frequently, however, bundles of isoglosses—or even a single isogloss of major importance—permit the division of a territory into regional dialects (see Figure 2 for the dialectal division of American English in the Atlantic states). The public is often aware of such divisions, usually associating them with names of geographic regions or provinces, or with some feature of pronunciation; *e.g.*, Southern English or Russian *o*-dialects and *a*-dialects. Especially clear-cut cases of division are those in which geographic isolation has played the principal role; *e.g.*, Australian English or Louisiana French.

Social dialects. Another important axis of differentiation is that of social strata. In many localities, dialectal differences are connected with social classes, educational levels, or both. More highly educated speakers and, often, those belonging to a higher social class tend to use more features belonging to the standard language, whereas the original dialect of the region is better preserved in the speech of the lower and less educated classes. In large urban centres, innovations unknown in the former dialect of the region frequently develop. Thus, in cities the social stratification of dialects is especially relevant and far reaching, whereas in rural areas, with a conservative way of life, the traditional geographic dialectal differentiation prevails.

Technical terminology and slang

Educational differences among speakers strongly affect the extent of their vocabulary. In addition, practically every profession has its own expressions, which include the technical terminology and sometimes also the casual words or idioms peculiar to the group. Slang, too, is characterized mainly by a specific vocabulary and is much more flexible than an ordinary dialect, as it is subject to fashion and depends strongly on the speaker's age group. Slang—just as a professional dialect—is used mainly by persons who are in a sense bidialectal; *i.e.*, they speak some other dialect or the standard language, in addition to slang. Dialectal differences also often run parallel with the religious or racial division of the population.

DEVELOPMENT OF DIALECTS

The basic cause of dialectal differentiation is linguistic change. Every living language constantly undergoes changes in its various elements. Because languages are extremely complex systems of signs, it is almost inconceivable that linguistic evolution could affect the same elements and even transform them in the same way in all localities where one language is spoken and for all speakers in the same locality. At first glance, differences caused by linguistic change seem to be slight, but they inevitably accumulate with time (*e.g.*, compare Chaucer's English with modern English or Latin with modern Italian, French, Spanish, or Romanian). Related languages usually begin as dialects of the same language.

Dialectal change and diffusion. When a change (an innovation) appears among only one section of the speakers of a language, this automatically creates a dialectal difference. Sometimes an innovation in dialect A contrasts with the unchanged usage (archaism) in dialect B. Sometimes a separate innovation occurs in each of the two dialects. Of course, different innovations will appear in different dialects, so that, in comparison with its contemporaries, no one dialect as a whole can be considered archaic in any absolute sense. A dialect may be characterized as relatively archaic, because it shows fewer innovations than the others; or it may be archaic in one feature only.

After the appearance of a new dialectal feature, interaction between speakers who have adopted this feature and those who have not leads to the expansion or the curtailment of its area or even to its disappearance. In a single social milieu (generally the inhabitants of the same locality, generation, and social class), the chance of the complete adoption or rejection of a new dialectal feature is very great; the intense contact and consciousness of membership within the social group fosters such uni-

formity. When several age groups or social strata live within the same locality and especially when people speaking the same language live in separate communities, dialectal differences are easily maintained.

The element of mutual contact plays a large role in the maintenance of speech patterns; that is why differences between geographically distant dialects are normally greater than those between dialects of neighbouring settlements. This also explains why bundles of isoglosses so often form along major natural barriers—impassable mountain ranges, deserts, uninhabited marshes or forests, or wide rivers—or along political borders. Similarly, racial or religious differences contribute to linguistic differentiation because contact between members of one faith or race and those of another within the same area is very often much more superficial and less frequent than contact between members of the same racial or religious group. An especially powerful influence is the relatively infrequent occurrence of intermarriages, thus preventing dialectal mixture at the point where it is most effective; namely, in the mother tongue learned by the child at home.

Natural barriers and language change

Unifying influences on dialects. Communication lines such as roads (if they are at least several centuries old), river valleys, or seacoasts often have a unifying influence. Also, important urban centres, such as Paris, Utrecht, or Cologne, often form the hub of a circular region in which approximately the same dialect is spoken. In such areas, the prestige dialect of the city has obviously expanded. As a general rule, those dialects, or at least certain dialectal features, with greater social prestige tend to replace those that are valued lower on the social scale.

In times of less frequent contact between populations, dialectal differences increase; in periods of greater contact, they diminish. The general trend in modern times is for dialectal differences to diminish, above all through the replacement of dialectal traits by those of the standard language. Mass literacy, schools, increased mobility of populations, and, in the last few decades, the ever-growing role of mass communications all contribute to this tendency. Naturally, the extent of such unifying action varies greatly in different linguistic domains, depending on the level of civilization. Nevertheless, the most thorough example of linguistic force exerted by a single dominating civilization belongs to ancient times: in the Hellenistic era, almost all ancient Greek dialects were replaced by the so-called koine, based on the dialect of Athens.

Mass migrations may also contribute to the formation of a more or less uniform dialect over broad geographic areas. Either the resulting dialect is that of the original homeland of a particular migrating population or it is a dialect mixture formed by the levelling of differences among migrants from more than one homeland. The degree of dialectal differentiation depends to a great extent on the length of time a certain population has remained in a certain place. Thus, it is understandable that the diversification of the English language is far greater in the British Isles than, for example, in North America (especially if the number of dialectal differences is considered on a comparable area basis, such as how many per 1,000 square miles). In the U.S. itself much greater diversity is evident among dialects in old colonial America—along the Atlantic coast—than among dialects west of the Appalachians. It is also typical that phonological differences are more far reaching in Switzerland among Swiss-German dialects than throughout the vast territory where the Russian language is spoken, extending from Leningrad to eastern Siberia. Such a situation results not only from migrations of the Russian population, (as compared to the centuries of Swiss stability) but also from the contrasting geographical configurations: in the U.S.S.R., there is unobstructed communication in all directions; in mountainous Switzerland, the territory is carved into small, isolated units.

Dialect differentiation and population movements

Migrations and, more rarely, geographical phenomena may in some areas cause a much stronger dialectal differentiation in one direction than in others. Isoglosses in the U.S., for example, run predominantly in an east–west direction, reflecting the westward stream of migration during the colonization of areas west of the Appalachians. Similarly, the majority of isoglosses in Russia follow latitude, but in the opposite (west–east) direction.

Focal, relic, and transitional areas. Dialectologists often distinguish between focal areas—which provide sources of numerous important innovations and usually coincide with centres of lively economic or cultural activity—and relic areas—places toward which such innovations are spreading but have not usually arrived. (Relic areas also have their own innovations, which, however, usually extend over a smaller geographical area.) Relic areas or relic phenomena are particularly common in out-of-the-way regional pockets or along the periphery of a particular language's geographical territory. An example of a focal area in the U.S. would be the Boston region, while rural Maine and New Hampshire and Cape Cod and Nantucket Island would be typical relic areas (see Figure 3).

Adapted from A.J. Bronstein, *The Pronunciation of American English—An Introduction to Phonetics* (1960), Appleton-Century-Crofts; with permission from the American Council of Learned Societies

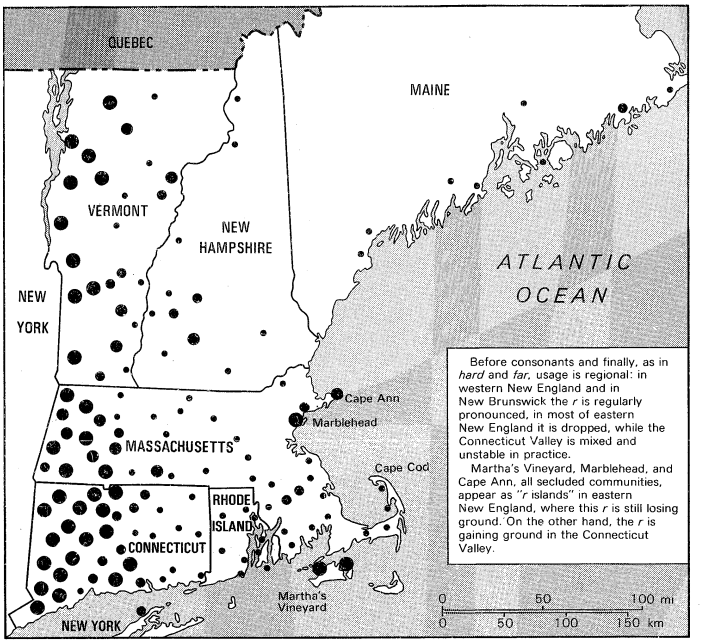


Figure 3: New England pronunciation of preconsonantal and final *r*. The largest circles indicate regular use of this *r*, the smallest ones sporadic use, and the two intermediate sizes rather evenly divided usage.

The borders of regional dialects often contain transitional areas that share some features with one neighbour and some with the other. Such mixtures result from unequal diffusion of innovations from both sides. Similar unequal diffusion in mixed dialects in any region also may be a consequence of population mixture created by migrations.

In regions with many bilingual speakers (*e.g.*, along the border between two languages) dialects of both languages will often undergo changes influenced by the other tongue. This is manifested not only in numerous loanwords but often also in the adoption of phonological or grammatical features. Such phenomena are particularly frequent in a population that once spoke one language and only later adopted the second language. In extreme cases, a so-called creolized language develops. (Creoles are pidgin languages that have become the only or major language of a speech community. See PIDGIN.)

Standard languages. Standard languages arise when a certain dialect begins to be used in written form, normally throughout a broader area than that of the dialect itself. The ways in which this language is used—in administrative matters, literature, economic life—lead to the minimization of linguistic variation. The social prestige attached to the speech of the richest, most powerful, and most highly educated members of a society trans-

Written form of dialects

forms their language into a model for others; it also contributes to the elimination of deviating linguistic forms. Dictionaries and grammars help to stabilize linguistic norms, as do the activity of scholarly institutions and, sometimes, governmental intervention. The base dialect for a country's standard language is very often the original dialect of its capital—in France, Paris; in England, London; in Russia, Moscow. Or the base may be a strong economic and cultural centre—in Italy, Florence. Or the language may be a combination of several regional dialects; *e.g.*, German or Polish.

Even a standard language that was originally based on one local dialect changes, however, as elements of other dialects infiltrate into it over the years. The actual development in any one linguistic area depends on historical events. Sometimes even the distribution of standard languages may not correspond to the dialectal situation. Dutch and Flemish dialects are a part of the Low German dialectal area, which embraces all of northern Germany, as well as the Netherlands and part of Belgium. In one part of the dialectal area, however, the standard language is based on High German, and, in the other part, the standard language is Dutch or Flemish, depending on the nationality of the respective populations. In the U.S., where there is no clearly dominant political or cultural centre—such as London or Paris—and where the territory is enormous, the so-called standard language shows perceptible regional variations in pronunciation.

In most developed countries, the majority of the population has an active (speaking, writing) or at least passive (understanding) command of the standard language. Very often the rural population, and not uncommonly the lower social strata of the urban population as well, are in reality bidialectal. They speak their maternal dialect at home and with friends and acquaintances in casual contacts, and they use the standard language in more formal situations. Even the educated urban population in some regions uses the so-called colloquial language informally. In the German-, Czech-, and Slovene-speaking areas of middle Europe, for example, a basically regional dialect from which the most striking local features have been eliminated is spoken. The use of this type of language is supported by psychological factors, such as feelings of solidarity with a certain region and pride in its traditions or the relaxed mood connected with informal behaviour.

DIALECT GEOGRAPHY

Dialect study as a discipline—dialectology—dates from the first half of the 19th century, when local dialect dictionaries and dialect grammars first appeared in western Europe. Soon thereafter, dialect maps were developed; most often they depicted the division of a language's territory into regional dialects. The 19th-century Romantic view of dialects and folklore as manifestations of the ethnic soul furnished a great impetus for dialectology.

19th-century impetus for dialect study

Early dialect studies. The first dialect dictionaries and grammars were most often written by scholars describing the dialect of their birthplace or by fieldworkers whose main method of investigation was free conversation with speakers of the dialect, usually older persons and, preferably, those who showed the least degree of literacy and who had travelled as little as possible. Many of these grammars and dictionaries recorded dialectal traits that deviated from the standard language. In the second half of the 19th century, when historical and comparative linguistic study was flourishing, it became customary to focus attention on the fate of particular elements of the archaic language in a given dialect; *e.g.*, the changes that Latin vowels and consonants underwent when used in different positions in a particular Romance dialect.

With the accumulation of dialectal data, investigators became increasingly conscious of the inadequacy of viewing dialects as internally consistent units that were sharply differentiated from neighbouring dialects. It became more and more clear that each dialectal element or phenomenon refused to stay neatly within the borders of a single dialect area and that each had its own isogloss; consequently, maps of dialects would have to be replaced by maps showing the distribution of each particular feature. While sound scientifically, the preparation and compilation of such maps, called linguistic atlases, is a difficult, costly, voluminous, and time-consuming job.

Dialect atlases. Dialect atlases are compiled on the basis of investigations of the dialects of a large number of places; a questionnaire provides uniform data. There are two basic methods of data collection: fieldwork and survey by correspondence. Fieldwork, in which a trained investigator transcribes dialectal forms directly (or on tape), affords more precise data and enables the questionnaire to include a greater number of diverse questions; but it implies a necessarily limited number of points to be covered. The advantage of the correspondence method lies in its ability to encompass more points at less cost and with less time expended in gathering the data. On the other hand, rural schoolteachers, normally the persons who complete such questionnaires, can answer only a relatively small number of questions and often imperfectly.

The first large-scale enterprise in linguistic geography was the preparation of the German linguistic atlas. In the 1880s, the initiator of this great undertaking, Georg Wenker, composed 40 test sentences that illustrated most of the important ways in which dialects differed and sent them to schoolmasters in over 40,000 places in the German Empire. The sentences were to be translated into the local dialect. Publication of the results was not begun until 1926; the main cause of the delay was the enormous quantity of material to be arranged and analyzed.

The French atlas

The famous French linguistic atlas of Jules Gilliéron and Edmond Edmont was based on a completely different concept. Using a questionnaire of about 2,000 words and phrases that Gilliéron had composed, Edmont surveyed 639 points in the French-speaking area. The atlas, compiled under the direction of Gilliéron, was published in fascicles from 1902 to 1912 and furnished both a strong stimulus and the basic model for work on linguistic atlases elsewhere in the world. European linguists, especially in Romance- and Germanic-speaking countries, were the first to participate in such atlas projects. One of the most significant contributions is the linguistic atlas of Italy and southern Switzerland by Karl Jaberg and Jakob Jud; it appeared from 1928 to 1940. Particularly noteworthy in its attention to precise definitions of meaning, this atlas often used illustrations and described objects and actions of village life denoted by the questionnaire's words.

At present, dialects of virtually all European languages have been treated in linguistic geography studies. In some countries, data are still being collected and classified and maps are being drawn, but in others a second generation of atlases is already underway. French dialectologists, for instance, are now working on regional atlases that will complement data contained in the *Atlas linguistique de la France*. In England, work began in 1946, under the direction of Harold Orton and Eugene Dieth; the first volume of the *Survey of English Dialects* was published in 1962. In Slavic-speaking countries, work is now underway both on atlases of separate Slavic languages and on the large general Slavic linguistic atlas that will cover nearly 1,000 locales in all parts of European territory where Slavic languages are spoken. Outside Europe, the greatest amount of work in linguistic geography has been completed in Japan and in the United States.

As early as 1905–06, a committee of Japanese dialectologists published the first linguistic atlas of Japan in two volumes, one devoted to phonology and one to morphology. Subsequent work has been done on a new atlas of Japan as a whole and on several regional atlases. The extensive activity of Chinese specialists has concentrated on descriptions of particular local and regional dialects. The Chinese situation is a peculiar one because of the enormous number of people who speak Chinese, the very significant dialectal differentiation (certain dialects, particularly those in the South of China, would be con-

sidered by Western standards as separate languages), and the nature of the Chinese script. Chinese characters do not represent sounds but concepts. Because of this, the written language can be read without difficulty in many different dialect areas, although its spoken form varies greatly from one region to another.

Because of the enormous size of the U.S., atlas surveys were done by region. Between 1931 and 1933, fieldworkers under the direction of the linguist Hans Kurath surveyed 213 New England communities; the results were published in the *Linguistic Atlas of New England* (with 734 maps) in 1939–43. Based on the methodological experience of Jaberg and Jud in their atlas of Italy and southern Switzerland, this work involved systematic investigations not only among the relatively uneducated but also among better educated, more cultured informants and among the very well educated, cultured, and informed members of a community. Thus the dimension of social stratification of language was introduced into linguistic geography, and valuable material about regional linguistic standards became available.

After 1933, fieldwork was extended to the other Atlantic states. Lack of financial support, however, has hindered the publication of these atlases. Nevertheless, several works based on the material gathered have appeared, among them Kurath's *Word Geography of the Eastern United States*, E. Bagby Atwood's *Survey of Verb Forms in the Eastern United States*, and Kurath's and McDavid's *Pronunciation of English in the Atlantic States*. Independent work was carried out in other U.S. regions, mainly with an adapted form of the questionnaire developed for the Atlantic states; only introductions or summaries of material in the files have been published, however, because of lack of funds.

Works on American English

The most effective and thorough—as well as the most expensive—way of presenting data in linguistic atlases is by printing the actual responses to questionnaire items right on the maps. Phenomena of linguistic geography, however, are usually represented by geometric symbols or figures at the proper points on the map, or, even more summarily, by the drawing of isoglosses (linguistic boundaries) or by shading or colouring the areas of particular features.

Only dialect atlases can furnish the complexity of data of the major dialectal phenomena in a multitude of geographic locations in a manner that both assures commensurability of the data and allows a panoramic examination of the whole gamut of data. The inventory of linguistic phenomena is so rich, however, that no one questionnaire can encompass it all. Moreover, the use of a questionnaire unavoidably brings about a schematization of answers that is lacking in spontaneity. For these reasons, other kinds of publications, such as dialect dictionaries or monographs based on extensive free conversation with speakers of local dialects, are indispensable complements to linguistic atlases.

The value and applications of dialectology. The scientific interest of dialectology lies in the fact that dialects are a valuable source of information about popular culture. They reflect not only the history of a language but, to a great extent, the ethnic, cultural, and even political history of a people as well. A knowledge of dialectal facts provides practical guidance to school systems that are trying to teach the standard language to an ever greater number of pupils.

In the 1930s, the value of dialectology to the study of language types became apparent. Because dialects greatly outnumber standard languages, they provide a much greater variety of phenomena than languages and thus have become the main source of information about the types of phenomena possible in linguistic systems. Also, in some languages, but not in others, an extremely wide structural variation among dialects has been found. In Yugoslavia, where two closely related Slavic languages, Serbo-Croatian and Slovene, are spoken, dialects are found with synthetic declension (case endings, as in Latin) and analytic declension (use of prepositions and word order, as in English). In addition, there are among these dialects complex systems of verbal tenses contrasting with simple ones, as well as dialects with or without the dual number or the neuter gender. The dialects of Serbo-Croatian and Slovene also exhibit almost every type of prosodic structure (*e.g.*, tone, stress, length) found in European languages. Some dialects differentiate long and short vowels or rising and falling accents, while others do not; and in some, but not all of them, stress fulfills a grammatical function. Of the several dozen vowel and diphthong sounds that occur in these dialects, only five are common to all of them; all the rest are restricted to relatively small areas. All of this rich variety contrasts sharply with the relative structural uniformity of the English language—not only in the U.S. but wherever it is spoken. (The outstanding exceptions are the creolized dialects, which are distinguished by far-reaching structural pecularities.)

Variations in Serbo-Croatian and Slovene dialects

SOCIAL DIALECTOLOGY

The methodology of generative grammar (see GRAMMAR) was first applied to dialectology in the 1960s, when the use of statistical means to measure the degree of similarity or difference between dialects also became increasingly common. The most important development of that time, however, was the rapid growth of methods for investigating the social variations of dialects; this type of variation, in contrast to geographical variation, is especially great in the U.S., above all in large urban centres. In cities such as New York, a whole scale of speech variations can be found to correlate with the social status and educational levels of the speakers. In addition, age groups exhibit different patterns, but such age-group patterns of variation differ from one social stratum to another. Still another dimension of variation, especially important in the U.S., is connected with the race and the ethnic origin of a speaker as well as his date of immigration. So-called Black English has been influenced by the fact that most of the black population in non-southern U.S. regions originally came from the southeastern states; thus, many Black English peculiarities are in reality transplanted southeastern dialectal traits.

Social correlations with speech variation

Normally, speakers of one of the social dialects of a city possess at least some awareness of the other dialects. In this way, speech characteristics also become subjectively integrated into the system of signs indicating social status. And in seeking to enhance their social status, poorer and less educated speakers may try to acquire the dialect of the socially prestigious. Certain groups—*e.g.*, blacks and the working class—however, will, under certain conditions, show a consciousness of solidarity and a tendency to reject members who imitate either the speech or other types of behaviour of models outside their own social group.

As a consequence of an individual's daily contacts with speakers of the various social dialects of a city, elements of the other dialects are imperceptibly drawn into his dialect. The collective result of such experiences is the spread of linguistic variables; *i.e.*, groups of variants (sounds or grammatical phenomena) primarily determined by social (educational, racial, age, class) influences, an example being the existence of the two forms "He don't know" and the standard "He doesn't know." Traits representing variables in intergroup relations can become variable features in the speech of individuals as well; *i.e.*, an individual may employ two or more variants for the same feature in his own speech, such as "seeing" and "seein' " or "he don't" and "he doesn't." The frequency of usage for each variable varies with the individual speaker as well as with the social group. There are intermediate stages of frequency between different social groups and entire scales of transitions between different age groups, thus creating even greater variation within the dialect of an individual. The variables also behave differently in the various styles of written or spoken language used by each speaker.

The study of variables is one of the central tasks of any investigation of the dialects of American cities. Applying the statistical methods of modern sociology, linguists have worked out investigative procedures sharply different from those of traditional dialectology. The chief

Application of modern statistical methods to dialectology

contributor has been William Labov, the pioneer of social dialectology in the U.S. The basic task is to determine the correlation between a group of linguistic variables—such as the different ways of pronouncing a certain vowel—and extralinguistic variables, such as education, social status, age, and race. For a reasonable degree of statistical reliability, one must record a great number of speakers. In general, several examples of the same variable must be elicited from each individual in order to examine the frequency and probability of its usage. Accordingly, the number of linguistic variables that can be examined is quite limited, in comparison with the number of dialectal features normally recorded by traditional fieldworkers in rural communities; in these situations, the investigator is often satisfied with one or two responses for each feature.

A completely new, flexible, and imaginative method of interviewing is needed for such work in urban centres, as well as new ways of finding and making contact with informants. One example is Labov's method for testing the fate of final and preconsonantal *r* in speakers of different social levels. Choosing three New York City department stores, each oriented to a completely different social stratum, he approached a large number of salesladies, asking each of them about the location of a certain department that he knew to be on the fourth floor. Thus, their answers always contained two words with potential *r*'s—"fourth" and "floor." This shortcut enabled Labov to establish in a relatively short time that the salesladies in the store with richer customers clearly tended to use "*r*-full" forms, whereas those in the stores geared to the poorer social strata more commonly used "*r*-less" forms.

Social dialectology has focussed on the subjective evaluation of linguistic features and the degree of an individual's linguistic security, phenomena that have considerable influence on linguistic change. Linguistic scientists, in studying the mechanism of such change, have found that it seems to proceed gradually from one social group to another, always attaining greater frequency among the young. Social dialectology also has great relevance for a society as a whole, in that the data it furnishes will help deal with the extremely complex problems connected with the speech of the socially underprivileged, especially of minority groups. Thus, the recent emphasis on the speech of minority groups, such as the Black English of American cities, is not a chance phenomenon. Specific methods for such investigation are being developed, as well as ways of applying the results of such investigation to educational policies.

BIBLIOGRAPHY. General works include: ALBERT DAUZAT, *La géographie linguistique* (1922); and SEVER POP, *La dialectologie: Aperçu historique et méthodes d'enquêtes linguistiques*, 2 vol. (1950), both in French. Among the atlases are: FERDINAND WREDE *et al.*, *Deutscher Sprachatlas auf Grund des Sprachatlas des Deutschen Reichs von Georg Wenker* (1926–), which treats German dialects; JULES GILLIERON and EDMOND EDMONT, *Atlas linguistique de la France* (1902–12), a work on French dialects; KARL JABERG and JAKOB JUD, *Sprach- und Sachatlas Italiens und der Südschweiz*, dialect recordings by P. SCHEUERMEYER, G. ROHLFS, and M.L. WAGNER, 8 vol. (1928–40), which gives the linguistic geography of Italy (written in German). Atlases on English dialects include: HANS KURATH *et al.*, *Linguistic Atlas of New England*, 3 vol. (1939–43); HANS KURATH and RAVEN I. MCDAVID, JR., *The Pronunciation of English in the Atlantic States* (1961); and HAROLD ORTON *et al.*, *Survey of English Dialects*, 4 vol. (1962–67). The following books discuss American English dialects: CARROLL E. REED, *Dialects of American English* (1967); and WILLIAM LABOV, *The Social Stratification of English in New York City* (1966). Articles on dialectology appear in these collections: HAROLD B. ALLEN (ed.), *Readings in Applied English Linguistics*, 2nd ed. (1964), see esp. pt. 3, "Linguistic Geography"; and HAROLD HUNGERFORD, JAY ROBINSON, and JAMES SLEDD (eds.), *English Linguistics: An Introductory Reader* (1970), see esp. pt. 2, "Dialectology."

(P.I.)

Dias, Bartolomeu

As leader of the first expedition to round the Cape of Good Hope, thus opening the seaway from Europe to Asia, Bartolomeu Dias is usually considered to be the greatest of the Portuguese pioneers who explored the Atlantic in the 15th century.

Almost nothing is known of Dias' early life. His supposed descent from one of Prince Henry the Navigator's pilots is unproved, and his rank was the comparatively modest one of squire of the royal household. The name "Dias de Novais" does not appear in contemporary documents but only in the deed of appointment of his grandson as governor of Angola in 1571.

In 1474, King Afonso V entrusted his son, Prince John (later John II), with the supervision of Portugal's trade with Guinea and the exploration of the western coast of Africa. John sought to close the area to foreign shipping and after his accession in 1481 ordered new voyages of discovery to ascertain the southern limit of the African continent. The navigators were given stone pillars (*padrões*) to stake the claims of the Portuguese crown. Thus one of them, Diogo Cão, reached the Congo and sailed down the coast of Angola to Cape Santa Maria at 13°26′ south, where he planted one of John's markers supposing that he had attained the southernmost tip of Africa. Cão was ennobled and rewarded and sailed again: this time he left a marker at 15°40′ and another at Cape Cross, continuing to 22°10′ south. Royal hopes that he would reach the Indian Ocean were disappointed, and nothing more is heard of Cão. John II entrusted command of a new expedition to Dias. In 1486 a Portuguese had heard of a great ruler, the Ogané, far to the east, who was identified with the legendary Christian ruler Prester John. John II then sent Pêro da Covilhã (*q.v.*) and one Afonso Paiva overland to locate India and Abyssinia and ordered Dias to find the southern limit of Africa.

Dias' fleet consisted of three ships, his own "São Cristóvão," the "São Pantaleão" under his associate João Infante, and a supply ship under Dias' brother, whose name is variously given as Pêro or Diogo. The company included some of the leading pilots of the day, among them Pêro de Alenquer and João de Santiago, who earlier had sailed with Cão. A 16th-century historian, João de Barros, places Dias' departure in August 1486 and says that he was away 16 months and 17 days, but since two other contemporaries, Duarte Pacheco and Christopher Columbus, put his return in December 1488, it is now usually supposed that he left in August 1487.

Dias passed Cão's marker, reaching the "Land of St. Barbara" on December 4, Walfisch Bay on December 8, and the "Gulf of St. Stephen" (Elizabeth Bay) on December 26. After January 6, 1488, he was prevented by storms from proceeding along the coast and sailed south out of sight of land for several days. When he again turned to port, no land appeared, and it was only on sailing north that he sighted land on February 3. He had thus rounded the Cape without having seen it. He called the spot Angra de São Brás ("Bay of St. Blaise," whose feast day it was) or the Bay of Cowherds, from the people he found there. Dias' Negro companions were unable to understand these people, who fled but later returned to attack the Portuguese. The expedition went on to Angra da Roca (present-day Algoa Bay). The crew was unwilling to continue, and Dias recorded the opinions of all his officers, who were unanimously in favour of returning. They agreed to go on for a few days, reaching Rio do Infante, named after the pilot of "São Pantaleão"; this is almost certainly the present Groot-Vis River.

Sighting of the Cape of Good Hope

It was now clear that India could be reached by the Cape route, and Dias turned back. He sighted the Cape itself in May. Barros says that he named it Cape of Storms and that John II renamed it Cape of Good Hope. Duarte Pacheco, however, attributes the present name to Dias himself, and this is likely since Pacheco joined Dias at the island of Príncipe. Little is known of the return journey, except that Dias touched at Príncipe, the Rio do Resgate (in the present Liberia), and the fortified trading post of Mina. One of Dias' markers, at Padrão de São Gregório, was retrieved from False Island, about 30 miles short of the Great Fish River, in 1938. Another marker formerly stood at the western end of the "Gulf of St. Christopher," now Dias Point.

Nothing is known of Dias' reception by John II. Although plans are said to have been made for a voyage to India, none was attempted for nine years, perhaps pending news of Pêro da Covilhã. John's successor, Manuel I, authorized Vasco da Gama's celebrated voyage of 1497. Bartolomeu Dias accompanied that expedition as far as Mina.

On Gama's return to Portugal, after successfully making contact with the seaports of western India, a further fleet was at once organized; it consisted of a dozen ships and was intended to impress the Indians and to open commerce on a large scale. The whole fleet was under Pedro Álvares Cabral, and Dias was given one of the smaller ships. The fleet sailed far into the western Atlantic on its way to the Cape and sighted land at Espírito Santo in Brazil. Thought to be an island, it was named the Land of the True Cross. Dias thus participated in the discovery of Brazil. He was lost at sea when they reached the Cape, thus perishing in the very waters he had been the first to navigate.

No portrait of Dias is known. He had a son, António, and his grandson, Paulo Dias de Novais, governed Angola and became the founder of the first European city in southern Africa, São Paulo de Luanda, in 1576.

BIBLIOGRAPHY. Sources for Dias are the 16th-century historians: JOAO DE BARROS, GALVAO, and DUARTE PACHECO PEREIRA. See also ERIC AXELSON, *South-East Africa, 1488–1530* (1940).

(H.V.L.)

Dice and Dice Games

Dice are small cubes used as implements for gambling and the playing of social games, each individual cube being called a die. Each side of a standard die is marked with from one to six small dots (spots). The spots are arranged in conventional patterns and placed so that the spots on the opposite sides always total seven: one and six, two and five, three and four. The combinations of the six spots plus the number of dice in play determine the mathematical probabilities.

In most games played with dice, the dice are thrown (rolled, flipped, shot, tossed, or cast) from the hand or from a receptacle called a dice cup, in such a way that they will fall at random. The spots that face upward when the dice come to rest are the deciding spots. The combination of the spots on the topmost surfaces of the dice decides, according to the rules of the game being played, whether the thrower (called the shooter) wins, loses, continues to throw, or loses possession of the dice.

DICE AND THEIR USE: FROM ANCIENT TO MODERN TIMES

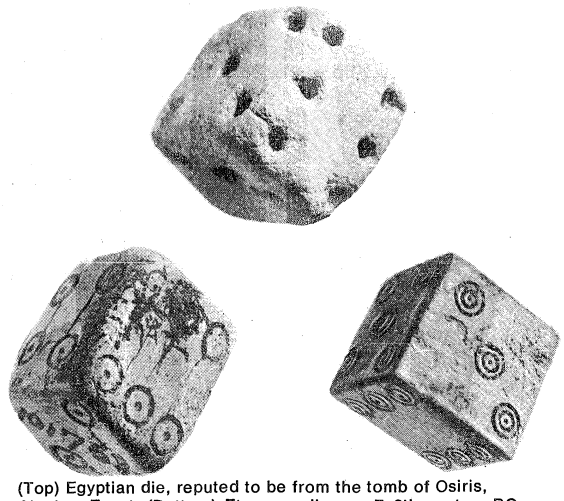
Antiquity of dice

Sophocles reported that dice were invented by one Palamedes, a Greek, during the siege of Troy, while Herodotus maintained that they were invented by the Lydians in the days of King Atys. Both "inventions" have been discredited by numerous archaeological finds demonstrating that dice were used in many earlier societies. Dice, before they became gambling implements, were magical devices that primitive men used in *sortilege*, the casting of lots to divine the future. The probable forerunners of dice were knucklebones (the anklebones of sheep) marked on four faces. In Arabic, the word for knucklebone is the same as that for dice.

Primitive people all over the globe—the North American Indian, the Aztec and Maya, the South Sea islander, the Eskimo, the African—have gambled with dice of many materials and curious shapes and markings. They have used dice made of plum and peach stones, of seeds, of buffalo, caribou, and moose bone, of deer horn, of pebbles, of pottery, of walnut shells, and of beaver and woodchuck teeth. In later Greek and Roman times, although most dice were made of bone and ivory, others were of bronze, agate, rock crystal, onyx, jet, alabaster, marble, amber, porcelain, and other materials.

Man's earliest written records mention not only dice but crooked dice as well. Cubical dice with markings practically equivalent to those of modern dice have been found in Chinese excavations dated as early as 600 BC and in Egyptian tombs dated earlier than 2000 BC. It is in India, more than 2,000 years ago, that the first written records of dice are found in the ancient Sanskrit epic, the *Mahābhārata*.

By courtesy of the Department of Anthropology, Smithsonian Institution, Washington, D.C.



(Top) Egyptian die, reputed to be from the tomb of Osiris, Abydos, Egypt. (Bottom) Etruscan dice, *c.* 7–6th century BC, from Chiusi, Italy. In the Smithsonian Institution, Washington, D.C.

Dice in various forms are the oldest gaming implements known to man, and countless games are and have been played with them. Craps, the most popular gambling-house game, is played with two dice. Chuck-a-Luck, Hazard, and 4,5,6 are played with three dice. In more social play, there are many Poker Dice games played with five dice, and various counter and bar games such as Twenty-six played with ten. In Backgammon and hundreds of board games two or more dice are thrown to determine moves.

Modern dice are almost all made of a cellulose or other plastic material. There are two kinds, perfect or casino dice made by hand and true to a tolerance of 1/5,000 inch, which are used mostly in gambling casinos to play Craps, and round-cornered or imperfect dice called drugstore or candystore dice, which are machine made and are generally used to play social and board games. Modern casino dice are sawed from extruded plastic rods. The spots are drilled approximately 17/1,000 inch into the faces of the die. Then all recesses are filled with a paint the same weight as the plastic that has been drilled out. The dice are then buffed and polished, and since no recesses remain, they are known as flush-spot dice.

Most casinos use red flush-spot dice, which are transparent and come in sets of five; the standard size used in most casinos the world over is 0.750 inch. The dice edges are generally square and known as razor edge, or slightly turned and known as feather edge. Casino dice usually carry the casino's own special monogram and coded serial numbers as a means of thwarting dice cheats. Perfect dice used in various dice games range from a 0.250-inch celluloid or bone "Pee Wee" to an extra large size 0.770 inch. Perfect concave-spot dice, although still in use, are rarely seen in casinos.

Pyramidal, pentahedral, and octahedral dice with all sorts of face designs also are and have been used.

Crooked dice

Dice specially made for cheating have been found in the tombs of ancient Egypt and the Orient and in prehistoric graves of North and South America. Any die that is not a perfect cube will not act according to correct mathematical odds and is called a shape. Shapes are cubes that have been shaved down on one or more sides so that they are slightly brick-shaped and will tend to settle down most often on their larger surfaces. Shapes are the most common of all crooked dice.

Loaded dice when measured with calipers may prove to be perfect cubes; but extra weight just below the surface

on some sides will make the opposite sides come up more often than they should.

Dice with one or more faces each duplicated on its opposite side and certain numbers omitted will tend to produce some numbers in disproportionate frequency and never to produce certain other numbers; for example, two dice marked respectively with duplicates of 3–4–5 and 1–5–6 can never produce combinations totalling 2,3,7 or 12, which are the only combinations with which one can lose in the game of Craps. Such dice, called tops and bottoms, are used as a rule only by accomplished dice cheats, who introduce them into the game by sleight of hand. Since it is impossible to see more than three sides of a cube at any one time, tops and bottoms are unlikely to be detected by the inexperienced gambler.

Cheating at dice is possible with honest dice also. The cheater throws the dice in a manner that causes one or two dice to spin instead of turning over at random so that a predetermined side or number will come to settle skyward.

Amateur cheats, using a blanket or soft surface to play on, roll one or both dice straight ahead in a spoollike fashion so that the two faces at the sides of each die cannot appear at the top—thus preventing certain combinations of numbers from appearing.

In many dice games, a person, usually the operator of the game, wins by no other expedient than offering bets at odds less favourable than those dictated by the theory of probability.

Probabilities

Assuming the use of perfect dice, skill in most dice games consists in knowing the mathematical probability that any total number will appear face up when two or more dice are thrown at random. Ultimate probabilities are subject to the laws of combinations and permutations, but in most dice games they can be ascertained by the application of simple arithmetic. The basic assumption is that any one side on each die is as likely as any other side to appear face up. With a symmetrical die of six sides each side has an equal chance with each of the others, so the expectation is for any one side to be thrown an average of once in six times. Hence its probability is 1/6, which is to say that the odds are 5 to 1 against the appearance of any specified side on any one throw.

With two dice, each of the six sides of one die can be combined with each of the six sides of the other to form 6×6 or 36 combinations. The chance that any combination of two like numbers such as two 6s, two 5s, etc., will appear in one throw is 1 in 36 or 1/36, which is to say that the odds against the appearance of two 6s or any two like numbers in one throw is 35 to 1. When three dice are used, as in Chuck-a-Luck, the total combinations are $6 \times 6 \times 6 = 216$. When four dice are used the total combinations are $6 \times 6 \times 6 \times 6 = 1{,}296$, and so on. The chances of throwing a specific number in one throw of the dice is calculated by comparing the number of favourable chances with the number of unfavourable chances. For example, the only throws that will produce the total 5 with four dice is 1–1–1–2, 1–1–2–1, 1–2–1–1, and 2–1–1–1, so 5 can be expected to appear 4 times in 1,296 throws, making the odds 1,292 to 4, or 323 to 1, against such an event.

GAMES PLAYED WITH DICE

Hazard, sometimes now known as English Hazard, a two-dice game for two or more players, became a mania in London's swank gaming establishments in the 17th and 18th centuries. The name comes from the Spanish *azar*, unlucky throw at dice, misfortune, which, in turn, is derived from the Arabic *az-zahr*, the die. A Chuck-a-Luck variant, Grand Hazard or Three Dice Hazard, is sometimes also called Hazard (see below). The modern game of Craps was derived from the old English game, and its name from a nickname of the losing cast of 1–1 or 1–2, which was called Crabs at least as early as the 16th century.

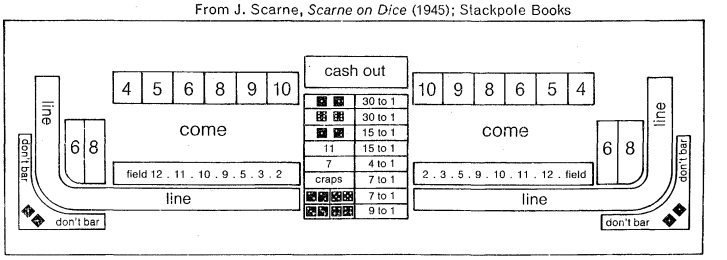
In Hazard, after the bets are made, one player, the shooter, rolls the dice to determine the "main" point (any number from 5 through 9). This done, he then rolls again, winning if he throws the main point or certain other numbers (11 if the main is 7, or 12 if it is 6 or 8), and losing if he throws certain others (crabs to any main; 11 to 5, 6, 8, or 9; or 12 to 7). If he throws any other number from 4 through 10 it becomes his "chance"; he continues to throw the dice until his chance comes up, when he wins, or the main comes up, when he loses. If there is a banker for the game the rules may differ.

Craps is of American origin. According to tradition, some time after 1800, blacks living around New Orleans tried their hands at Hazard, which the English sometimes nicknamed Crabs or Craps. In the course of time they modified the rules and playing procedure so greatly that they ended up inventing the game of Craps (in the U.S. idiom known as Crapshooting or Shooting Craps and here identified as Private Craps to distinguish it from Open Craps and the more formalized variants offered in gambling casinos). In 1907 John H. Winn, a New York City dice maker, became the first Craps bookmaker or banker in history. Winn invented the game of Open Craps, a game in which the players are permitted to bet among themselves or with a bookmaker, who charges a fee. They may bet that the shooter will win or lose, paying, for example, a quarter for a $5 bet and 50 cents for a $10 bet. Shortly afterward, Winn introduced Bank Craps, in which all bets must be made against the house, charging 5 percent of the amount of any bet. Later, other gamblers eliminated the direct charge, substituting shorter house odds, and added more bets to the layout.

Craps not only replaced Faro as the great American banking game but also outdistanced all other casino games in popularity. Since the legalization of gambling in Nevada in 1931, the Las Vegas version of Bank Craps has been introduced into countless casinos the world over, in the Caribbean, Europe, South America, Africa, and Asia. Its spread undoubtedly was encouraged by the belief of the various governments that Las Vegas Craps attracts American tourists. The popularity of the private game of Craps with the U.S. military personnel during World Wars I and II helped to spread that game to many parts of the world. There are four ways to play Craps. Private Craps is a friendly, social game that does not use a casino, Crap table, or banker. The only requisites for Private Craps are two or more persons with cash in their pockets and a pair of dice. It can be played on a street corner, in a back alley, private club, army barracks, living room—anywhere the players have room to roll the dice.

Bank Craps is played on a special table, similar to a billiard table, covered with a "Craps layout" divided into spaces representing the various bets that can be placed. Although there are various layouts, the actual differences among them are small. The layout is planned so that the house has a mathematical advantage on every bet.

From J. Scarne, *Scarne on Dice* (1945); Stackpole Books



Caribbean style double-side dealer layout on a Las Vegas dice table.

Open or Money Craps is a game in which players are permitted to bet among themselves, but a houseman called a banker or bookie is present to accept any bet within the house limit that a player is unable to place with another player. For this privilege, the player must pay the banker a charge, usually 5 percent of the amount wagered. New York Craps is a version of Bank Craps found in gambling houses in the eastern part of the United States, the Bahamas, and in England. The table and the layout differ from the Las Vegas or Bank Craps

table, known as a double-end dealer, and the dealer charges 5 percent of the amount wagered on the point numbers (4,5,6,8,9,10).

Principles of Craps

The following principles apply to all forms of Craps: A standard pair of dice is used. Any number may play. Any player, by consent of the others, may start the game by becoming the first shooter (person throwing the dice).

The dice are thrown by the shooter and the two numbers that face skyward when the dice come to rest, added together, are the deciding numbers. If on the first roll, the shooter throws a natural (7 or 11), it is a winning decision called a pass; if he throws a crap (2,3, or 12) it is a losing decision called a missout; if he throws a 4, 5,6,8,9, or 10, that number becomes his point and he continues throwing until he either throws his point again (for a pass), or throws a 7 (for a missout). When the shooter fails to make his point, the dice pass to the next player on his left and it becomes his turn to shoot. The shooter also has the option of handing the dice to the next player after a pass if he wishes. The next player also may refuse to shoot in his turn and pass the dice on.

Chuck-a-Luck is a very old dice game originally called Sweat-cloth in England and known in the United States, where it appeared about 1800, as Sweat. Later it came to be known as Chucker-Luck and simply Chuck, and, more recently, as the Bird Cage. Chuck-a-Luck has disappeared from many gambling establishments and is played mainly at carnivals and bazaars. The standard equipment is simplicity itself: a cloth layout numbered from 1 to 6 inclusive, a table, and a 20-inch wire cage containing three dice. Players may bet any amount within agreed minimum and maximum limits on one or more of the six numbers. When all bets are down, the wire cage is inverted and the three dice tumble down and come to rest at the bottom of the cage. The faceup numbers of the three dice determine the outcome.

If, for example, one unit was bet on 6, then if there is one 6 up after the cage is inverted, the operator pays even money and will return two units for a net gain of one unit to the bettor; if two 6s are up, the operator will pay 2 to 1, or three units; and if all three dice show a 6, the operator pays 3 to 1, or four units. When one unit is wagered on each of the six numbers, the operator breaks even when three different numbers are thrown, makes a profit of one unit when a pair and a single are thrown, makes a profit of two units when three of a kind are thrown. The operator's overall advantage amounts to 7.87 percent on each unit bet.

Crown and Anchor, a variant of Chuck-a-Luck, played in England, Australia, and especially Canada, uses the same equipment as Chuck-a-Luck except that each of the three dice carries six symbols: crown, anchor, heart, spade, diamond, and club, instead of spots. The layout carries the same symbols. The players place their bets on the symbols on the layout and the banker throws three dice from a dice cup. The payoff and the advantage for the operator are the same as in Chuck-a-Luck.

Grand Hazard, another variant of Chuck-a-Luck, also known as Three-Dice Hazard or simply Hazard, was one of America's most popular gambling house games in the early 1930s. The three dice were dropped through a Hazard chute containing a series of inclined planes that tripped the dice as they fell, or they were thrown from a cup. The layout included, in addition to the six numbers, spaces for bets on high, low, odd, and even numbers, and on three of a kind (triplets or raffles); some layouts permitted bets on specific combinations of the dice.

Barbooth or Barbudi is a two-dice game of Middle Eastern origin. It is still being played in Greece, Turkey, and other countries of the Middle East and North Africa, including the United Arab Republic. In the United States and elsewhere it is played chiefly by persons of Greek, Jewish, and other Middle Eastern ancestry. It is often played in gambling houses for high stakes.

In Barbooth, unlike Craps, the shooter does not specify the amount of his wager. The player on his right sets the stakes, betting that the shooter will not win. The other players may make side bets on whether the shooter or the fader will win. The game, which provides no mathematical advantage for either the shooter or fader, is known as a dead-even game. A house employee known as a cutter takes a percentage of each winning bet. A bookmaker or banker is usually available to accept side bets for a charge, which is divided equally between the bookmaker and house cutter.

The shooter and fader—beginning with the shooter—alternate throwing two pee wee dice from a dice cup until a decision has been achieved. If either the shooter or the fader throws one of four specified combinations (*e.g.*, 3–3 or 5–6) he wins; if he throws one of four others (*e.g.*, 1–1 or 1–2) he loses. All other throws are meaningless. After each decision the shooter retains or passes the dice according to specified winning or losing combinations.

Four-Five-Six, or See Low, is a popular gambling-house game in the U.S. Northwest, Western Canada, and Alaska. Three dice and a dice cup are used. Each player places the amount he desires to wager in front of him. The operator, or banker, then covers all bets and plays against each player in turn. The banker starts the game by throwing the dice from the cup once. If he throws any three of a kind (such as 1–1–1) or any pair plus 6 (such as 1–1–6), or the sequence, 4–5–6, he wins all bets. If he throws 1–2–3 or any pair plus 1 such as 2–2–1), he loses all bets. When any pair is thrown and the third die is a 2,3,4, or 5 (such as 1–1–2), the number on the third die becomes the point number. All combinations except the ones stated are meaningless and the banker throws again. If the banker throws a point number, each other player in turn to the banker's left throws the dice to determine the outcome against the banker. If the player fails to score a winning or losing decision and throws a point, the higher of the banker's or player's points wins. A tie is a standoff or no decision. When a player does not get a pair and does not throw either 4–5–6 or 1–2–3, the roll is meaningless, and he must continue throwing the dice until he wins, loses, or ties. The banker's advantage lies in the fact that there are more winning than losing combinations and he has the first chance to throw them.

One of the most popular U.S. bar and counter games is Poker Dice. It is usually played with five poker dice whose sides bear the playing card denominations: ace, king, queen, jack, ten, and nine, although the usual spotted dice are sometimes used. The ace is also sometimes played wild. Any number can play and each player throws one die to determine the order of play, highest man going first. The object is to throw the highest poker hand in either one or two throws as determined beforehand.

Indian dice, another popular bar and counter game, is played with five dice similar to poker dice, but in this game the player may take as many as three throws to score a hand.

Scarney Dice, a creation of John Scarne, author of *Scarne on Dice* and other works, is a poker-dice game played with a dice cup and five Scarney dice; that is, dice that are marked with the 1,3,4, and 6 spots plus the word "dead" on two opposite sides. (The usual spotted dice also may be used, with the 2 and 5 spots considered "dead.") A player may throw as many times as he wishes, providing he has not thrown five dead dice. After each throw, dead dice are put aside and the remaining dice are thrown again. On any throw made without a dead die, the total points facing up, plus any bonus, are credited to the shooter. Bonuses are awarded on a scale ranging from 10 points for any two pairs to 100 points for any five of a kind.

Yacht is another poker-dice variant, in which five dice are thrown from a cup and a score is kept. Each player in rotation may throw the dice three times in each of the 12 rounds. After each of his first two casts in each round, he may leave standing such dice as satisfy him and throw the rest. Various dice combinations (categories) are assigned scoring values, the highest being yacht (five of a kind) for 50 points. The other categories are big straight, little straight, full house, four of a kind, choice (any five

dice), and sixes, fives, threes, twos, and aces. Before each turn a player must select a category not previously selected.

Liar Dice, which approaches real poker with its element of bluffing, is popular in the U.S. Army. The shooter shakes five dice in a cup, turns the cup upside down and lifts it, shielding the dice from view with his hand. He then announces the best poker hand he can make, but his announcement need not be true. His opponent may either roll for a higher hand, or challenge the announcement by saying, "Liar," in which case, the player must reveal his dice. He wins if his announcement was correct, loses if it was not.

Aces is one of the most popular dice games played in the Far East by U.S. military personnel. Any number may play, but each player must have a dice cup and five dice. Each player puts an agreed stake into a pot and then throws his five dice—the player with the highest poker hand becoming the first shooter, the player throwing the second highest sits on his left and shoots second and so on; tying players throw again.

The first shooter begins by throwing his five dice. Each 1 is placed in the centre of the table; all 2s are passed to the player on his left; all 5s to the player on his right. The player continues to throw until he fails to throw a 1, a 2, or a 5, or until he has no dice left. The player on the left then begins his throw. Players with no dice are still in the game as they may receive dice from the players on either side of them. When all the dice but one have been placed in the centre of the table, the player throwing the last 1 with the last die is the winner and takes the pot.

In Buck Dice, any number can play and three dice are used. Each player throws the dice to determine the order of play, with the highest total first. The low man then throws one die and the number thrown becomes the point. The starter throws all three dice, scoring one point for each point number thrown. He continues to throw as long as he throws point numbers, which are added as he goes along. When he fails to throw a point number on any throw, the dice pass to the next player. The object is to score exactly 15 points, called buck or game. If a throw puts a player's score above 15, the throw does not count and he must throw again. As each player reaches that score he drops out until only one player remains; he is the loser and pays everyone a predetermined amount. Any three of a kind (not point numbers) is a little buck and counts five points. The point number on all three dice is big buck or the general, which counts 15 points and eliminates the player no matter what score he has previously made.

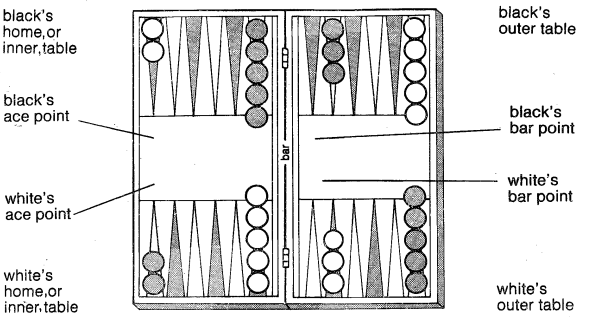
High Dice is a two-dice game also called Beat the Banker and Counter Klondike in which the banker and player each throw once, the banker first. The player must throw a higher number than the banker to win. The banker takes all ties.

In the 1950s 26 was a favourite counter game in many stores and taverns throughout the U.S. Midwest, in which customers would play for a fee such as 25 cents to win checks or tokens that could be cashed for drinks, cigarettes, or other merchandise. Federal and state anti-gambling drives during the late 1950s eliminated most of these games. A high rail dice board, score pad, dice cup, and ten dice are used. The player selects any number from 1 through 6 as his point. He throws the ten dice 13 times and totals the number of times he has thrown his point number. The object is to throw 26 or more point numbers.

Backgammon

Backgammon is the parent of all track board games and perhaps the oldest game with dice still being played. The French and Germans call it Tric-Trac, the Italians *Tavola reale* (royal table). The Romans who played it with three dice knew it as *ludus duodecim scripta* (game of the 12 lines). The derivation of the word Backgammon has been ascribed both to the Welsh words *back* and *gammon* (little battle) and to the Saxon *bac* and *gamen* (back game).

Two persons play and each has a pair of dice and a cup plus one doubling die. There are 30 playing counters called men, 15 of one colour and 15 of another (usually black and white). The players sit on opposite sides of a board that contains 24 alternately coloured triangles called points, 12 opposite each player. A bar divides the board into two equal rectangles one of which, by agreement, is designated the inner, or home, table, and the other the outer table. The placement of the men on the backgammon board to start the game is shown in the figure below.



Modern Backgammon board at the beginning of play.

Each player throws one die from the cup. The player throwing the higher number plays first by using both his and his opponent's numbers. Thereafter, each player throws his own two dice. The numbers on the two dice, taken separately, show the number of points over which the men may be moved. When one man has been moved the number of points indicated by one die, the other number may be used either to move the same man farther or to move another man. The player must try to use both numbers. If he can use only one number he must use the higher number of the two if possible. When both dice show the same number, the player has thrown doublets and each number is played twice. Therefore if a double 3 is thrown it would be played the same as four 3s.

A single man on a point is a blot. When a man of the opposite colour lands on that point the blot is hit and the man that was originally there is taken off and placed on the bar. A man on the bar may enter again when a number is thrown on the dice that will place the man on an unoccupied point of his opponent's home table. All men on the bar must be re-entered before any other moves may be made.

When a player has two or more men on a point, it is said to be blocked and an opponent's man may not come to rest on that point although it can move past the blocked point. If a player blocks six adjacent points and has one or more of his opponent's men behind it, he has made a prime. If he blocks all six points of his home table when his opponent still has one or more men on the bar, he is said to have a shutout.

When a players's men have all reached his own home table, he begins removing them from the board. This is called bearing or bearing off the men. For each number thrown a man is removed from the corresponding point. On a throw of 5–3 for instance, one man is removed from the 5 and one from the 3 point. A throw of doublets allows the player to remove twice as many men. On a throw of double 3, four men may be taken from the 3 point. When the number thrown is higher than any point on which there are men, the player may remove a man from the next highest point. If low numbers are thrown and there are no men located on those points, men from higher points are moved down toward the one point according to the numbers thrown on the dice. Men may also be moved down instead of being borne off if desired.

The first player to bear off all his men wins a single game. If he bears off all men before his opponent has borne off any, he wins a double game or gammon and collects double stakes. If the loser has not borne off any men and still has one or more men in his opponent's

home table or on the bar, the winner has scored a triple game or backgammon and collects triple stakes.

At any time after his opponent has finished a throw a player may double the stake. If his opponent rejects the double, he forfeits the games and the stake. The privilege of making the next double falls to his opponent. There is no limit to the number of doubles that may be made but the option of offering doubles alternates between players.

Acey Deucey is the favourite Backgammon game of the U.S. Navy, Marine Corps, and Merchant Marine. The equipment and rules of play are similar to Backgammon except that the pieces (men) are not placed on the board at the start of the game but must be played onto the board, and when ace-deuce is thrown the player moves one man three spaces, or one man one space and another two spaces, and then also has the choice of any doublets and also may take an additional throw of the dice.

In Russian Backgammon, no men are placed upon the board at the start, but each player enters his men by throws of the dice—both players enter in the same home table and both move in the same direction around the board to the opposite table. Bearing off is the same as in Backgammon. After having entered two or more men, a player is at liberty either to continue entering his men with any subsequent throws, or to move the men already entered. When a blot is hit, the owner must re-enter it before he makes any other move. Except on a player's first throw of the game, doublets are used twice over. He not only can play the upper faces of the dice twice over, but the bottom (opposite) faces as well; he also can throw again before his opponent. After each thrown doublet the player continues to throw until he fails to throw a doublet, in which case he plays the numbers thrown and the throw passes to his opponent.

BIBLIOGRAPHY. JOHN SCARNE and CLAYTON RAWSON, *Scarne on Dice*, 8th ed. (1962), the definitive work; JOHN SCARNE, *Scarney Dice* (1969).

(J.S.)

Dickens, Charles

Generally regarded as the greatest English novelist, Charles Dickens enjoyed a wider popularity than any previous author had done during his lifetime. Much in his work could appeal to simple and sophisticated, to the poor and to the Queen, and technological developments as well as the qualities of his work enabled his fame to spread worldwide very quickly. His long career saw fluctuations in the reception and sales of individual novels, but none of them was negligible or uncharacteristic or disregarded, and, though he is now admired for aspects and phases of his work that were given less weight by his contemporaries, his popularity has never ceased and his present critical standing is higher than ever before. The most abundantly comic of English authors, he was much more than a great entertainer. The range, compassion, and intelligence of his apprehension of his society and its shortcomings enriched his novels and made him both one of the great forces in 19th-century literature and an influential spokesman of the conscience of his age.

EARLY YEARS

Charles John Huffam Dickens was born February 7, 1812, in Portsmouth but left it in infancy. His happiest childhood years were spent in Chatham (1817–22), an area to which he often reverts in his fiction. From 1822 he lived in London, until, in 1860, he moved permanently to a country house, Gad's Hill, near Chatham. His origins were middle class, if of a newfound and precarious respectability; one grandfather had been a domestic servant, and the other an embezzler. His father, a clerk in the navy pay office, was well paid, but his extravagance and ineptitude often brought the family to financial embarrassment or disaster. (Some of his failings and his ebullience are dramatized in Mr. Micawber in the partly autobiographical *David Copperfield*.) In 1824 the family reached bottom. Charles, the eldest son, had been withdrawn from school and was now set to manual work in a factory, and his father went to prison for debt. These shocks deeply affected Charles. Though abhorring this brief descent into the working class, he began to gain that sympathetic knowledge of their life and privations that informed his writings. Also, the images of the prison and of the lost, oppressed, or bewildered child recur in many novels. Much else in his character and art stems from this period, including, as the 20th-century novelist Angus Wilson has argued, his later difficulty, as man and author, in understanding women: this may be traced to his bitter resentment against his mother, who had, he felt, failed disastrously at this time to appreciate his sufferings. She had wanted him to stay at work when his father's release from prison and an improvement in the family's fortunes made the boy's return to school possible. Happily the father's view prevailed.



Dickens, 1859.
By courtesy of the Gernsheim Collection, the University of Texas at Austin

Factory experiences

His schooling, interrupted and unimpressive, ended at 15. He became a clerk in a solicitor's office, then a shorthand reporter in the lawcourts (thus gaining a knowledge of the legal world often used in the novels), and finally, like other members of his family, a parliamentary and newspaper reporter. These years left him with a lasting affection for journalism and contempt both for the law and for Parliament. His coming to manhood in the reformist 1830s, and particularly his working on the Liberal Benthamite *Morning Chronicle* (1834–36), greatly affected his political outlook. Another influential event now was his rejection as suitor to Maria Beadnell because his family and prospects were unsatisfactory; his hopes of gaining and chagrin at losing her sharpened his determination to succeed. His feelings about Maria then and at her later brief and disillusioning re-entry into his life are reflected in David Copperfield's adoration of Dora Spenlow and in the middle-aged Arthur Clennam's discovery (in *Little Dorrit*) that Flora Finching, who had seemed enchanting years ago, was "diffuse and silly," that Flora "whom he had left a lily, had become a peony."

Beginning of literary career. Much drawn to the theatre, Dickens nearly became a professional actor in 1832. In 1833 he began contributing stories and descriptive essays to magazines and newspapers; these attracted attention and were reprinted as *Sketches by "Boz"* (February 1836). The same month, he was invited to provide a comic serial narrative to accompany engravings by a well-known artist; seven weeks later the first installment of *Pickwick Papers* appeared. Within a few months *Pickwick* was the rage and Dickens the most popular author of the day. During 1836 he also wrote two plays and a pamphlet on a topical issue (how the poor should be allowed to enjoy the Sabbath) and, resigning from his newspaper job, undertook to edit a monthly magazine, *Bentley's Miscellany*, in which he serialized *Oliver Twist*

Early stories and essays

(1837–39). Thus, he had two serial installments to write every month. Already the first of his nine surviving children had been born; he had married (in April 1836) Catherine, eldest daughter of a respected Scottish journalist and man of letters, George Hogarth.

For several years his life continued at this intensity. Finding serialization congenial and profitable, he repeated the *Pickwick* pattern of 20 monthly parts in *Nicholas Nickleby* (1838–39); then he experimented with shorter weekly installments for *The Old Curiosity Shop* (1840–41) and *Barnaby Rudge* (1841). Exhausted at last, he then took a five-month vacation in America, touring strenuously and receiving quasi-royal honours as a literary celebrity but offending national sensibilities by protesting against the absence of copyright protection. A radical critic of British institutions, he had expected more from "the republic of my imagination," but he found more vulgarity and sharp practice to detest than social arrangements to admire. Some of these feelings appear in *American Notes* (1842) and *Martin Chuzzlewit* (1843–44).

American tour

First novels. His writing during these prolific years was remarkably various and, except for his plays, resourceful. *Pickwick* began as high-spirited farce and contained many conventional comic butts and traditional jokes; like other early works, it was manifestly indebted to the contemporary theatre, the 18th-century English novelists, and a few foreign classics, notably *Don Quixote*. But, besides giving new life to old stereotypes, *Pickwick* displayed, if sometimes in embryo, many of the features that were to be blended in varying proportions throughout his fiction: attacks, satirical or denunciatory, on social evils and inadequate institutions; topical references; an encyclopaedic knowledge of London (always his predominant fictional locale); pathos; a vein of the macabre; a delight in the demotic joys of Christmas; a pervasive spirit of benevolence and geniality; inexhaustible powers of character creation; a wonderful ear for characteristic speech, often imaginatively heightened; a strong narrative impulse; and a prose style that, if here overdependent on a few comic mannerisms, was highly individual and inventive. Rapidly improvised and written only weeks or days ahead of its serial publication, *Pickwick* contains weak and jejune passages and is an unsatisfactory whole—partly because Dickens was rapidly developing his craft as a novelist while writing and publishing it. What is remarkable is that a first novel, written in such circumstances, not only established him overnight and created a new tradition of popular literature but also survived, despite its crudities, as one of the best known novels in the world.

Popularity of *Pickwick Papers*

His self-assurance and artistic ambitiousness had appeared in *Oliver Twist*, where he rejected the temptation to repeat the successful *Pickwick* formula. Though containing much comedy still, *Oliver Twist* is more centrally concerned with social and moral evil (the workhouse and the criminal world); it culminates in Bill Sikes's murdering Nancy and Fagin's last night in the condemned cell at Newgate. The latter episode was memorably depicted in George Cruikshank's engraving; the imaginative potency of Dickens' characters and settings owes much, indeed, to his original illustrators (Cruikshank for *Sketches by "Boz"* and *Oliver Twist*, "Phiz" [Hablot K. Browne] for most of the other novels until the 1860s). The currency of his fiction owed much, too, to its being so easy to adapt into effective stage versions. Sometimes 20 London theatres simultaneously were producing adaptations of his latest story; so even nonreaders became acquainted with simplified versions of his works. The theatre was often a subject of his fiction, too, as in the Crummles troupe in *Nicholas Nickleby*. This novel reverted to the *Pickwick* shape and atmosphere, though the indictment of the brutal Yorkshire schools (Dotheboys Hall) continued the important innovation in English fiction seen in *Oliver Twist*—the spectacle of the lost or oppressed child as an occasion for pathos and social criticism. This was amplified in *The Old Curiosity Shop*, where the death of Little Nell was found overwhelmingly powerful at the time, though a few decades later it became a byword for "Victorian sentimentality." In *Barnaby Rudge* he attempted another genre, the historical novel. Like his later attempt in this kind, *A Tale of Two Cities*, it was set in the late 18th century and presented with great vigour and understanding (and some ambivalence of attitude) the spectacle of large-scale mob violence.

To create an artistic unity out of the wide range of moods and materials included in every novel, with often several complicated plots involving scores of characters, was made even more difficult by Dickens' writing and publishing them serially. In *Martin Chuzzlewit* he tried "to resist the temptation of the current Monthly Number, and to keep a steadier eye upon the general purpose and design" (1844 Preface). Its American episodes had, however, been unpremeditated (he suddenly decided to boost the disappointing sales by some America-baiting and to revenge himself against insults and injuries from the American press). A concentration on "the general purpose and design" was more effective in the next novel, *Dombey and Son* (1846–48), though the experience of writing the shorter, and unserialized, Christmas books had helped him obtain greater coherence.

A Christmas Carol (1843), suddenly conceived and written in a few weeks, was the first of these Christmas books (a new literary genre thus created incidentally). Tossed off while he was amply engaged in writing *Chuzzlewit*, it was an extraordinary achievement—the one great Christmas myth of modern literature. His view of life was later to be described or dismissed as "Christmas philosophy," and he himself spoke of "*Carol* philosophy" as the basis of a projected work. His "philosophy," never very elaborated, involved more than wanting the Christmas spirit to prevail throughout the year, but his great attachment to Christmas (in his family life as well as his writings) is indeed significant and has contributed to his popularity. "Dickens dead?" exclaimed a London costermonger's girl in 1870. "Then will Father Christmas die too?"—a tribute both to his association with Christmas and to the mythological status of the man as well as in his work. The *Carol* immediately entered the general consciousness; Thackeray, in a review, called it "a national benefit, and to every man and woman who reads it a personal kindness." Further Christmas books, essays, and stories followed annually (except in 1847) through 1867. None equalled the *Carol* in potency, though some achieved great immediate popularity. Cumulatively they represent a celebration of Christmas attempted by no other great author.

How he struck his contemporaries in these early years appears in R.H. Horne's *New Spirit of the Age* (1844). Dickens occupied the first and longest chapter, as

> . . . manifestly the product of his age . . . . a genuine emanation from its aggregate and entire spirit. . . . He mixes extensively in society, and continually. Few public meetings in a benevolent cause are without him. He speaks effectively. . . . His influence upon his age is extensive—pleasurable, instructive, healthy, reformatory . . . .
>
> Mr. Dickens is, in private, very much what might be expected from his works. . . . His conversation is genial . . . [He] has singular personal activity, and is fond of games of practical skill. He is also a great walker, and very much given to dancing Sir Roger de Coverley. In private, the general impression of him is that of a first-rate practical intellect, with "no nonsense" about him.

Status as a public figure

He was indeed very much a public figure, actively and centrally involved in his world, and a man of confident presence. He was reckoned the best after-dinner speaker of the age; other superlatives he attracted included his having been the best shorthand reporter on the London press and his being the best amateur actor on the stage. Later he became one of the most successful periodical editors and the finest dramatic recitalist of the day. He was splendidly endowed with many skills. "Even irrespective of his literary genius," wrote an obituarist, "he was an able and strong-minded man, who would have succeeded in almost any profession to which he devoted himself" (*Times*, June 10, 1870). Few of his extraliterary skills and interests were irrelevant to the range and mode of his fiction.

Privately in these early years, he was both domestic and social. He loved home and family life and was a proud and efficient householder; he once contemplated writing a cookbook. To his many children, he was a devoted and delightful father, at least while they were young; relations with them proved less happy during their adolescence. Apart from periods in Italy (1844–45) and Switzerland and France (1846–47), he still lived in London, moving from an apartment in Furnival's Inn to larger houses as his income and family grew. Here he entertained his many friends, most of them popular authors, journalists, actors, or artists, though some came from the law and other professions or from commerce and a few from the aristocracy. Some friendships dating from his youth endured to the end, and, though often exasperated by the financial demands of his parents and other relatives, he was very fond of some of his family and loyal to most of the rest. Some literary squabbles came later, but he was on friendly terms with most of his fellow authors, of the older generation as well as his own. Necessarily solitary while writing and during the long walks (especially through the streets at night) that became essential to his creative processes, he was generally social at other times. He enjoyed society that was unpretentious and conversation that was genial and sensible but not too intellectualized or exclusively literary. High society he generally avoided, after a few early incursions into the great houses; he hated to be lionized or patronized.

He had about him "a sort of swell and overflow as of a prodigality of life," an American journalist said. Everyone was struck by the brilliance of his eyes and his smart, even dandyish, appearance ("I have the fondness of a savage for finery," he confessed). John Forster, his intimate friend and future biographer, recalled him at the *Pickwick* period:

> the quickness, keenness, and practical power, the eager, restless, energetic outlook on each several feature [of his face] seemed to tell so little of a student or writer of books, and so much of a man of action and business in the world. Light and motion flashed from every part of it.

He was proud of his art and devoted to improving it and using it to good ends (his works would show, he wrote, that "Cheap Literature is not behind-hand with the Age, but holds its place, and strives to do its duty"), but his art never engaged all his formidable energies. He had no desire to be narrowly literary.

Unsuccessful journalistic ventures

A notable, though unsuccessful, demonstration of this was his being founder-editor in 1846 of the *Daily News* (soon to become the leading Liberal newspaper). His journalistic origins, his political convictions and readiness to act as a leader of opinion, and his wish to secure a steady income independent of his literary creativity and of any shifts in novel readers' tastes made him attempt or plan several periodical ventures in the 1840s. The return to daily journalism soon proved a mistake—the biggest fiasco in a career that included few such misdirections or failures. A more limited but happier exercise of his practical talents began soon afterward: more than a decade he directed, energetically and with great insight and compassion, a reformatory home for young female delinquents, financed by his wealthy friend Angela Burdett-Coutts. The benevolent spirit apparent in his writings often found practical expression in his public speeches, fund-raising activities, and private acts of charity.

Dombey and Son (1846–48) was a crucial novel in his development, a product of more thorough planning and maturer thought and the first in which "a pervasive uneasiness about contemporary society takes the place of an intermittent concern with specific social wrongs" (Kathleen Tillotson). Using railways prominently and effectively, it was very up-to-date, though the questions posed included such perennial moral and religious challenges as are suggested by the child Paul's first words in the story: "Papa, what's money?" Some of the corruptions of money and pride of place and the limitations of "respectable" values are explored, virtue and human decency being discovered most often (as elsewhere in Dickens) among the poor, humble, and simple. In Paul's early death Dickens offered another famous pathetic episode; in Mr. Dombey he made a more ambitious attempt than before at serious and internal characterization. *David Copperfield* (1849–50) has been described as a "holiday" from these larger social concerns and most notable for its childhood chapters, "an enchanting vein which he had never quite found before and which he was never to find again" (Edmund Wilson). Largely for this reason and for its autobiographical interest, it has always been among his most popular novels and was Dickens' own "favourite child." It incorporates material from the autobiography he had recently begun but soon abandoned and is written in the first person, a new technique for him. David differs from his creator in many ways, however, though Dickens uses many early experiences that had meant much to him—his period of work in the factory while his father was jailed, his schooling and reading, his passion for Maria Beadnell, and (more cursorily) his emergence from parliamentary reporting into successful novel writing. In Micawber the novel presents one of the "Dickens characters" whose imaginative potency extends far beyond the narratives in which they figure; Pickwick and Sam Weller, Mrs. Gamp and Mr. Pecksniff, and Scrooge are some others.

Autobiographical interest of *David Copperfield*

MIDDLE YEARS

Journalism. Dickens' journalistic ambitions at last found a permanent form in *Household Words* (1850–59) and its successor, *All the Year Round* (1859–88). Popular weekly miscellanies of fiction, poetry, and essays on a wide range of topics, these had substantial and increasing circulations, reaching 300,000 for some of the Christmas numbers. Dickens contributed some serials—the lamentable *Child's History of England* (1851–53), *Hard Times* (1854), *A Tale of Two Cities* (1859), and *Great Expectations* (1860–61)—and essays, some of which were collected in *Reprinted Pieces* (1858) and *The Uncommercial Traveller* (1861, later amplified). Particularly in 1850–52 and during the Crimean War, he contributed many items on current political and social affairs; in later years he wrote less—much less on politics—and the magazine was less political, too. Other distinguished novelists contributed serials, including Mrs. Gaskell, Wilkie Collins, Charles Reade, and Bulwer Lytton. The poetry was uniformly feeble; Dickens was imperceptive here. The reportage, often solidly based, was bright (sometimes painfully so) in manner. His conduct of these weeklies shows his many skills as editor and journalist but also some limitations in his tastes and intellectual ambitions. The contents are revealing in relation to his novels: he took responsibility for all the opinions expressed (for articles were anonymous) and selected and amended contributions accordingly; so comments on topical events and so on may generally be taken as representing his opinions, whether or not he wrote them. No English author of comparable status has devoted 20 years of his maturity to such unremitting editorial work, and the weeklies' success was due not only to his illustrious name but also to his practical sagacity and sustained industry. Even in his creative work, as his eldest son said,

> No city clerk was ever more methodical or orderly than he; no humdrum, monotonous, conventional task could ever have been discharged with more punctuality, or with more businesslike regularity.

Novels. The novels of these years, *Bleak House* (1852–53), *Hard Times* (1854), and *Little Dorrit* (1855–57), were much "darker" than their predecessors. Presenting a remarkably inclusive and increasingly sombre picture of contemporary society, they were inevitably often seen at the time as fictionalized propaganda about ephemeral issues. They are much more than this, though it is never easy to state how Dickens' imagination transforms their many topicalities into an artistically coherent vision that transcends their immediate historical context. Similar questions are raised by his often basing fictional characters, places, and institutions on actual originals. He once spoke of his mind's taking "a fanciful photograph" of a scene, and there is a continual interplay between photographic realism and "fancy" (or imagination). "He describes London like a special correspondent

Themes of the later novels

for posterity" (Walter Bagehot, 1858), and posterity has certainly found in his fiction the response of an acute, knowledgeable, and concerned observer to the social and political developments of "the moving age." In the novels of the 1850s, he is politically more despondent, emotionally more tragic. The satire is harsher, the humour less genial and abundant, the "happy endings" more subdued than in the early fiction. Technically, the later novels are more coherent, plots being more fully related to themes, and themes being often expressed through a more insistent use of imagery and symbols (grim symbols, too, such as the fog in *Bleak House* or the prison in *Little Dorrit*). His art here is more akin to poetry than to what is suggested by the photographic or journalistic comparisons. "Dickensian" characterization continues in the sharply defined and simplified grotesque or comic figures, such as Chadband in *Bleak House* or Mrs. Sparsit in *Hard Times*, but large-scale figures of this type are less frequent (the Gamps and Micawbers belong to the first half of his career). Characterization also has become more subordinate to "the general purpose and design"; moreover, Dickens is presenting characters of greater complexity, who provoke more complex responses in the reader (William Dorrit, for instance). Even the juvenile leads, who had usually been thinly conceived conventional figures, are now often more complicated in their make-up and less easily rewarded by good fortune. With his secular hopes diminishing, Dickens becomes more concerned with "the great final secret of all life"—a phrase from *Little Dorrit*, where the spiritual dimension of his work is most overt. Critics disagree as to how far so worldly a novelist succeeds artistically in enlarging his view to include the religious. These novels, too, being manifestly an ambitious attempt to explore the prospects of humanity at this time, raise questions, still much debated, about the intelligence and profundity of his understanding of society.

Personal unhappiness. Dickens' spirits and confidence in the future had indeed declined: 1855 was "a year of much unsettled discontent for him," his friend Forster recalled, partly for political reasons (or, as Forster hints, his political indignation was exacerbated by a "discontent" that had personal origins). The Crimean War, besides exposing governmental inefficiency, was distracting attention from the "poverty, hunger, and ignorant desperation" at home. In *Little Dorrit*, "I have been blowing off a little of indignant steam which would otherwise blow me up . . .," he wrote, "but I have no present political faith or hope—not a grain." Not only were the present government and Parliament contemptible but "representative government is become altogether a failure with us, . . . the whole thing has broken down . . . and has no hope in it." Nor had he a coherent alternative to suggest. This desperation coincided with an acute state of personal unhappiness. The brief tragicomedy of Maria Beadnell's re-entry into his life, in 1855, finally destroyed one nostalgic illusion and also betrayed a perilous emotional immaturity and hunger. He now openly identified himself with some of the sorrows dramatized in the adult David Copperfield:

Marital grief

> Why is it, that as with poor David, a sense comes always crushing on me, now, when I fall into low spirits, as of one happiness I have missed in life, and one friend and companion I have never made?

This comes from the correspondence with Forster in 1854–55, which contains the first admissions of his marital unhappiness; by 1856 he is writing, "I find the skeleton in my domestic closet is becoming a pretty big one"; by 1857–58, as Forster remarks, an "unsettled feeling" had become almost habitual with him, "and the satisfactions which home should have supplied, and which indeed were essential requirements of his nature, he had failed to find in his home." From May 1858, Catherine Dickens lived apart from him. A painful scandal arose, and Dickens did not act at this time with tact, patience, or consideration. The affair disrupted some of his friendships and narrowed his social circle, but surprisingly it seems not to have damaged his popularity with the public.

Catherine Dickens maintained a dignified silence, and most of Dickens' family and friends, including his official biographer, Forster, were discreetly reticent about the separation. Not until 1939 did one of his children (Katey), speaking posthumously through conversations recorded by a friend, offer a candid inside account. It was discreditable to him, and his self-justifying letters must be viewed with caution. He there dated the unhappiness of his marriage back to 1838, attributed to his wife various "peculiarities" of temperament (including her sometimes labouring under "a mental disorder"), emphatically agreed with her (alleged) statement that "she felt herself unfit for the life she had to lead as my wife," and maintained that she never cared for the children nor they for her. In more temperate letters, where he acknowledged her "amiable and complying" qualities, he simply and more acceptably asserted that their temperaments were utterly incompatible. She was, apparently, pleasant but rather limited; such faults as she had were rather negative than positive, though family tradition from a household that knew the Dickenses well speaks of her as "a whiney woman" and as having little understanding of, or patience with, the artistic temperament.

Friendship with Ellen Ternan

Dickens' self-justifying letters lack candour in omitting to mention Ellen Ternan, an actress 27 years his junior, his passion for whom had precipitated the separation. Two months earlier he had written more frankly to an intimate friend:

> The domestic unhappiness remains so strong upon me that I can't write, and (waking) can't rest, one minute. I have never known a moment's peace or content, since the last night of The Frozen Deep.

The Frozen Deep was a play in which he and Nelly (as Ellen was called) had performed together in August 1857. She was an intelligent girl, of an old theatrical family; reports speak of her as having "a pretty face and well-developed figure"—or "passably pretty and not much of an actress." She left the stage in 1860; after Dickens' death she married a clergyman and helped him run a school. The affair was hushed up until the 1930s, and evidence about it remains scanty, but every addition confirms that Dickens was deeply attached to her and that their relationship lasted until his death. It seems likely that she became his mistress, though probably not until the 1860s; assertions that a child, or children, resulted remain unproved. Similarly, suggestions that the anguish experienced by some of the lovers in the later novels may reflect Dickens' own feelings remain speculative. It is tempting, indeed, to associate Nelly with some of their heroines (who are more spirited and complex, less of the "legless angel," than most of their predecessors), especially as her given names, Ellen Lawless, seem to be echoed by those of heroines in the three final novels—Estella, Bella, and Helena Landless—but nothing definite is known about how she responded to Dickens, what she felt for him at the time, or how close any of these later love stories were to aspects or phases of their relationship.

"There is nothing very remarkable in the story," commented one early transmitter of it, and this seems just. Many middle-aged men feel an itch to renew their emotional lives with a pretty young girl, even if, unlike Dickens, they cannot plead indulgence for "the wayward and unsettled feeling which is part (I suppose) of the tenure on which one holds an imaginative life." But the eventual disclosure of this episode caused surprise, shock, or piquant satisfaction, being related of a man whose rebelliousness against his society had seemed to take only impeccably reformist shapes. A critic in 1851, listing the reasons for his unique popularity, had cited "above all, his deep reverence for the household sanctities, his enthusiastic worship of the household gods." After these disclosures he was, disconcertingly or intriguingly, a more complex man; and, partly as a consequence, Dickens the novelist also began to be seen as more complex, less conventional, than had been realized. The stimulus was important, though Nelly's significance, biographically and critically, has proved far from inexhaustible.

Public readings. In the longer term, Kathleen Tillotson's remark is more suggestive: "his lifelong love-affair

with his reading public, when all is said, is by far the most interesting love-affair of his life." This took a new form, about the time of Dickens' separation from his wife, in his giving public readings from his works, and it is significant that, when trying to justify this enterprise as certain to succeed, he referred to "that particular relation (personally affectionate and like no other man's) which subsists between me and the public." The remark suggests how much Dickens valued his public's affection, not only as a stimulus to his creativity and a condition for his commercial success but also as a substitute for the love he could not find at home. He had been toying with the idea of turning paid reader since 1853, when he began giving occasional readings in aid of charity. The paid series began in April 1858, the immediate impulse being to find some energetic distraction from his marital unhappiness. But the readings drew on more permanent elements in him and his art: his remarkable histrionic talents, his love of theatricals and of seeing and delighting an audience, and the eminently performable nature of his fiction. Moreover, he could earn more by reading than by writing, and more certainly; it was easier to force himself to repeat a performance than create a book.

Histrionic talents

His initial repertoire consisted entirely of Christmas books but was soon amplified by episodes from the novels and magazine Christmas stories. A performance usually consisted of two items; of the 16 eventually performed, the most popular were "The Trial from *Pickwick*" and the *Carol*. Comedy predominated, though pathos was important in the repertoire, and horrifics were startlingly introduced in the last reading he devised, "Sikes and Nancy," with which he petrified his audiences and half killed himself. Intermittently, until shortly before his death, he gave seasons of readings in London and hardworking tours through the provinces and (in 1867–68) the United States. Altogether he performed about 471 times. He was a magnificent performer, and important elements in his art—the oral and dramatic qualities—were demonstrated in these renderings. His insight and skill revealed nuances in the narration and characterization that few readers had noticed. Necessarily, such extracts or short stories, suitable for a two-hour entertainment, excluded some of his larger and deeper effects—notably, his social criticism and analysis—and his later novels were underrepresented. Dickens never mentions these inadequacies. He manifestly enjoyed the experience until, near the end, he was becoming ill and exhausted. He was writing much less in the 1860s. It is debatable how far this was because the readings exhausted his energies, while providing the income, creative satisfaction, and continuous contact with an audience that he had formerly obtained through the novels. He gloried in his audiences' admiration and love. Some friends thought this too crude a gratification, too easy a triumph, and a sad declension into a lesser and ephemeral art. However the episode is judged, it was characteristic of him—of his relationship with his public, his business sense, his stamina, his ostentatious display of supplementary skills: also of his originality, for no important author (at least since Homer, as reviewers said) and no English author since of anything like his stature has devoted so much time and energy to this activity. The only comparable figure is his contemporary, Mark Twain, who acknowledged Dickens as the pioneer.

LAST YEARS

Final novels. Tired and ailing though he was, he remained inventive and adventurous in his final novels. *A Tale of Two Cities* (1859) was an experiment, relying less than before on characterization, dialogue, and humour. An exciting and compact narrative, it lacks too many of his strengths to count among his major works. Sydney Carton's self-sacrifice was found deeply moving by Dickens and by many readers; Dr. Manette now seems a more impressive achievement in serious characterization. The French Revolution scenes are vivid, if superficial in historical understanding. *Great Expectations* (1860–61) resembles *Copperfield* in being a first-person narration and in drawing on parts of Dickens' personality and experience. Compact like its predecessor, it lacks the panoramic inclusiveness of *Bleak House*, *Little Dorrit*, and *Our Mutual Friend*, but, though not his most ambitious, it is his most finely achieved novel. The hero Pip's mind is explored with great subtlety, and his development through a childhood and youth beset with hard tests of character is traced critically but sympathetically. Various "great expectations" in the book proved ill founded—a comment as much on the values of the age as on the characters' weaknesses and misfortunes. *Our Mutual Friend* (1864–65), a large inclusive novel, continues this critique of monetary and class values. London is now grimmer than ever before, and the corruption, complacency, and superficiality of "respectable" society are fiercely attacked. Many new elements are introduced into Dickens' fictional world, but his handling of the old comic-eccentrics (such as Boffin, Wegg, and Venus) is sometimes tiresomely mechanical. How the unfinished *Edwin Drood* (1870) would have developed is uncertain. Here again Dickens left panoramic fiction to concentrate on a limited private action. The central figure was evidently to be John Jasper, whose eminent respectability as a cathedral organist was in extreme contrast to his haunting low opium dens and, out of violent sexual jealousy, murdering his nephew. It would have been his most elaborate treatment of the themes of crime, evil, and psychological abnormality that had recurred throughout his novels; a great celebrator of life, he was also obsessed with death.

Change in Dickens' personality

How greatly Dickens personally had changed appears in remarks by friends who met him again, after many years, during the American reading tour in 1867–68. "I sometimes think . . .," wrote one, "I must have known two individuals bearing the same name, at various periods of my own life." But just as the fiction, despite many developments, still contained many stylistic and narrative features continuous with the earlier work, so, too, the man remained a "human hurricane," though he had aged considerably, his health had deteriorated, and his nerves had been jangled by travelling ever since his being in a railway accident in 1865. Other Americans noted that, though grizzled, he was "as quick and elastic in his movements as ever." His photographs, wrote a journalist after one of the readings, "give no idea of his genial expression. To us he appears like a hearty, companionable man, with a deal of fun in him." But that very day Dickens was writing, "I am nearly used up," and listing the afflictions now "telling heavily upon me." His pride and the old-trouper tradition made him conceal his sufferings. And, if sometimes by an effort of will, his old high spirits were often on display. "The cheerfullest man of his age," he was called by his American publisher, J.T. Fields; Fields's wife more perceptively noted, "Wonderful, the flow of spirits C.D. has for a sad man."

His fame remained undiminished, though critical opinion was increasingly hostile to him. Henry Wadsworth Longfellow, noting the immense enthusiasm for him during the American tour, remarked: "One can hardly take in the whole truth about it, and feel the universality of his fame." But in many respects he was "a sad man" in these later years. He never was tranquil or relaxed. Various old friends were now estranged or dead or for other reasons less available; he was now leading a less social life and spending more time with young friends of a calibre inferior to his former circle. His sons were causing much worry and disappointment; "all his fame goes for nothing," said a friend, "since he has not the one thing. He is very unhappy in his children." His life was not all dreary, however. He loved his country house, Gad's Hill, and he could still "warm the social atmosphere wherever he appeared with that summer glow which seemed to attend him." T.A. Trollope (contributor to Dickens' *All the Year Round* and brother of the novelist Anthony Trollope), who wrote that, despaired of giving people who had not met him any idea of

> the general charm of his manner. . . . His laugh was brimful of enjoyment. . . . His enthusiasm was boundless. . . . He was a *hearty* man, a large-hearted man, . . . a strikingly manly man.

Only a week before his death he was at the theatre,

> in high spirits, brim-full of *joie-de-vivre*. His talk had all the sparkle of champagne, and he himself kept laughing at the majesty of his own absurdities, as one droll thought followed another, . . . at times still so young and almost boyish in his gaiety. (Lord Redesdale, *Memories*, 1915)

Last speech

Farewell readings. His health remained precarious after the punishing American tour and was further impaired by his addiction to giving the strenuous "Sikes and Nancy" reading. His farewell readings tour was abandoned when, in April 1869, he collapsed. He began writing another novel and gave a short farewell season of readings in London, ending with the famous speech, "From these garish lights I vanish now for evermore . . ." —words repeated, less than three months later, on his funeral card. He died suddenly, on June 9, 1870, and was buried in Westminster Abbey. People all over the world mourned the loss of "a friend" as well as a great entertainer and creative artist and one of the acknowledged influences upon the spirit of the age.

ASSESSMENT

Contemporary opinion. Ralph Waldo Emerson, attending one of the readings in Boston, "laughed as if he must crumble to pieces," but, discussing Dickens afterward, he said:

> I am afraid he has too much talent for his genius; it is a fearful locomotive to which he is bound and can never be free from it nor set to rest. . . . He daunts me! I have not the key.

There is no simple key to so prolific and multifarious an artist nor to the complexities of the man, and interpretation of both is made harder by his possessing and feeling the need to exercise so many talents besides his imagination. How his fiction is related to these talents—practical, journalistic, oratorical, histrionic—remains controversial. Also the geniality and unequalled comedy of the novels must be related to the sufferings, errors, and self-pity of their author and to his concern both for social evils and for the perennial griefs and limitations of humanity. The novels cover a wide range, social, moral, emotional, and psychological. Thus, he is much concerned with very ordinary people but also with abnormality (*e.g.*, eccentricity, depravity, madness, hallucinations, dream states). He is both the most imaginative and fantastic and the most topical and documentary of great novelists. He is unequal, too; a wonderfully inventive and poetic writer, he can also, even in his mature novels, write with a painfully slack conventionality.

Biographers have only recently known enough to explore the complexity of his nature. Critics have always been challenged by his art, though from the start it contained enough easily acceptable ingredients, evident skill and gusto, to ensure popularity. The earlier novels were and have mostly remained the most popular: *Pickwick*, *Oliver Twist*, *Chuzzlewit*, the *Carol*, *Copperfield*. Critics began to demur against the later novels, deploring the loss of the freer comic spirit, baffled by the more symbolic mode of his art, and uneasy when the simpler reformism over isolated issues became a more radical questioning of social assumptions and institutions. Dickens was never neglected or forgotten and never lost his popularity, but for 70 years after his death he received remarkably little serious attention (George Gissing, G.K. Chesterton, and George Bernard Shaw being notable exceptions). F.R. Leavis, later to revise his opinion, was speaking for many, in 1948, when he asserted that "the adult mind doesn't as a rule find in Dickens a challenge to an unusual and sustained seriousness"; Dickens was indeed a great genius, "but the genius was that of a great entertainer."

Most popular novels

Modern criticism. Modern Dickens criticism dates from 1940–41, with the very different impulses given by George Orwell, Edmund Wilson, and Humphry House. In the 1950s, a substantial reassessment and re-editing of the works began, his finest artistry and greatest depth now being discovered in the later novels—*Bleak House*, *Little Dorrit*, and *Great Expectations*—and (less unanimously) in *Hard Times* and *Our Mutual Friend*. Scholars have explored his working methods, his relations with his public, and the ways in which he was simultaneously an eminently Victorian figure and an author "not of an age but for all time." Biographically, little had been added to Forster's massive and intelligent *Life* (1872–74), except the Ellen Ternan story, until Edgar Johnson's in 1952. Since then, no radically new view has emerged, though particular phases or aspects have received fuller attention. The centenary in 1970 demonstrated a critical consensus about his standing second only to William Shakespeare in English literature, which would have seemed incredible 40 or even 20 years earlier.

Dickens and Shakespeare compared

He was being compared to Shakespeare, for imaginative range and energy, while still in his twenties. He and Shakespeare are the two unique popular classics that England has given to the world, and they are alike in being remembered not for one masterpiece (as Dante, Cervantes, or John Milton are) but for a creative world, a plurality of works populated by a great variety of figures, in situations ranging from the sombre to the farcical. For the common reader, both Shakespeare and Dickens survive through their characterization, though they offer much else. Shakespeare ranges more deeply and widely. Dickens enjoys one temporary advantage in having lived when he did and thus being able to write of an urban industrial world, in which notions of representative government and social responsibility were current—a world containing many of the problems and hopes that persist a century after his death and far beyond the land of his birth.

MAJOR WORKS

NOVELS: *The Pickwick Papers* (1837); *Oliver Twist* (1838); *Nicholas Nickleby* (1839); *The Old Curiosity Shop* and *Barnaby Rudge* (1841), two novels first published in a "clock framework," later abandoned, under the title of *Master Humphrey's Clock; Martin Chuzzlewit* (1844); *Dombey and Son* (1848); *David Copperfield* (1850); *Bleak House* (1853); *Hard Times* (1854); *Little Dorrit* (1857); *A Tale of Two Cities* (1859); *Great Expectations* (1861); *Our Mutual Friend* (1865); *The Mystery of Edwin Drood* (1870, unfinished).

CHRISTMAS BOOKS: *A Christmas Carol* (1843); *The Chimes* (1845, for 1844); *The Cricket on the Hearth* (1846, for 1845); *The Battle of Life* (1846); *The Haunted Man* (1848).

STORIES (CHRISTMAS STORIES): The volume entitled *Christmas Stories* in collected editions includes "A Christmas Tree" (1850); "What Christmas Is as We Grow Older" (1851); "The Poor Relation's Story" (1852); "Nobody's Story" (1853); "The Seven Poor Travellers" (1854); "The Holly-Tree," sometimes called "The Holly-Tree Inn" (1855); "The Wreck of the Golden Mary" (1856); "The Perils of Certain English Prisoners" (1857); "Going into Society" (1858); "The Haunted House" (1859); "A Message from the Sea" (1860); "Tom Tiddler's Ground" (1861); "Somebody's Luggage" (1862); "Mrs. Lirriper's Lodgings" (1863); Mrs. Lirriper's Legacy" (1864); "Doctor Marigold" (1865); "Mugby Junction" (1866); "No Thoroughfare" (1867). (OTHER STORIES): in collected editions generally appended to the volume entitled *Reprinted Pieces*, ["The Lamplighter" (1841);] "To Be Read at Dusk" (1852); "Hunted Down" (1859); "George Silverman's Explanation" (1867); "Holiday Romance" (1868; children's story in 4 parts; pt. 2, "The Magic Fishbone," often reprinted separately).

OTHER WORKS: *Sketches by "Boz,"* 2 series (1836, together, 1839, included Dickens' first published work, "A Dinner at Poplar Walk," 1833); *Sketches of Young Gentlemen* (1838) and *Sketches of Young Couples* (1840), both usually appended to the *Sketches by "Boz"* volume, in collected editions, which also usually contains "The Mudfog Papers" (contributed to *Bentley's Miscellany*, 1837–38); *American Notes* (1842); *Pictures from Italy* (1846); *The Life of Our Lord* (completed 1849, for his children; published 1934); *A Child's History of England* (1852–54); "The Lazy Tour of Two Idle Apprentices" (with Wilkie Collins, contributed to *Household Words* [1857]; often included in the volume entitled *Christmas Stories*); *Reprinted Pieces* (1858; contributed to *Household Words*, 1850–56); *The Uncommercial Traveller* (1861, amplified 1868, 1875; contributed to *All the Year Round*, 1860–69); *Plays and Poems*, ed. by R.H. Shepherd (1885); *Miscellaneous Papers*, ed. by B.W. Matz (1908; the most substantial posthumous collection, mainly essays contributed to *Household Words*, 1850–59; 16 further items, in the volume retitled *Collected Papers*, in

The Nonesuch Dickens, 1937); *Uncollected Writings from Household Words 1850–1859*, ed. by Harry Stone (1968).

BIBLIOGRAPHY

Bibliographies: K.J. FIELDING, *Charles Dickens* (1953); ADA NISBET, "Charles Dickens," in LIONEL STEVENSON (ed.), *Victorian Fiction: A Guide to Research*, pp. 44–153 (1964), full discussion of materials for Dickens studies and of writings about him in many languages; PHILIP COLLINS, *A Dickens Bibliography* (1970), off printed from GEORGE WATSON (ed.), *New Cambridge Bibliography of English Literature*, vol. 3, col. 779–850.

Most of the manuscripts and proof sheets of the novels are in the Victoria and Albert Museum, London. Other important collections of manuscripts and letters are in Dickens House, London; the British Museum; New York Public Library; Pierpont Morgan Library, New York; Free Library of Philadelphia; Henry E. Huntington Library and Art Gallery, San Marino, California; the University of Texas Libraries; and Yale University Library. The Dickens Fellowship (Dickens House, London) has branches all over the world and publishes the *Dickensian* (3/yr.). *Dickens Studies Newsletter* (quarterly) and *Dickens Studies Annual* are published from Carbondale, Illinois, where the Dickens Society is based.

Collected editions: *The New Oxford Illustrated Dickens* (1947–58) is the most cited but will be replaced by the *Clarendon* edition (1966–). See also *Speeches*, ed. by K.J. FIELDING (1960); and *Public Readings*, ed. by PHILIP COLLINS (forthcoming).

Letters: The most complete collection, *The Letters of Charles Dickens*, ed. by W. DEXTER, 3 vol. (1938), is being replaced by *The Letters of Charles Dickens*, ed. by M. HOUSE *et al.* (1965–). See also *The Heart of Charles Dickens, As Revealed in His Letters to Angela Burdett-Coutts*, ed. by E. JOHNSON (1952).

Biographies: JOHN FORSTER, *The Life of Charles Dickens*, 3 vol. (1872–74), remains indispensible, though EDGAR JOHNSON, *Charles Dickens: His Tragedy and Triumph*, 2 vol. (1952, reprinted 1965), is now the standard biography.

Criticism: GEORGE R. GISSING, *Charles Dickens: A Critical Study* (1898); G.K. CHESTERTON, *Charles Dickens* (1906); GEORGE ORWELL, "Dickens," in *Critical Essays*, pp. 7–56 (1946); EDMUND WILSON, "Dickens: The Two Scrooges," in *The Wound and the Bow*, pp. 1–104 (1941); HUMPHRY HOUSE, *The Dickens World*, 2nd ed. (1960), the best discussion of Dickens and his age; G.H. FORD, *Dickens and His Readers* (1955); JOHN E. BUTT and KATHLEEN TILLOTSON, *Dickens at Work* (1957); J. HILLIS MILLER, *Charles Dickens: The World of His Novels* (1958), the most influential critical study since Edmund Wilson's; PHILIP COLLINS, *Dickens and Crime* (1962); ROBERT GARIS, *The Dickens Theatre* (1965); ANGUS WILSON, *The World of Charles Dickens* (1970); F.R. and Q.D. LEAVIS, *Dickens, the Novelist* (1971).

Anthologies of Dickens criticism: G.H. FORD and L. LANE (eds.), *The Dickens Critics* (1961); STEPHEN WALL (ed.), *Charles Dickens: A Critical Anthology* (1970); and PHILIP COLLINS (ed.), *Dickens, the Critical Heritage* (1971), on his reception 1836–82.

(Ph.C.)

Dickinson, Emily

Emily Dickinson is one of the world's masters of the short lyric poem. The subjects of her poems, expressed in intimate, domestic figures of speech, include love, death, and nature. The contrast between her quiet, secluded life in the house in which she was born and died, and the depth and intensity of her terse poems, has provoked much speculation about her personality and personal relationships. Her 1,775 poems and her letters, which survive in almost as great a number, reveal a passionate, witty woman and a scrupulous craftsman who made an art not only of her poetry but also of her correspondence and her life.

Early life

Emily Elizabeth Dickinson was born December 10, 1830, at Amherst, Massachusetts, the second of three children. The three remained close throughout their adult lives: her younger sister, Lavinia, stayed in the family home and did not marry, and her older brother, Austin, lived in the house next door after his marriage to a friend of Emily's. Her grandfather, Samuel Fowler Dickinson, had been one of the founders of Amherst College, and her father, Edward Dickinson, served as treasurer of the college from 1835 to 1872. A lawyer who served one term (1853–1855) in Congress, Edward Dickinson was an austere and somewhat remote father, but not an unkind one. Emily's mother, too, was not close to her children.



Emily Dickinson, daguerreotype, *c.* 1847.
By courtesy of The Harvard College Library

Emily was educated at Amherst Academy and Mount Holyoke Female Seminary. Mount Holyoke, which she attended from 1847 to 1848, insisted on religious as well as intellectual growth, and Emily was under considerable pressure to become a professing Christian. She resisted, however, and although many of her poems deal with God, she remained all her life a skeptic. Despite her doubts, she continued to hold strong religious feelings, a conflict that lent tension to her writings.

Emily began to write verse about 1850, apparently while under the spell of the poems of Ralph Waldo Emerson and Emily Brontë, and under the tutelage of Benjamin F. Newton, a young man studying law in her father's office. Only a handful of her poems can be dated before 1858, when she began to collect them into small, handsewn booklets. Her letters of the 1850s reveal a vivacious, humorous, somewhat shy young woman. In 1855 Emily went to Washington, D.C., with her sister to visit their father, who was serving in Congress. During the trip they stopped off at Philadelphia, where she heard the preaching of the noted clergyman, the Rev. Charles Wadsworth, who was to become her "dearest earthly friend." He was something of a romantic figure: a man said to have known great sorrow, whose eloquence in the pulpit contrasted with his solitary broodings. He and Emily exchanged letters on spiritual matters, his Calvinist orthodoxy perhaps serving as a useful foil for her own speculative reasoning. She may also have found in his stern, rigorous beliefs a welcome corrective to the easy assumption of a benign universe made by Emerson and the other Transcendentalists.

In the 1850s Emily began two of her significant correspondences—with Dr. and Mrs. Josiah G. Holland and with Samuel Bowles. The two men were editors of the *Springfield* (Massachusetts) *Republican*, a paper that took an interest in literary matters and even published verse. The correspondence continued over the years, although in the case of the Hollands, most of the letters after the 1850s went to Mrs. Holland, a woman intelligent enough to comprehend Emily's subtleties and witticisms. Emily tried to interest Bowles in her poetry, and it was a crushing blow to her that he, a man of quick mind but conventional literary tastes, failed to appreciate it.

By the late 1850s, when she was writing poems at a steadily increasing pace, Emily Dickinson loved a man whom she called "Master" in three drafts of letters. "Master" does not exactly resemble any of Emily's known friends, but may have been Bowles or Wadsworth. This love shines forth in several lines from her poems: "I'm ceded—I've stopped being Theirs," " 'Tis so much joy! 'Tis so much joy," and "Dare you see a Soul *at the White*

Heat?" to name only a few. Other poems reveal the frustration of this love and its gradual sublimation into a love for Christ and a celestial marriage to him.

Character of her poetry

The poems of the 1850s are fairly conventional in sentiment and form, but beginning about 1860, they become experimental both in language and prosody, though they owe much to the metres of the English hymn writer Isaac Watts and to Shakespeare and the King James Version of the Bible. Emily's prevailing poetic form was the quatrain of three iambic feet, a type described in one of the books by Watts in the family library. She used many other forms as well, and to even the simpler hymnbook measures she gave complexity by constantly altering the metrical beat to fit her thought: now slow, now fast, now hesitant. She broke new ground in her wide use of off-rhymes, varying from the true in a variety of ways that also helped to convey her thought and its tensions. In striving for an epigrammatic conciseness, she stripped her language of superfluous words and saw to it that those that remained were vivid and exact. She tampered freely with syntax and liked to place a familiar word in an extraordinary context, shocking the reader to attention and discovery.

On April 15, 1862, Emily Dickinson wrote a letter, enclosing four poems, to a literary man, Thomas Wentworth Higginson, asking whether her poems were "alive." Higginson, although he advised Emily not to publish, recognized the originality of her poems and remained her "preceptor" for the rest of her life. After 1862 Emily Dickinson resisted all efforts by her friends to put her poems before the public. As a result, only seven poems by Emily Dickinson were published during her lifetime, five of them in the *Springfield Republican*.

Period of greatest creativity

The years of Emily Dickinson's greatest poetic output, about 800 poems, coincide with the Civil War. Although she looked inward and not to the war for the substance of her poetry, the tense atmosphere of the war years may have contributed to the urgency of her writing. The year of greatest stress was 1862, when distance and danger threatened Emily's friends—Samuel Bowles, in Europe for his health; Charles Wadsworth, who had moved to a new pastorate at the Calvary Church in San Francisco; and T.W. Higginson, serving as an officer in the Union Army. Emily also had persistent eye trouble, which led her, in 1864 and 1865, to spend several months in Cambridge, Massachusetts, for treatment. Once back in Amherst, she never travelled again, and after the late 1860s never left the boundaries of the family's property.

After the Civil War, Emily Dickinson's poetic tide ebbed, but she sought increasingly to regulate her life by the rules of art. Her letters, some of them equal in artistry to her poems, classicize daily experience in an epigrammatic style. For example, when a friend affronted Emily by sending a letter jointly to her and her sister, she replied: "A Mutual plum is not a plum. I was too respectful to take the pulp and do not like a stone." By 1870 Emily Dickinson dressed only in white and saw few of the callers who came to the homestead; her seclusion was fiercely guarded by her devoted sister. In August 1870, Higginson visited Amherst and described Emily as "a little plain woman" with reddish hair, dressed in white, bringing him flowers as her "introduction" and speaking in a "soft frightened breathless childlike voice."

Her later years were marked with sorrow at the deaths of many people she loved. The most prostrating of these were the deaths of her father in 1874 and her eight-year-old nephew Gilbert in 1883, which occasioned some of her finest letters. She also mourned the loss of Bowles in 1878, Holland in 1881, Charles Wadsworth and her mother in 1882, Otis P. Lord in 1884, and Helen Hunt Jackson in 1885. Lord, a judge from Salem, Massachusetts, with whom Emily fell in love about 1878, had been the closest friend of Emily's father. Emily's drafts of letters to Lord reveal a tender, mature love, which Lord returned. Mrs. Jackson, a poet and popular novelist, discerned the greatness of Emily's poetry and tried unsuccessfully to get her to publish it.

Emily Dickinson died on May 15, 1886. Soon after her death her sister Lavinia determined to have Emily's poems published. In 1890 *Poems by Emily Dickinson*, edited by T.W. Higginson and Mabel Loomis Todd, appeared. Other volumes of Dickinson poems, edited chiefly by Mabel Loomis Todd, Martha Dickinson Bianchi (Emily's niece), and Millicent Todd Bingham, were published 1891–1945, and in 1955 Thomas H. Johnson edited all the surviving poems and their variant versions.

MAJOR WORKS

No collection of poems by Emily Dickinson was published in her lifetime. The first selection, *Poems by Emily Dickinson* (1890), was followed by *Poems: Second Series* (1891), and *Poems: Third Series* (1896). Additional poems were included in *Letters of Emily Dickinson*, 2 vol. (1894). Later volumes of poems were: *The Single Hound: Poems of a Lifetime* (1914), *Further Poems of Emily Dickinson: Withheld from Publication by Her Sister Lavinia* (1929), *Unpublished Poems of Emily Dickinson* (1935), and *Bolts of Melody: New Poems of Emily Dickinson* (1945).

BIBLIOGRAPHY. S.T. CLENDENNING, *Emily Dickinson: A Bibliography, 1850–1966* (1968), is the most recent and most comprehensive bibliography. The great majority of Dickinson manuscripts, both poems and letters, are in the libraries of Harvard University and Amherst College. Emily Dickinson's home, the property of Amherst College, contains some memorabilia. The basic text of the poems is *The Poems of Emily Dickinson, Including Variant Readings Critically Compared with All Known Manuscripts*, 3 vol., ed. by T.H. JOHNSON (1955); the most complete edition of the letters, *The Letters of Emily Dickinson*, 3 vol., ed. by T.H. JOHNSON and THEODORA WARD (1958). There is as yet no definitive biography of Emily Dickinson. Biographical studies include: M.T. BINGHAM, *Emily Dickinson's Home: Letters of Edward Dickinson and His Family* (1955), the best account to date of Emily Dickinson's early years; T.H. JOHNSON, *Emily Dickinson: An Interpretive Biography* (1955), an extended critical biography; and JAY LEYDA, *The Years and Hours of Emily Dickinson*, 2 vol. (1960), a day-by-day guidebook to the life of Emily Dickinson. Two critical studies of the poems are C.R. ANDERSON, *Emily Dickinson's Poetry: Stairway of Surprise* (1960); and G.F. WHICHER, *This Was a Poet* (1938). Although Whicher's book is no longer wholly reliable as biography, both of these works are critically excellent.

(D.J.M.H.)

Dictionary

The word "dictionary" is used to describe a wide variety of reference works. Basically, a dictionary lists a set of words with information about them. The list may attempt to be a complete inventory of a language or may be only a small segment of it. A short list, sometimes at the back of a book, is often called a glossary. When a word list is an index to a limited body of writing, with references to each passage, it is called a concordance. Theoretically, a good dictionary could be compiled by bringing together and organizing into one list a large number of concordances. When a word list is of geographic names only, it takes the special name gazetteer.

The word "lexicon" designates a wordbook, but it also has a special abstract meaning among linguists, referring to the body of separable structural units of which the language is made up. In this sense a preliterate culture has a lexicon long before its units are written down in a dictionary. Scholars in England sometimes use "lexis" to designate this lexical element of language.

The compilation of a dictionary is lexicography; lexicology is a branch of linguistics in which, with the utmost scientific rigour, the theories that lexicographers make use of in the solution of their problems are developed.

The common phrase "dictionary order" takes for granted that the alphabetical order will be followed, and yet the alphabetical order has been called a tyranny that makes dictionaries less useful than they might be if compiled in some other order. The assembling of words into groups related by some principle, as by their meanings, can be done, and such a work is often called a thesaurus or synonymy. But such works need an index for ease of reference, and it is unlikely that alphabetical order will be superseded except in specialized works.

The distinction between a dictionary and an encyclopaedia is easy to state but difficult to carry out in a practical way: a dictionary explains words, whereas an encyclopaedia explains things. Because words achieve

their usefulness by referring to things, however, it is difficult to construct a dictionary without considerable attention to the objects and abstractions designated.

Distinction between dictionary and encyclopaedia

A monolingual dictionary has both the word list and the explanations in the same language, whereas bilingual or multilingual (polyglot) dictionaries have the explanations in another language or different languages. The word "dictionary" is also extended, in a loose sense, to reference books with entries in alphabetical order, such as a dictionary of biography, a dictionary of heraldry, or a dictionary of plastics.

The present article, after an account of the development of dictionaries from classical times to the recent past, treats the kinds of dictionaries and their features and problems. It concludes with a brief section on some of the major dictionaries that are available. Examples and illustrations for the sections on the types of dictionaries and on their features and problems are drawn primarily from the products of English lexicographers.

HISTORICAL BACKGROUND

From classical times to 1604. In the long perspective of human evolutionary development, dictionaries have been known through only a slight fraction of language history. People at first simply talked without having any authoritative backing from reference books. A short Akkadian word list, from central Mesopotamia, has survived from the 7th century BC. The Western tradition of dictionary making began among the Greeks, although not until the language had changed so much that explanations and commentaries were needed. After a 1st-century-AD lexicon by Pamphilus of Alexandria, many lexicons were compiled in Greek, the most important being those of the Atticists in the 2nd century, that of Hesychius of Alexandria in the 5th century, and those of Photius and Suidas in the Middle Ages. (The Atticists were compilers of lists of words and phrases thought to be in accord with the usage of the Athenians.)

Because Latin was a much-used language of great prestige well into modern times, its monumental dictionaries were important and later influenced English lexicography. In the 1st century BC, Marcus Terentius Varro wrote a treatise *De lingua Latina*; the extant books of its section of etymology are valuable for their citations from Latin poets. At least five medieval scholastics—Papias the Lombard, Alexander Neckam, Johannes de Garlandia (John Garland), Hugo of Pisa, and Giovanni Balbi of Genoa—turned their attention to dictionaries. The mammoth work of Ambrogio Calepino, published at Reggio (now Reggio nell'Emilia), in 1502, incorporating several other languages besides Latin, was so popular that "calepin" came to be an ordinary word for a dictionary. A Lancashire will of 1568 contained the provision: "I wyll that Henry Marrecrofte shall have my calapyne and my parafrasies." This is an early instance of the tendency that, several centuries later, caused people to say, "Look in Johnson" or "Look in Webster."

Interlingual dictionaries

Because language problems within a single language do not loom so large to ordinary people as those that arise in the learning of a different language, the interlingual dictionaries developed early and had great importance. The corporation records of Boston, Lincolnshire, have the following entry for the year 1578:

> That a dictionarye shall be bought for the scollers of the Free Scoole, and the same boke to be tyed in a cheyne, and set upon a deske in the scoole, whereunto any scoller may have accesse, as occasion shall serve.

The origin of the bilingual lists can be traced to a practice of the early Middle Ages, that of writing interlinear glosses—explanations of difficult words—in manuscripts. It is but a step for these glosses to be collected together at the back of a manuscript and then for the various lists—glossaries—to be assembled in another manuscript. Some of these have survived from the 7th and 8th centuries—and in some cases they preserve the earliest recorded forms in English.

The first bilingual glossary to find its way into print was a French–English vocabulary for the use of travellers, printed in England by William Caxton without a title page, in 1480. The words and expressions appeared in parallel columns on 26 leaves. Next came a Latin–English vocabulary by a noted grammarian, John Stanbridge, published by Richard Pynson in 1496 and reprinted frequently. But far more substantial in character was an English–Latin vocabulary called the *Promptorius puerorum* ("Storehouse [of words] for Children") brought out by Pynson in 1499. It is better known under its later title of *Promptorium parvulorum sive clericorum* ("Storehouse for Children or Clerics") commonly attributed to Geoffrey the Grammarian (Galfridus Grammaticus), a Dominican friar of Norfolk, who is thought to have composed it about 1440.

The next important dictionary to be published was an English–French one by John (or Jehan) Palsgrave in 1530, *Lesclaircissement de la langue francoise* ("Elucidation of the French Tongue"). Palsgrave was a tutor of French in London, and a letter has survived showing that he arranged with his printer that no copy should be sold without his permission,

> lest his proffit by teaching the Frenche tonge myght be mynished by the sale of the same to suche persons as, besids hym, wern disposed to studye the sayd tongue.

A Welsh–English dictionary by William Salesbury in 1547 brought another language into requisition: *A Dictionary in Englyshe and Welshe moche necessary to all suche Welshemen as wil spedly learne the Englyshe tôgue.* The encouragement of Henry VIII was responsible for an important Latin–English dictionary that appeared in 1538 from the hand of Sir Thomas Elyot. Thomas Cooper enlarged it in subsequent editions and in 1565 brought out a new work based upon it—*Thesaurus Linguae Romanae et Britannicae* ("Thesaurus of the Roman Tongue and the British"). A hundred years later John Aubrey, in *Brief Lives*, recorded Cooper's misfortune while compiling it:

Thomas Cooper and his *Thesaurus*

> His wife . . . was irreconcileably angrie with him for sitting-up late at night so, compileing his Dictionary. . . . When he had halfe-donne it, she had the opportunity to gett into his studie, tooke all his paines out in her lap, and threw it into the fire, and burnt it. Well, for all that, that good man had so great a zeale for the advancement of learning, that he began it again, and went through with it to that perfection that he hath left it to us, a most usefull worke.

More important still was Richard Huloet's work of 1552, *Abecedarium Anglo-Latinum*, for it contained a greater number of English words than had before appeared in any similar dictionary. In 1556 appeared the first edition by John Withals of *A shorte Dictionarie for Yonge Beginners*, which gained greater circulation (to judge by the frequency of editions) than any other book of its kind. Many other lexicographers contributed to the development of dictionaries. Certain dictionaries were more ambitious and included a number of languages, such as John Baret's work of 1573, *An Alvearie: or triple Dictionarie, in Englishe, Latin, and French.* In his preface Baret acknowledged that the work was brought together by his students in the course of their exercises, and the title *Alvearie* was to commemorate their "beehive" of industry. The first rhyming dictionary, by Peter Levens, was produced in 1570—*Manipulus Vocabulorum. A Dictionarie of English and Latine wordes, set forthe in suche order, as none heretofore hath ben.*

The first rhyming dictionary

The interlingual dictionaries had a far greater stock of English words than were to be found in the earliest all-English dictionaries, and the compilers of the English dictionaries, strangely enough, never took full advantage of these sources. It may be surmised, however, that people in general sometimes consulted the interlingual dictionaries for the English vocabulary. The anonymous author of *The Arte of English Poesie*, thought to be George Puttenham, wrote, in 1589, concerning the adoption of southern speech as the standard:

> herein we are already ruled by th' English Dictionaries and other bookes written by learned men, and therefore is needeth none other direction in that behalfe.

The mainstream of English lexicography is the word list explained in English. The first known English–English glossary grew out of the desire of the supporters of

the Reformation that even the most humble Englishman should be able to understand the Scriptures. William Tyndale, when he printed the Pentateuch on the Continent in 1530, included "A table expoundinge certeyne wordes." The following entries are typical:

> Albe, a longe garment of white lynen.
> Boothe, an housse made of bowes.
> Brestlappe or brestflappe, is soche a flappe as thou seist in the brest or a cope.
> Consecrate, to apoynte a thinge to holy uses.
> Dedicate, purifie or sanctifie.
> Firmament: the skyes.
> Slyme was . . . a fattenesse that osed out of the erth lyke unto tarre/And thou mayst call it cement/if thou wilt.
> Tabernacle, an house made tentwise, or as a pauelion.
> Vapor/a dewymiste/as the smoke of a sethynge pott.

Spelling reformers long had a deep interest in producing English dictionaries. In 1569 one such reformer, John Hart, lamented that the "disorders and confusions" of spelling were so great that "there can be made no perfite Dictionarie nor Grammer." But a few years later the phonetician William Bullokar promised to produce such a work and stated, "A dictionary and grammar may stay our speech in a perfect use for euer."

The schoolmasters also had a strong interest in the development of dictionaries. In 1582 Richard Mulcaster, of the Merchant Taylors' school and later of St. Paul's, expressed the wish that some learned and laborious man "wold gather all the words which we vse in our English tung," and in his book commonly referred to as *The Elementarie* he listed about 8,000 words, without definitions, in a section called "The Generall Table." Another schoolmaster, Edmund Coote, of Bury St. Edmund's, in 1596 brought out *The Englishe Scholemaister, teachinge all his schollars of what age soever the most easie short & perfect order of distinct readinge & true writinge our Englishe tonge*, with a table that consisted of about 1,400 words, sorted out by different typefaces on the basis of etymology. This is important, because what is known as the "first" English dictionary, eight years later, was merely an adaptation and enlargement of Coote's table.

First purely English dictionary

From 1604 to 1828. In 1604 at London appeared the first purely English dictionary to be issued as a separate work, entitled *A Table Alphabeticall, conteyning and teaching the true writing and understanding of hard usuall English wordes, borrowed from the Hebrew, Greeke, Latine,* or *French &c.*, by Robert Cawdrey, who had been a schoolmaster at Oakham, Rutland, about 1580, and in 1604 was living at Coventry. He had the collaboration of his son Thomas, a schoolmaster in London. This work contained about 3,000 words but was so dependent upon three sources that it can rightly be called a plagiarism. The basic outline was taken over from Coote's work of 1596, with 87 percent of his word list adopted. Further material was taken from the Latin–English dictionary by Thomas Thomas, *Dictionarium linguae Latinae et Anglicanae* (1588). But the third source is most remarkable. In 1599 a Dutchman known only as A.M. translated from Latin into English a famous medical work by Oswald Gabelkhouer, *The Boock of Physicke*, published at Dort, in the Netherlands. As he had been away from England for many years and had forgotten much of his English, A.M. sometimes merely put English endings on Latin words. When friends told him that Englishmen would not understand them, he compiled a list of them, explained by a simpler synonym, and put it at the end of the book. Samples are: "Puluerisated, reade beaten; Frigifye, reade coole; Madefye, reade dipp; Calefye, reade heat; Circumligate, reade binde; Ebulliated, read boyled." Thus, the fumblings of a Dutchman who knew little English (in fact, his errata) were poured into Cawdrey's word list. But other editions of Cawdrey were called for—a second in 1609, a third in 1613, and a fourth in 1617.

The next dictionary, by John Bullokar, *An English Expositor*, is first heard of on May 25, 1610, when it was entered in the Stationers' Register (which established the printer's right to it), but it was not printed until six years later. Bullokar introduced many archaisms, marked with a star ("onely used of some ancient writers, and now growne out of use"), such as "aye," "eld," "enewed," "fremd," "gab," and "glee." The work had 14 editions, the last as late as 1731.

Still in the tradition of hard words was the next work, in 1623, by Henry Cockeram, the first to have the word dictionary in its title: *The English Dictionarie: or, an Interpreter of hard English Words.* It added many words that have never appeared anywhere else—adpugne, adstupiate, bulbitate, catillate, fraxate, nixious, prodigity, vitulate, and so on. Much fuller than its predecessors was Thomas Blount's work of 1656, *Glossographia: or, a dictionary Interpreting all such hard words . . . as are now used in our refined English tongue.* He made an important forward step in lexicographical method by collecting words from his own reading that had given him trouble; and he often cited the source. Much of Blount's material was appropriated two years later by Edward Phillips, a nephew of the poet Milton, for a work called *The New World of English Words*, and Blount castigated him bitterly.

Thus far, the English lexicographers had all been men who made dictionaries in their leisure time or as an avocation, but in 1702 appeared a work by the first professional lexicographer, John Kersey the Younger. This work, *A New English Dictionary*, incorporated much from the tradition of spelling books and discarded most of the fantastic words that had beguiled earlier lexicographers. As a result, it served the reasonable needs of ordinary users of the language. Kersey later produced some bigger works, but all of these were superseded in the 1720s, when Nathan Bailey, a schoolmaster in Stepney, issued several innovative works. In 1721 he produced *An universal etymological English Dictionary*, which for the rest of the century was more popular even than Dr. Johnson's. A supplement in 1727 was the first dictionary to mark accents for pronunciation. Bailey's imposing *Dictionarium Britannicum* of 1730 was used by Samuel Johnson as a repository during the compilation of the monumental dictionary of 1755.

Kersey's *New English Dictionary*

Many literary men felt the inadequacy of English dictionaries, particularly in view of the continental examples. The Accademia della Crusca, of Florence, founded in 1582, brought out its *Vocabolario* at Venice in 1612, filled with copious quotations from Italian literature. The Académie Française produced its dictionary in 1694, but two other French dictionaries were actually more scholarly—that of César-Pierre Richelet in 1680 and that of Antoine Furetière in 1690. In Spain the Royal Spanish Academy (Real Academia Española), founded in 1713, produced its *Diccionario de la lengua Castellana*, 1726–1739, in six thick volumes. The foundation work of German lexicography, by Johann Leonhard Frisch, *Teutsch-Lateinisches Wörterbuch*, in 1741, freely incorporated quotations in German. The Russian Academy of Arts (St. Petersburg) published the first edition of its dictionary somewhat later, from 1789 to 1794. Both the French and the Russian academies arranged the first editions of their dictionaries in etymological order but changed to alphabetical order in the second editions.

In England, in 1707, the antiquary Humphrey Wanley set down in a list of "good books wanted," which he hoped the Society of Antiquaries would undertake: "A dictionary for fixing the English language, as the French and Italian." A number of noted authors made plans to fulfill this aim (Joseph Addison, Ambrose Philips, Alexander Pope, and others), but it remained for a promising poet and critic, Samuel Johnson, to bring such a project to fulfillment. Five leading booksellers of London banded together to support his undertaking, and a contract was signed on June 18, 1746. Next year Johnson's *Plan* was printed, a prospectus of 34 pages, consisting of a discussion of language that can still be read as a masterpiece in its judicious consideration of linguistic problems.

Samuel Johnson's *Plan*

With the aid of six amanuenses to copy quotations, Johnson read widely in the literature up to his time and gathered the central word-stock of the English language. He included about 43,500 words (a few more than the number in Bailey), but they were much better selected and represented the keen judgment of a man of letters.

He was sympathetic to the desire of that age to "fix" the language, but he realized as he went ahead that "language is the work of man, of a being from whom permanence and stability cannot be derived." At most, he felt that he could curb "the lust for innovation."

The chief glory of Johnson's dictionary was its 118,000 illustrative quotations. No doubt some of these were included for their beauty, but mostly they served as the basis for his sense discriminations. No previous lexicographer had the temerity to divide the verb "take," transitive, into 113 senses and the intransitive into 21 more. The definitions often have a quaint ring to modern readers because the science of the age was either not well developed or was not available to him. But mostly the definitions show a sturdy common sense, except when Johnson used long words sportively. His etymologies reflect the state of philology in his age. Usually they were an improvement on those of his predecessors, because he had as a guide the *Etymologicum Anglicanum* of Franciscus Junius, as edited by Edward Lye, which became available in 1743 and which provided guidance for the important Germanic element of the language.

Four editions were issued during Dr. Johnson's lifetime, the fourth in 1773 having received much personal care in revision. The *Dictionary* retained its supremacy for many decades and received lavish praise, but some would-be rivals were bitter in criticism. A widely heralded work of the 1780s and 1790s was the projected dictionary of Herbert Croft, in a manuscript of 200 quarto volumes, to be called the *Oxford English Dictionary*. Croft was however, unable to get it into print.

By courtesy of the Newberry Library, Chicago

OA'TMEAL. *n. ſ.* [*oat* and *meal.*] Flower made by grinding oats.

Oatmeal and butter, outwardly applied, dry the ſcab on the head. *Arbuthnot on Aliment.*

Our neighbours tell me oft, in joking talk,
Of aſhes, leather, *oatmeal*, bran, and chalk. *Gay.*

OA'TMEAL. *n. ſ.* An herb. *Ainſworth.*

OATS. *n. ſ.* [aten, Saxon.] A grain, which in England is generally given to horſes, but in Scotland ſupports the people.

It is of the graſs leaved tribe; the flowers have no petals, and are diſpoſed in a looſe panicle: the grain is eatable. The meal makes tolerable good bread. *Miller.*

The *oats* have eaten the horſes. *Shakeſpeare.*

It is bare mechaniſm, no otherwiſe produced than the turning of a wild *oatbeard*, by the inſinuation of the particles of moiſture. *Locke.*

For your lean cattle, fodder them with barley ſtraw firſt, and the *oat* ſtraw laſt. *Mortimer's Huſbandry.*

His horſe's allowance of *oats* and beans, was greater than the journey required. *Swift.*

OA'TTHISTLE. *n. ſ.* [*oat* and *thiſtle.*] An herb. *Ainſ.*

The definition of "Oats" (top) by Samuel Johnson in his *Dictionary* of 1755 exemplifies his prejudice against the Scots and shows his divergence from his source, Nathan Bailey (bottom), who interspersed idiomatic examples throughout his entries (1736).

OARS, [ora̅n. *Sax.* aura *Su.*] a boat for carrying paſſengers, with two men to row it; alſo inſtruments wherewith boats are rowed.

To have an OAR in every Man's Boat.

That is, to meddle with every man's concerns.

OATS [of aten or etan. *Sax.* to eat] a grain, food for horſes.

To ſow one's wild OATS.

That is to play one's youthful pranks.

OAT *Thiſtle*, an herb.

OA'TEN, of or pertaining to oats

OATH [aꝺ, *Sax.* Eeꝺ, *Dan* and *Su.* Eeꝺt, *Da.* Eꝺ, G.] a ſwearing, or confirming a thing by ſwearing.

OATH [in a legal ſenſe] a ſolemn action, whereby God is called to witneſs the truth of an affirmation, given before one or more perſons impowered to receive the ſame.

OAT-MEAL [of aten and mealepe, *Sax.*] meal or flour made of oats.

Pronouncing dictionaries

The practice of marking word stress was taken over from the spelling books by Bailey in his *Dictionary* of 1727, but a full-fledged pronouncing dictionary was not produced until 1757, by James Buchanan; his was followed by those of William Kenrick (1773), William Perry (1775), Thomas Sheridan (1780), and John Walker (1791), whose decisions were regarded as authoritative, especially in the United States.

The attention to dictionaries was thoroughly established in U.S. schools in the 18th century. Benjamin Franklin, in 1751, in his pamphlet "Idea of the English School," said, "Each boy should have an English dictionary to help him over difficulties." The master of an English grammar school in New York in 1771, Hugh Hughes, announced: "Every one of this Class will have Johnson's Dictionary in Octavo." These were imported from England, because the earliest dictionary printed in the U.S. was in 1788, when Isaiah Thomas of Worcester, Massachusetts, issued an edition of Perry's *Royal Standard English Dictionary*. The first dictionary compiled in America was *A School Dictionary* by Samuel Johnson, Jr. (not a pen name), printed in New Haven, Connecticut, in 1798. Another, by Caleb Alexander, was called *The Columbian Dictionary of the English Language* (1800) and on the title page claimed that "many new words, peculiar to the United States," were inserted. It received abuse from critics who were not yet ready for the inclusion of American words.

In spite of such attitudes, Noah Webster, already well-known for his spelling books and political essays, embarked on a program of compiling three dictionaries of different sizes that included Americanisms. In his announcement on June 4, 1800, he entitled the largest one *A Dictionary of the American Language*. He brought out his small dictionary for schools, the *Compendious*, in 1806 but then engaged in a long course of research into the relation of languages, in order to strengthen his etymologies. At last, in 1828, at the age of 70, he published his master work, in two thick volumes, with the title *An American Dictionary of the English Language*. His change of title reflects his growing conservatism and his recognition of the fundamental unity of the English language. His selection of the word list and his well-phrased definitions made his work superior to previous works, although he did not give illustrative quotations but merely cited the names of authors. The dictionary's worth was recognized, although Webster himself was always at the centre of a whirlpool of controversy.

New trends in dictionary making

Since 1828. It was Noah Webster's misfortune to be superseded in his philology in the very decade that his masterpiece came out. He had spent many years in compiling a laborious "Synopsis" of 20 languages, but he lacked an awareness of the systematic relationships in the Indo-European family of languages. Germanic scholars such as Jacob Grimm, Franz Bopp, and Rasmus Rask had developed a rigorous science of "comparative philology," and a new era of dictionary making was called for. Even as early as 1812 Franz Passow had published an essay in which he set forth the canons of a new lexicography, stressing the importance of the use of quotations arranged chronologically in order to exhibit the history of each word. The brothers Jacob and Wilhelm Grimm developed these theories in their preparations for the *Deutsches Wörterbuch* in 1838. The first part of it was printed in 1852, but the end was not reached until more than a century later, in 1960. French scholarship was worthily represented by Maximilien-Paul-Émile Littré, who began working on his *Dictionnaire de la langue française* in 1844, but, with interruptions of the Revolution of 1848 and his philosophical studies, he did not complete it until 1873.

Among scholars in England the historical outlook took an important step forward in 1808 in the work of John Jamieson on the language of Scotland. Because he did not need to consider the "classical purity" of the language, he included quotations of humble origin; in his *Etymological Dictionary of the Scottish Language*, his use of

"mean" sources marked a turning point in the history of lexicography. Even as late as 1835 the critic Richard Garnett said that "the only good English dictionary we possess is Dr. Jamieson's Scottish one." Another collector, James Jermyn, showed by his publications between 1815 and 1848 that he had the largest body of quotations assembled before that of the *Oxford English Dictionary* (*OED*). Charles Richardson was also an industrious collector, presenting his dictionary, from 1818 on, distributed alphabetically throughout the *Encyclopaedia Metropolitana* (vol. 14 to 25) and then reissued as a separate work in 1835–37. Richardson was a disciple of the benighted John Horne Tooke, whose 18th-century theories long held back the development of philology in England. Richardson excoriated Noah Webster for ignoring "the learned elders of lexicography" such as John Minsheu (whose *Guide into the Tongues* appeared in 1617), Gerhard Johannes Vossius (who published his *Etymologicum linguae Latinae* in 1662), and Franciscus Junius (*Etymologicum Anglicanum*, written before 1677). Richardson did collect a rich body of illustrative quotations, sometimes letting them show the meaning without a definition, but his work was largely a monument of misguided industry that met with the neglect it deserved.

The beginnings of the *OED*

Scholars more and more felt the need for a full historical dictionary that would display the English language in accordance with the most rigorous scientific principles of lexicography. The Philological Society, founded in 1842, established an "unregistered words committee," but, upon hearing two papers by Richard Chenevix Trench in 1857—*On Some Deficiencies in Our English Dictionaries*—the society changed its plan to the making of "A New English Dictionary." Forward steps were taken under two editors, Herbert Coleridge and Frederick James Furnivall, until, in 1879, James Augustus Henry Murray, a Scot known for his brilliance in philology, was engaged as editor. A small army of voluntary readers were inspirited to contribute quotation slips, which reached the number of 5,000,000 in 1898, and no doubt 1,000,000 were added after that. Only 1,827,306 of them were used in print. The copy started going to the printer in 1882; Part I was finished in 1884. Later, three other editors were added, each editing independently with his own staff—Henry Bradley, from the north of England, in 1888, William Alexander Craigie, another Scot, in 1901, and Charles Talbut Onions, the only "Southerner," in 1914. So painstaking was the work that it was not finished until 1928, in over 15,500 pages with three long columns each. A supplement, presenting new material collected since the work began, was issued in 1933. An extraordinarily high standard was maintained throughout. As the name *New English Dictionary* proved to be inadequate, the work came to be called The *Oxford English Dictionary* (*OED*) after 1895.

In the United States, lexicographical activity has been unceasing since 1828. In the middle years of the 19th century, a "war of the dictionaries" was carried on between the supporters of Noah Webster and those of his rival, Joseph Emerson Worcester. To a large extent, this was a competition between publishers who wished to pre-empt the market in the lower schools, but literary people took sides on the basis of other issues. In particular, the contentious Noah Webster had gained a reputation as a reformer of spelling and a champion of American innovations, while the quiet Worcester followed traditions.

In 1846 Worcester brought out an important new work, *A Universal and Critical Dictionary of the English Language*, which included many neologisms of the time, and in the next year Webster's son-in-law, Chauncey Allen Goodrich, edited an improved *American Dictionary* of the deceased Webster. In this edition the Webster interests were taken over by an aggressive publishing firm, the G. & C. Merriam Co. Their agents were very active in the "war of the dictionaries" and sometimes secured an order, by decree of a state legislature, for their book to be placed in every schoolhouse of the state. Worcester's climactic edition of 1860, *A Dictionary of the English Language*, gave him the edge in the "war," and James Russell Lowell declared: "From this long conflict Dr. Worcester has unquestionably come off victorious." The Merriams, however, brought out their answer in 1864, popularly called "the unabridged," with etymologies supplied by a famous German scholar, Karl August Friedrich Mahn. Thereafter, the Worcester series received no major re-editing, and its faltering publishers allowed it to pass into history.

The Century Dictionary

One of the best English dictionaries ever compiled was issued in 24 parts from 1889 to 1891 as *The Century Dictionary*, edited by William Dwight Whitney. It contained much encyclopaedic material but bears comparison even with the *OED*. Isaac Kauffman Funk, in 1893, brought out *A Standard Dictionary of the English Language*, its chief innovation being the giving of definitions in the order of their importance, not the historical order. Thus, at the turn of the new century, the U.S. had four reputable dictionaries—Webster's, Worcester's (already becoming moribund), the *Century*, and Funk's *Standard*. England was also well served by many (the original dates given here)—John Ogilvie (1850), P. Austin Nuttall (1855), Robert Gordon Latham (1866, re-editing Todd's Johnson of 1818), Robert Hunter (1879), and Charles Annandale (1882).

KINDS OF DICTIONARIES

General-purpose dictionaries. Although one may speak of a "general-purpose" dictionary, it must be realized that every dictionary is compiled with a particular set of users in mind. In turn, the public has come to expect certain conventional features (see below *Features and problems*), and a publisher departs from the conventions at his peril. One of the chief demands is that a dictionary should be "authoritative," but the word authoritative is ambiguous. It can refer to the quality of scholarship, the employment of the soundest information available, or it can describe a prescriptive demand for compliance to particular standards. Many people ask for arbitrary decisions in usage choices, but most linguists feel that, when a dictionary goes beyond its function of recording accurate information on the state of the language, it becomes a bad dictionary.

Most people encounter dictionaries in the abridged sizes, commonly called "desk" or "college-size" dictionaries. Such handy abridgments go back to the 18th century; Dr. Johnson issued an octavo size in 1756. Their form had become stultified until, in the 1930s, Edward Lee Thorndike, drawing upon the principles of the psychology of learning, produced a series for schools (*Beginning*, *Junior*, and *Senior*). His dictionaries were not "museums" but tools that encouraged schoolchildren to learn about language. He drew upon his word counts and his "semantic counts" to determine inclusions. The new mode was carried on to the college level by Clarence L. Barnhart in *The American College Dictionary* (*ACD*), in 1947, and in the later college-size works that were revised to meet that competition—the Merriam-Webster *Seventh New Collegiate* (1963), the *Standard College Dictionary* (1963), and *Webster's New World Dictionary* (1953, and second edition 1970). An especially valuable addition was *The Random House Dictionary* (1966), edited by Jess Stein in a middle size called "the unabridged" and by Laurence Urdang in a smaller size (1968).

The Merriam-Webster *New International* of 1909 had a serene, uncluttered air that suited a simpler age. The second edition, completely re-edited, appeared in 1934, and it, in turn, was superseded in 1961 by the *Third New International*, edited by Philip Babcock Gove. Because its competitors of similar size have not been kept up to date, it stands alone among American dictionaries in giving a full report on the lexicon of present-day English. Unfortunately, the advance publicity, before publication, emphasized the quotations from ephemeral writers such as Polly Adler, Ethel Merman, and Mickey Spillane and the statement about "ain't" as "used orally in most parts of the U.S. by many cultivated speakers." Such reports aroused a storm of denunciation in newspapers and magazines by writers who, others asserted, revealed a shocking ignorance of the nature of language. The comments

The *Third New International*

were collected in a "casebook" entitled *Dictionaries and That Dictionary*, edited by James H. Sledd and Wilma R. Ebbitt (1962).

In 1969 came *The American Heritage Dictionary*, edited by William Morris, who was known for his valuable small dictionary *Words* (1947). The *American Heritage* was designed to take advantage of the reaction against the Merriam-Webster *Third*. A "usage panel" of 104 members, chosen mostly from the conservative "literary establishment," provided material for a set of "usage notes." Their pronouncements, found by scholars to be inconsistent, were supposed to provide "the essential dimension of guidance," as the editor put it, "in these permissive times." The etymological material was superior to that in comparable dictionaries.

In England, Henry Cecil Wyld produced his *Universal Dictionary of the English Language* (1932), admirable in every way except for its social class elitism. The smaller sized dictionaries of the Oxford University Press deserve their wide circulation.

Scholarly dictionaries. Beyond the dictionaries intended for practical use are the scholarly dictionaries, with the scientific goal of completeness and rigour in their chosen area. Probably the most scholarly dictionary in the world is the *Thesaurus Linguae Latinae*, being edited in Germany. Its main collections were made from 1883 to 1900, but by 1969 its publication had reached only the letter *O*. A number of countries have "national dictionaries" under way—projects that often take many decades. Two have already been mentioned—the Grimm dictionary for German (a new edition begun in 1965) and the Littré for French (re-edited 1956–58); but, in addition, there are the *Woordenboek der Nederlandsche taal* for Dutch, begun in 1882 and now very close to completion; the *Ordbok öfver svenska språket*, for Swedish, begun in 1882, reaching *S* in 1965; the *Slovar sovremennogo russkogo literaturnogo yazika* ("Dictionary of Modern Literary Russian," begun in 1950); the *Nynorsk ordbok* projected for Norwegian; and *Det Norske litteraere ordboksverk* projected for Danish. Of outstanding scholarship are the *Dictionary of Sanskrit on Historical Principles* being prepared at Pune (Poona), India, and *The Historical Dictionary of the Hebrew Language*, now getting under way in Jerusalem. The most ambitious project of all is located at the Centre National in Nancy, France, directed by Paul Imbs, preparing for a *Trésor de la langue française*. In the decade following 1960, over 250,000,000 word examples were collected, the latest techniques of computerization being used. It remains to be seen how much of this can be printed. A laboratory at Besançon, under the direction of B. Quemada, for contemporary French, has extensive collections.

Thesaurus Linguae Latinae

The *Oxford English Dictionary* remains as the supreme completed achievement in all lexicography. Its size makes its revision impractical and the decision was therefore made for supplementation rather than revision. In 1919 plans were pushed forward for a set of "period dictionaries." After the completion of the *OED* in 1928, the remaining quotations, both used and unused, were divided up for use in each project. The prime mover of this plan, Sir William Craigie, undertook *A Dictionary of the Older Scottish Tongue* himself, covering the period from the 14th to the 17th century in Scottish speech. Enough material was amassed under his direction so that editing could begin in 1925, and before his death in 1957 he arranged that it should be carried on at the University of Edinburgh. By 1971 it had reached the word "natural." The work on the older period spurred the establishment of a project on modern Scots, which got under way in 1925, called *The Scottish National Dictionary*, giving historical quotations after the year 1700. By 1971 the project had passed the word "stane."

Period dictionaries for English

For the mainstream of English, a period dictionary for Old English (before 1100) was planned for many decades by a dictionary committee of the Modern Language Association of America (Old English section), but only in the late 1960s did it get under way at the Centre for Medieval Studies at the University of Toronto. Plans are for the dictionary to be based on a combining of computerized concordances of bodies of Old English literature. A *Middle English Dictionary* has fared much better, covering the period 1100 to 1475. Started in 1925, it had reached the middle *L*'s by 1971, with an overwhelming fullness of detail. For the period 1475 to 1700, an *Early Modern English Dictionary* has not fared as well. It got under way in 1928 at the University of Michigan, and over 3,000,000 quotation slips were amassed, but the work could not be continued in the decade of the Great Depression, and only in the middle 1960s was it revived again. The *OED* supplement of 1933 is again being supplemented—this time in three large volumes, the first of which was published in 1972.

Craigie, in 1925, proposed a dictionary of American English. Support was found for the project, and he transferred from Oxford University to the University of Chicago in order to become its editor. The aim of the work, he wrote, was that of "exhibiting clearly those features by which the English of the American colonies and the United States is distinguished from that of England and the rest of the English-speaking world." Thus, not only specific Americanisms were dealt with but words that were important in the natural history and cultural history of the New World. After a 10-year period of collecting, publication began in 1936 under the title *A Dictionary of American English on Historical Principles*, and the 20 parts (four volumes) were completed in 1944. This was followed in 1951 by a work that limited itself to Americanisms only—*A Dictionary of Americanisms*, edited by Mitford M. Mathews.

The English language, as it has spread widely over the world, has come to consist of a group of coordinate branches, each expressing the needs of its speakers in communication; further scholarly dictionaries are needed to record the particular characteristics of each branch. Both Canada and Jamaica were treated in 1967—*A Dictionary of Canadianisms on Historical Principles*, Walter Spencer Avis, editor in chief, and *Dictionary of Jamaican English*, edited by Frederic G. Cassidy and R.B. LePage A historical dictionary of South African English is under way at Rhodes University, Grahamstown, South Africa, edited by William Branford, and some day full dictionaries must be compiled for Australian English, New Zealand English, and so on. Such dictionaries are valuable in displaying the intimate interrelations of the language to the culture of which it is a part.

Specialized dictionaries. Specialized dictionaries are overwhelming in their variety and their diversity. Each area of lexical study, such as etymology, pronunciation, and usage, can have a dictionary of its own. The earliest important dictionary of etymology for English was Stephen Skinner's *Etymologicon Linguae Anglicanae* of 1671, in Latin, with a strong bias for finding a classical origin for every English word. In the 18th century, a number of dictionaries were published that traced most English words to Celtic sources, because the authors did not realize that the words had been borrowed into Celtic rather than the other way around. With the rise of a soundly based philology by the middle of the 19th century, a scientific etymological dictionary could be compiled, and this was provided in 1879 by Walter William Skeat. It has been kept in print in re-editions ever since but was superseded in 1966 by *The Oxford Dictionary of English Etymology*, by Charles Talbut Onions, who had worked many decades on it until his death. Valuable in its particular restricted area is J.F. Bense's *Dictionary of the Low-Dutch Element in the English Vocabulary* (1926–39).

Earliest English etymological dictionary

Two works are especially useful in showing the relation between languages descended from the ancestral Indo-European language—Carl Darling Buck's *Dictionary of Selected Synonyms in the Principal Indo-European Languages* (1949) and Julius Pokorny's *Indogermanisches etymologisches Wörterbuch* (1959). The Indo-European roots are well displayed in the summary by Calvert Watkins, published as an appendix to *The American Heritage Dictionary* mentioned earlier. Interrelations are also dealt with by Eric Partridge in his *Origins* (1958).

The pronouncing dictionary, a type handed down from

the 18th century, is best known in the present day by two examples, one in England and one in America. That of Daniel Jones, *An English Pronouncing Dictionary*, represents what is "most usually heard in everyday speech in the families of Southern English persons whose men-folk have been educated at the great public boarding-schools." Although he called this the Received Pronunciation (RP), he had no intention of imposing it on the English-speaking world. It originally appeared in 1917 and was repeatedly revised during the author's long life. Also strictly descriptive was a similar U.S. work by John S. Kenyon and Thomas A. Knott, *A Pronouncing Dictionary of American English*, published in 1944 and never revised but still valuable for its record of the practices of its time.

The "conceptual dictionary," in which words are arranged in groups by their meaning, had its first important exponent in Bishop John Wilkins, whose *Essay towards a Real Character and a Philosophical Language* was published in 1668. A plan of this sort was carried out by Peter Marc Roget with his *Thesaurus*, published in 1852 and many times reprinted and re-edited. Although philosophically oriented, Roget's work has served the practical purpose of another genre, the dictionary of synonyms.

Dictionaries of usage

The dictionaries of usage record information about the choices that a speaker must make among rival forms. In origin, they developed from the lists of errors that were popular in the 18th century. Many of them are still strongly puristic in tendency, supporting the urge for "standardizing" the language. The work with the most loyal following is Henry Watson Fowler's *Dictionary of Modern English Usage* (1926), ably re-edited in 1965 by Sir Ernest Gowers. It represents the good taste of a sensitive, urbane litterateur. It has many devotees in the U.S. and also a number of competitors. Among the latter, the most competently done is *A Dictionary of Contemporary American Usage* (1957), by Bergen Evans and Cornelia Evans. Usually the dictionaries of usage have reflected the idiosyncrasies of the compilers; but, from the 1920s to the 1960s, a body of studies by scholars emphasized an objective survey of what is in actual use, and these were drawn upon by Margaret M. Bryant for her book *Current American Usage* (1962). A small corner of the field of usage is dealt with by Eric Partridge in *A Dictionary of Clichés* (1940).

The regional variation of language has yielded dialect dictionaries in all the major languages of the world. In England, after John Ray's issuance of his first glossary of dialect words in 1674, much collecting was done, especially in the 19th century under the auspices of the English Dialect Society. This collecting culminated in the splendid *English Dialect Dictionary* of Joseph Wright in six volumes (1898–1905). American regional speech was collected from 1774 onward; John Pickering first put a glossary of Americanisms into a separate book in 1816. The American Dialect Society, founded in 1889, made extensive collections, with plans for a dictionary, but this came to fruition only in 1965, when Frederic G. Cassidy embarked on *A Dictionary of American Regional English* (known as *DARE*).

The many "functional varieties" of English also have their dictionaries. Slang and cant in particular have been collected in England since 1565, but the first important work was published in 1785, by Capt. Francis Grose, *A Classical Dictionary of the Vulgar Tongue*, reflecting well the low life of the 18th century. In 1859 John Camden Hotten published the 19th-century material, but a full historical, scholarly survey was presented by John Stephen Farmer and W.E. Henley in their *Slang and Its Analogues*, in seven volumes, 1890–1904, with a revised first volume in 1909 (all reprinted in 1971). For the present century, the dictionaries of Eric Partridge are valuable. Slang in the United States is so rich and varied that collectors have as yet only scratched the surface, but the work by Harold Wentworth and Stuart B. Flexner, *Dictionary of American Slang* (1960), can be consulted. The argot of the underworld has been treated in many studies by David W. Maurer.

Bilingual dictionaries

Of all specialized dictionaries, the bilingual group are the most serviceable and frequently used. With the rise of the vernacular languages during the Renaissance, translating to and from Latin had great importance. The Welshman in England was provided with a bilingual dictionary as early as 1547, by William Salesbury. Scholars in their analyses of language, as well as practical people for everyday needs, are anxious to have bilingual dictionaries. Even the most exotic and remote languages have been tackled, often by religious missionaries with the motive of translating the Bible. The finding of exact equivalents is more difficult than is commonly realized, because every language slices up the world in its own particular way.

Dictionaries dealing with special areas of vocabulary are so overwhelming in number that they can merely be alluded to here. In English, the earliest was a glossary of law terms published in 1527 by John Rastell. His purpose, he said, was "to expown certeyn obscure & derke termys concernynge the lawes of thys realme." The dictionaries of technical terms in many fields often have the purpose of standardizing the terminology; this normative aim is especially important in newly developing countries where the language has not yet become accommodated to modern technological needs. In some fields, such as philosophy, religion, or linguistics, the terminology is closely tied to a particular school of thought or the individual system of one writer, and, consequently, a lexicographer is obliged to say, "according to Kant," "in the usage of Christian Science," "as used by Bloomfield," and so on.

FEATURES AND PROBLEMS

Establishment of the word list. The goal of the big dictionaries is to make a complete inventory of a language, recording every word that can be found. The obsolete and archaic words must be included from the earlier stages of the language and even the words attested to only once (nonce words). In a language with a large literature, many "uncollected words" are likely to remain, lurking in out-of-the-way sources. The *OED* caught many personal coinages, but not "head-over-heelishness" (1882), "odditude" (1860), "pigstyosity" (1869), "white-chokerism" (1866), and other graceless jocularities. Also, the so-called latent words are a problem, when a lexicographer knows that a derivative word probably has been used, but he has no evidence for it. The *OED* had three quotations for "kindheartedness" but none for "kindheartedly," which any speaker of English would feel free to use. Some "ghost words" have arisen from the misreading of manuscripts and from misprints, and the lexicographer attempts to cast these out.

Problems in selecting a word list

Various large blocks of words have a questionable status. Both geographic names and biographical entries are selectively included in some dictionaries but are really encyclopaedic. More than 2,000,000 insects have been identified and named by entomologists, while names of chemical compounds and drugs may be almost as numerous. Trade names and proprietary names may number in the hundreds of thousands. Vogue suffixes like "-ism," "-ology," "-scope," or "-wise" are used by some with the freedom of a grammatical construction. These millions are beyond what any dictionary can be expected to include.

For the smaller-sized dictionaries, the editors attempt to choose the words that are likely to be looked up. They comb the scholarly works carefully and supplement them from files that they may have collected. They may decide to put derivative words at the end of entries as "run-ons" or to have all words strictly as separate alphabetical entries. The size is ultimately decided by the commercial consideration of how much can be put into a work that can be sold for a reasonable price and held readily in the hand. (Bulk also influences the size of the word list for unabridged dictionaries.)

The establishment of a word list involves many difficult technical problems. Linguists tend to use the terms morpheme, free form, bound form, lexeme, and so on, inasmuch as "word" is a popular term not suited to technical

use. A safe compromise is to use "lexical unit." This term allows the inclusion of set phrases (established groups) and idioms. Words having different etymological sources must be considered as different words. Thus "calf" in the sense of the young of a bovine animal came from Common Germanic, whereas "calf" for the fleshy back of the lower part of the leg came from Old Norse, perhaps from a Celtic source. A more difficult problem is found when a word entered the language at different points—such as "cookie," from the Dutch *koekje* "little cake," recorded in Scottish in 1701 in the form *cuckie*, then independently taken from the Dutch of the Hudson Valley in the form *cockie* in 1703, and perhaps independently taken into South African English from Afrikaans in the mid-19th century.

Spelling. Dictionaries have probably played an important role in establishing the conventions of English spelling. Dr. Johnson has received much credit for this, though he differed very little from his predecessors. He used the spelling "smoak" in the early part of his dictionary, but when he came to the entry itself, he changed it to "smoke," and this has prevailed. Noah Webster introduced some simplifications that have become accepted in American English. American dictionaries usually label the distinctive British spellings, such as "centre" and its class, "honour" and its class, "connexion," "gaol," "kerb," "tyre," "waggon," and a few others.

British and American spelling

The desire for uniformity is so great that popular variants are not welcomed; the very common "alright" is not yet approved, nor is the widespread "miniscule" for "minuscule." The *OED* is exceptional in listing the early variant spellings, showing that a common word like "good" has been spelled in 13 different ways, with seven more from Scottish usage. When the spelling reform movement was at its height, from the 1880s to about 1910, the dictionaries included the new forms, but in recent years these have been expunged. The graphic dress of the language is now so sacrosanct that dictionaries are used as authoritarian "style manuals" in matters of spelling, hyphenation, and syllabification.

Pronunciation. Dictionaries are more responsive to usage in the matter of pronunciation than they are in spelling. It is claimed that in the 19th century the Merriam–Webster dictionaries foisted a New England pronunciation on the United States, but in recent years many regional variations have been recorded. *Webster's Third New International* (1961) went to surprising lengths in its variants; perhaps its record is in giving 132 different ways of pronouncing "a fortiori."

The former practice of giving pronunciations as if the words were pronounced in isolation in a formal manner represented an artificiality that distorted language in use; recent dictionaries have marked pronunciation as it appears in continuous discourse. Furthermore, there has been a trend toward what has been called "democratization." In the word "government," for instance, it is recognized that many people do not pronounce an *n*, and some people actually say something like "gubb-munt." There is a constant battle between traditional spoken forms and spelling pronunciations.

Systems for indicating sounds

Since the alphabet is notoriously inadequate for recording the sounds of English, dictionaries are forced to adopt additional symbols. A system of using numerals over vowels was handed down from the 18th century, but that gave way to the diacritic markings of the Merriam–Webster series. The rise of the International Phonetic Alphabet (IPA) has offered another possibility, but the general public as yet finds it abstruse. Even more detailed symbols are needed in linguistic atlases and phonetic research. With considerable courage, Clarence L. Barnhart introduced the symbol schwa (ə) into *The American College Dictionary* (1947) for the neutral mid-central vowel, as at the beginning and end of "America," and the symbol has now become widely accepted. Although some systems are clumsier than others, the key does not matter much if it is applied consistently.

Etymology. The supplying of etymologies involves such difficult decisions for a lexicographer as whether words should be carried back into prehistory by means of reconstructed forms or the degree to which speculation should be permitted. A U.S. Romance scholar, Yakov Malkiel, has presented the notion that words follow "trajectories"—by finding certain points in the history of a word, one can link up the developments in form and meaning. The austere treatment of some words consists in saying "derivation unknown," and yet this sometimes causes interesting possibilities to be ignored.

Native stock and loanwords

A fundamental distinction is made in word history between the "native stock" and the "loanwords." There have been so many borrowings into English that the language has been called "hypertrophied." The traditional view is to regard the borrowings as a source of "richness." A historical dictionary does its best to ascertain the date at which a word was adopted from another language, but the word may have to go through a period of probation. Murray, the editor of the *OED*, listed four stages of the word "citizenship": the casual, the alien, the denizen, and the natural. The casuals may not be part of the language, as they appear only in travel writings and accounts of foreign countries, but a lexicographer must collect citations for them in order to record the early history of a word that may later become naturalized. Some words may remain "denizens" for centuries, Murray pointed out, such as "phenomenon" treated as Greek, "genus" as Latin, and "aide-de-camp" as French. When a word is borrowed, its etymology may be traced through its descent in its original language.

Some early philosophies have assumed that there is a mystic relation between the present use of a word and its origin and that etymology is a search for the "true meaning." The recognition of continuous linguistic change establishes, however, that etymology is no more than early history, sometimes as reconstructed on the basis of relationships and known sound changes. Ingenuity in etymologizing is dangerous, and even plausibility can be misleading, but ascertained fact has overriding importance. It is curious that recent slang is often more uncertain in its origin than words of long history.

Grammatical information. Dictionaries are obliged to contain the two basic kinds of words of a language—the "function words" (those that perform the grammatical functions in a language, such as the articles, pronouns, prepositions, and conjunctions) and the "referential words" (those that symbolize entities outside the language system). Each kind must be treated in a suitable way. Dictionaries have been much criticized for not including a sufficiency of grammatical information. It is usual to mark the part of speech, but not the categories of mass noun and count noun. (A mass noun, such as "milk" or "oxygen," cannot ordinarily be used in the plural, while a count noun is any noun that can be pluralized.) Such information is given in some dictionaries designed for teaching, and the technique could well be adopted more generally. The irregular inflections must be given, showing that one says "goose," "geese," but not "moose," "meese." Or in the verbs one says "walk," "walked," but "ride," "rode." It is usual to treat the different parts of speech as separate lexical entries, as in "to walk" and "to take a walk," requiring a parallel list of senses, but Edward Lee Thorndike, in his school dictionaries, experimented with grouping the parts of speech together when they had a similar sense.

Grammar and vocabulary

The relation of grammar to the vocabulary is the subject of considerable controversy among linguists. If one considers the analysis of language as one unified enterprise, then the grammar is central and the lexical units are inserted at some point in the analysis. Another view is that the division is into coordinate branches, such as phonology, syntax, and lexicon. Certainly lexicographers try to take advantage of all findings made by grammarians.

Sense division and definition. A language like English has so many complex developments in the senses—*i.e.*, the particular meanings—of its words that the task of the lexicographer is difficult. It is generally accepted that "meaning" is a suffusing characteristic of all language by definition, and the attempt to slice meaning into "senses" must be done arbitrarily by the person analyzing the lan-

guage. This is where collected contexts form the basis of the lexicographer's judgment. He sorts the quotations into piles on the basis of similarities and differences and he may have to discard "transitional" examples. Figurative developments, such as the "mouth" of a river or the "foot" of a hill, make complications in the relationships.

For the order in which the senses of words are given, the order of historical development has been chiefly used. For an old word like "earth," the information may be insufficient. The editors of the *OED* had to give up, because, they said, "Men's notions of the shape and position of the earth have so greatly changed since Old Teutonic times"; they were obliged to compromise with a logical order. Sometimes, but not always, a word seems to have a "core," or central, meaning from which other meanings develop. If the historical order is followed, the obsolete and archaic meanings may have to appear first; and, therefore, some popular dictionaries give the most important meaning first and work down to the rare and occasional meanings at the end. The so-called "semantic count," giving senses in order of frequency, has also been used.

There seems to be no one method that is best for defining all words. The lexicographer must use artistry in selecting the ways that will convey a sense accurately and succinctly. He attempts to find what is "criterial" in a particular meaning, but he can also give further detail until he runs into the area of the encyclopaedic.

In logical theory it would be ideal to have a "metalanguage" in which definitions could be stated, but nothing of the sort is available for popular use. A "defining vocabulary" can be established, and in school dictionaries the definitions use simple words. In the last analysis all definitions have to fall back on undefined terms (to be accepted like axioms) that symbolize first-order experience of life. In this connection the logician Willard Quine has argued that lexicography is basically concerned with synonymy.

Variations in language use

Usage labels. Part of the information that a dictionary should give concerns the restrictions and constraints on the use of words, commonly called usage labelling. There is great variation in language use in many dimensions—temporal, geographical, and cultural. The people who make a two-part division into "correct" and "incorrect" show that they do not understand how language works. The valuation does not lie in the word itself but in the appropriateness of the context. Therefore, it is preferable to be sparing in the use of labels and to allow the tone to become apparent from the illustrative examples. An important distinction was put forward in 1948 by an American philologist, John S. Kenyon, when he discriminated between "cultural levels," which refer to the degree of education and cultivation of a person, and "functional varieties," which refer to the styles of speech suitable to particular situations. Thus a cultivated person rightly uses informal or colloquial language when at ease with friends.

A lexicographer is faced with the difficult task of selecting a suitable set of labels. In the temporal categories, labels such as obsolete, obsolescent, archaic, and old-fashioned are dangerous, because some speakers have long memories and might use old words very naturally. The national labels are problematical, because words move easily from one branch of the language to another. The word "blizzard," for instance, is no doubt an Americanism in origin, but, since the 1880s, it has been so well-known over the English-speaking world that a national label would be misleading. The label "dialect" or "regional," either for England or America, offers many problems, for alleged "boundaries" are permeable. The label "colloquial" was much misunderstood, and now "informal" is often used in its place. There may be a "poetic vocabulary" that needs labelling, and few people will agree on any definition of "slang."

It is revealing that in the Merriam-Webster *Third New International* under the word "cock-eyed," marked "slang," one of the quotations is by a careful stylist named Jacques Barzun; in order to use effective English, as he does, this cultivated writer is willing to draw upon slang. Some would argue that in marking the use as "slang," the Merriam–Webster staff was not sufficiently "permissive."

Some dictionaries wisely include special paragraphs on the constraints of usage, sometimes as a "synonymy" and sometimes as a "usage note."

Illustrative quotations. Dictionaries of the past have copied shamelessly from one to another, but the collecting of a file of illustrative quotations makes possible a fresh, original treatment. Scholarly works like the OED and its supplementations follow the canon of always using the earliest quotation and the latest for an obsolete word; in between, the quotations are selected for revealing facets of usage or for "forcing" a meaning. The criterion of use by only the best writers does not hold for a truly historical dictionary, because a "low" source may be especially revealing. The giving of exact source citations is not a matter of pedantry but establishes the scientific basis by which others can check the evidence. A different set of quotations, accurately attested, might have led to a different treatment. Thus the phrase "illustrative quotation" is something of a misnomer, for the quotations are more than "illustrative"; they form the basic evidence from which conclusions are drawn. It is the work of the editor to decide when the collections are sufficient—"ripe," as it were—to move from the collecting stage to the editing stage.

Citation of sources

A small-sized dictionary may advantageously use made-up sentences, because an aptly framed "forcing" context can tell more than a definition. In fact, the habitual collocations of a word (the surrounding words with which it usually appears) may be revealing of the nature of a word. "Dictionaries of collocations" may be a step forward in future lexicography.

Technological aids. The development of machine aids, such as the computer, has been heralded by some as ushering in a new era in lexicography. Although the computer can do well in many tasks of great drudgery—mechanical excerpting of texts, alphabetizing, and classifying by designated descriptors—it is limited to what a human being tells it to do. It is difficult for a computer to sort out homographs—separate words that are spelled alike; and, at the editing stage, the delicate decisions must be humanly made.

Uses for computers

The computer can be used to good advantage in the compilation of concordances of individual authors or of limited texts, and then one type of dictionary could be made by a summation of concordances. Such a procedure, with a large body of literature like that of the Renaissance, would overwhelm an editor. More feasible, perhaps, is the establishment of a computerized archive that would never be published, but would serve as a storehouse from which, by advanced retrieval methods, the desired information could be called forth at will. The *Trésor de la langue française* of Nancy, already mentioned, is a step in this direction.

Attitudes of society. Without a doubt, dictionaries have been a conservative force for many hundreds of years, not only in countries that have had an official academy that has the national language as part of its province but also in the English-speaking countries, in which academies have been spurned. Well-entrenched popular attitudes account for this. A Neoplatonic outlook assumes that there exists an ideal form of language from which faltering human beings have departed and that dictionaries might bring people closer to the perfect language. Also, there is a widespread "yearning for certainty," a seeking for guidance amid the wilderness of possible forms. Thus, people welcome self-proclaimed "supreme authorities."

Americans have had additional reasons for their homage to the dictionary. In colonial times Americans felt themselves to be far from the centre of civilization and were willing to accept a book standard in order to learn what they thought prevailed in England. This linguistic colonialism lasted a long time and set the pattern of accepting the dictionary as a "lawgiver." In 1869, a cultural leader declared: "Upon the proper spelling, pronunciation, etymology, and definition of words, a dictionary

might be made to which high and almost absolute authority might justly be awarded." In this vein teachers have taken pains to inculcate "the dictionary habit" in their pupils. Rather than observe the language around them, as Englishmen commonly do, Americans give up their autonomy and fly to a dictionary to settle questions on language. This call for dogmatic prescription has been a source of uneasiness to lexicographers, most of whom now argue that all they can do legitimately is to describe how the language has been used.

Observance of taboos

Social attitudes have affected the dictionaries also in the enforcement of certain taboos. Certain words commonly called obscene have been omitted, and, thus, irrational taboos have been strengthened. If the sex words were given in their alphabetical place, with suitable labels, the false attitudes in society would more readily be cleansed. A perennial problem in lexicography is the treatment of the terms of ethnic insult, such as "dago," "kike," and "wop." There is constant social pressure for leaving them out, and some dictionaries have succumbed to it, but it may be that an enlightened attitude shows that the open discussion of prejudices is the best way of getting rid of them.

The greatest value of a dictionary is in giving access to the full resources of a language and as a source of information that will enhance free enjoyment of the mother tongue.

MAJOR DICTIONARIES

For the English language the important dictionaries have already been cited in the appropriate sections; but the supreme achievement represented by the *OED* should be emphasized again. The major dictionaries in some other languages may be mentioned here.

Dictionaries of the romance languages

For the French language, the Académie's dictionary is now in its eighth edition (1931–35) and manifests conservative views about the vocabulary, but three other works are actually more serviceable—the *Petit Larousse: dictionnaire encyclopédique pour tous* (1959); a new edition of the famous Littré, *Dictionnaire de la langue française* (1956–58), seven volumes; and a splendid new work, Paul Robert, *Dictionnaire alphabétique et analogique de la langue française* (1960–64), six volumes. For French etymology alone, the standard work is Walther von Wartburg, *Französisches etymologisches Wörterbuch*, nearing completion in 1970 in volume 18, with a few gaps to be filled.

Among other Romance tongues, Italian has had many dictionaries. The Accademia della Crusca of Florence furnished its *Vocabolario* in a first edition in 1612, but the edition begun in 1863 bogged down at the letter *O* in 1923, and a successor work, begun in 1941, has not gone far. There is also the dictionary by G. Devoto and G.C. Oli, *Dizionario della lingua Italiana* (1971). Following the model of the OED is the still uncompleted *Grande dizionario della lingua italiana* (1961), edited by Salvatore Battaglia. Very serviceable to English speakers is the *Italian Dictionary* of Alfred Hoare (1915; second edition, 1925) and that of Barbara Reynolds, begun in 1962, and still under way. For Spanish, the Real Academia Española in Madrid has done well since its first edition in 1726–39. At present the 18th edition, from 1956, is available. Contributions from New World Spanish need further scholarly treatment.

For the German language, the great dictionary begun by the brothers Grimm, completed in 1960, is to be reedited in a project that will take many years; but, meanwhile, a standard work is that of Hermann Paul, *Deutsches Wörterbuch*, which first appeared in 1897 but is now available in a sixth edition (1968). The national dictionaries in the Scandinavian countries were mentioned above, but a work done with special scholarly skill is noteworthy: Einar Haugen, editor in chief, *Norwegian English Dictionary* (Madison, Wisconsin [Oslo printed], 1965), dealing with the two official languages of Norway, Bokmål and Nynorsk. Another form of a Germanic language, Afrikaans, which developed from the Dutch transplanted to South Africa in the 17th century, has a big dictionary under way, *Woordeboek van die Afrikaanse taal*, begun at Pretoria in 1950 as a collaboration of the best scholars in South Africa. A full dictionary of Yiddish requires profound scholarship, and this was provided by Uriel Weinreich in *Modern English-Yiddish, Yiddish-English Dictionary* (1968).

Special problems of Greek dictionaries

Greek lexicography offers special difficulties because of the long range of illustrious literature that must be covered and the split in recent centuries between Katharevusa, the literary language, and Demotic, the language of everyday life. For the English-speaking world, the standard work for Ancient Greek is by Henry George Liddell and Robert Scott, *A Greek-English Lexicon*, published in a first edition in 1843, but now available in a ninth edition, 1925–40. A full dictionary of Demotic, edited by Demetrius Georgacas at the University of North Dakota, is still in the project stage. For Russian the Soviet Academy of Arts has produced a useful work in four volumes (1957–61), but a more detailed one has been in progress since 1950, reaching *F* in 1964. The Royal Irish Academy is at work on a definitive dictionary of Irish, but only "contributions" and certain parts are so far available. Many linguists have attempted to cover Arabic; probably the most useful work is that of Hans Wehr, as translated and edited by J. Milton Cowan, *A Dictionary of Modern Written Arabic* (1961). For Japanese, the standard source is the *Dai-jiten* ("Great Dictionary"), issued at Tokyo in 26 volumes (1934–36). The best known Chinese dictionary, *Tz'u hai*, was revised in 1969 and published in Taipei, Taiwan.

Titles of other works are to be found in the bibliographies listed below, especially in Constance Winchell's *Guide*.

BIBLIOGRAPHY. For the best list of dictionaries, see CONSTANCE M. WINCHELL, *Guide to Reference Books*, 8th ed., pp. 91–132 (1967; 1st suppl., *1965–1966*, 1968; and 2nd suppl., *1967–1968*, 1970). See also ROBERT L. COLLISON, *Dictionaries of Foreign Languages* (1957); A.J. WALFORD (ed.), *A Guide to Foreign Language Grammars and Dictionaries*, 2nd ed. (1967); WOLFRAM ZAUNMULLER, *Bibliographisches Handbuch der Sprachwörterbücher* (1958); GERT A. ZISCHKA, *Index Lexicorum* (1959); and *Foreign Language-English Dictionaries*, 2 vol. (Library of Congress, Reference Department, 1955). For dictionaries published in Communist countries, see DANUTA RYMSZA-ZALEWSKA (ed.), *Bibliography of Dictionaries* (1965). Dictionaries of Americanisms and of slang are well covered by W.J. BURKE, *The Literature of Slang*, pp. 2–11 (1939).

History: For the history of classical dictionaries, see JOHN EDWIN SANDYS, *A History of Classical Scholarship*, 3rd ed., vol. 1, pp. 295–407 (1921). An old standard survey is JAMES A.H. MURRAY, *The Evolution of English Lexicography* (1900). See also MITFORD M. MATHEWS, *A Survey of English Dictionaries* (1933, reprinted 1966); and JAMES ROOT HULBERT, *Dictionaries: British and American*, rev. ed. (1968). For excellent scholarly details in their areas, see DEWITT T. STARNES, *Renaissance Dictionaries: English-Latin and Latin-English* (1954); DEWITT T. STARNES and GERTRUDE E. NOYES, *The English Dictionary from Cawdrey to Johnson, 1604–1755* (1946); and JAMES H. SLEDD and GWIN J. KOLB, *Dr. Johnson's Dictionary: Essays in the Biography of a Book* (1955). For the history of the *Oxford English Dictionary*, see WILLIAM A. CRAIGIE, "Historical Introduction," in the *Supplement* (1933), transferred to vol. 1 in the re-issue of 1933; HANS AARSLEFF, "The Early History of the *Oxford English Dictionary*," *Bulletin of the New York Public Library*, 66:417–439 (1962), and *The Study of Language in England, 1780–1860* (1967), especially ch. 6. For American dictionaries, see JOSEPH HAROLD FRIEND, *The Development of American Lexicography, 1798–1864* (1967). The documents on the controversy over the Merriam-Webster *Third New International Dictionary* are collected by JAMES H. SLEDD and WILMA R. EBBITT in *Dictionaries and That Dictionary* (1962). For discussions of the technical problems arising in lexicography, see FRED W. HOUSEHOLDER and SOL SAPORTA (eds.), *Problems in Lexicography*, 2nd ed. (1967), papers of a conference held in 1960—specially practical is the paper by CLARENCE L. BARNHART, "Problems in Editing Commercial Monolingual Dictionaries," pp. 161–181; LADISLAV ZGUSTA, *Manual of Lexicography* (1971); and ALLEN WALKER READ, "Approaches to Lexicography and Semantics," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 10, pp. 145–205 (1972). An "International Conference on Lexicography in English" was held in New York City, June 5–7, 1972; its proceedings are published in the *Ann. N.Y. Acad. Sci.* (1973).

(A.W.Re.)

Diderot, Denis

Man of letters and philosopher, Denis Diderot straddled the 18th century as the chief editor of the French *Encyclopédie*, the testament of the Age of Enlightenment, and by virtue of the originality of his contributions in the fields of philosophy and ethics, dramatic and aesthetic theory, literary criticism, fiction, scientific speculation, and politics.

Giraudon



Diderot, oil painting by Louis-Michel van Loo, 1767. In the Louvre, Paris.

Youth and marriage. He was born at Langres on October 5, 1713, the son of a widely respected master cutler, whose worth he recognized in later life. He was tonsured in 1726, though he did not in fact enter the church, and was first educated by the Jesuits at Langres. From 1729 to 1732 he studied in Paris at the Collège d'Harcourt or at the Lycée Louis-le-Grand or possibly at both these institutions, and he was awarded the degree of master of arts in the University of Paris on September 2, 1732. He then studied law as an articled clerk in the office of Clément de Ris but was more interested in languages, literature, philosophy, and higher mathematics. Of the period 1734 to 1744 comparatively little is known. He dropped an early ambition to enter the theatre and, instead, taught for a living, led a penurious existence as a publisher's hack, and wrote sermons for missionaries at 50 *écus* each. At one time he seems to have entertained the idea of taking up an ecclesiastical career, but it is most unlikely that he entered a seminary. Yet his work testifies to his having gone through a religious crisis, and he progressed relatively slowly from faith to deism and to atheism. That he led a disordered existence is made clear in his posthumously published novel, *Le Neveu de Rameau* ("Rameau's Nephew"). He frequented the coffeehouses, particularly the Régence and the Procope, where he met the philosopher Jean-Jacques Rousseau in 1741 and established a friendship with him that was to last for 15 years, until it was broken by a quarrel.

Meeting with Rousseau

In 1741 he also met Antoinette Champion, daughter of a linendraper, and on November 6, 1743, married her—secretly, because of his father's disapproval. The relationship, as surviving letters show, was based on romantic love, but the marriage was not a happy one owing to incompatible interests. The bond held, however, partly through a common affection for their daughter, Angélique, sole survivor of three children, who was born in 1753 and whom Diderot eventually married to Albert de Vandeul, a man of some standing at Langres. Diderot lavished care over her education, and she eventually wrote a short account of his life and classified his manuscripts. There is a bourgeois solidity about the pattern of his family life—he always kept himself informed of events at Langres, and he remained deeply attached to his sister Denise—which is in marked contrast with that of the aristocratic Voltaire and the plebeian Rousseau.

Mature career. In order to earn a living, he undertook translation work and in 1745 published a free translation of the *Inquiry concerning Virtue* by the 3rd Earl of Shaftesbury, whose fame and influence he spread in France. His *Pensées philosophiques* (1746; *Philosophic Thoughts*, 1916), an original work with new and explosive anti-Christian ideas couched in a vivid prose, contains many passages directly translated from or inspired by Shaftesbury. The proceeds of this publication, as of his allegedly indecent novel *Les Bijoux indiscrets*, were used to meet the demands of his mistress, Madeleine de Puisieux, with whom he broke a few years later during his detention at Vincennes. In 1755 he met "Sophie" Volland, with whom he formed an attachment that was to last over 20 years. The liaison was founded on common interests, natural sympathy, and a deepening friendship. His correspondence with Sophie, together with his other letters, forms one of the most fascinating documents on Diderot's personality, enthusiasms and ideas and on the intellectual society of Louise d'Épinay, F.M. Grimm, the Baron d'Holbach, Ferdinando Galiani, and other deistic writers and thinkers (Philosophes) with whom he felt most at home. Through Rousseau, Diderot met Condillac, the philosopher, and for a time the three friends dined together at the Panier Fleuri. Diderot set great store, in his personal life, by friendship as a basis for the establishment of moral values. This was brought out in his short story "Les Deux Amis de Bourbonne" ("The Two Friends of Bourbonne"), in which the Germans Goethe and Schiller were to become interested, and in the conclusion of "Ceci n'est pas un conte" ("This Is No Yarn").

Correspondence with "Sophie" Volland

The Encyclopédie. In 1745 the publisher André Le Breton approached Diderot with a view to bringing out a French translation of Ephraim Chambers' *Cyclopaedia*, after two other translators had withdrawn from the project. Diderot undertook the task with the distinguished mathematician Jean Le Rond d'Alembert as coeditor and profoundly changed the nature of the publication, broadening its scope and turning it into an important organ of radical and revolutionary opinion. He gathered around him a team of dedicated litterateurs, scientists, and even priests, many of whom, as yet unknown, were to make their mark in later life. All were fired with a common purpose: to further knowledge and, by so doing, strike a resounding blow against reactionary forces in church and state. As a *dictionnaire raisonné* ("rational dictionary"), the *Encyclopédie* was to bring out the essential principles and applications of every art and science. The underlying philosophy was rationalism and a qualified faith in the progress of the human mind. In 1749 he published the *Lettre sur les aveugles* (trans. *c.* 1750, *An Essay on Blindness*), remarkable for its proposal to teach the blind to read through the sense of touch, along lines that Louis Braille was to follow in the following century, and for the presentation of the first step in his evolutionary theory of survival by superior adaptation. This daring exposition of the doctrine of materialist atheism with its stress on human dependence on sense impression led to Diderot's arrest and incarceration in the prison of Vincennes for three months. His work on the *Encyclopédie*, however, was not interrupted for long, and in 1750 he published his *Prospectus*, which d'Alembert expanded into the momentous *Discours préliminaire* (1751). The history of the *Encyclopédie*, from the publication of the first volume in 1751 to the distribution of the final volumes of plates in 1772, was checkered, but ultimate success was never in doubt. A critical moment occurred in 1758 on the publication of the seventh volume when d'Alembert resigned on receiving warning of trouble and after reading Rousseau's attack on his article "Genève." Another serious blow came when the philosopher Helvétius' book *De l'esprit* ("On the Mind"), said to be a summary of the *Encyclopédie*, was condemned to be burned by the Parlement of Paris; the work itself was suppressed. Untempted by Voltaire's offer to have the publication continued outside France, Diderot held on in Paris with great tenacity. His heart was nearly broken, however, by the discovery in 1764 that Le Breton

Arrest and incarceration in Vincennes

had secretly removed compromising material from the corrected proof sheets of about ten folio volumes. The censored passages, though of considerable interest, would not have made an appreciable difference on the impact of the work. To the 17 volumes of text and 11 volumes of plates (1751–72), Diderot contributed innumerable articles partly original, partly derived from varied sources, especially on the history of philosophy ("Eclectisme" ["Eclecticism"]), aesthetics ("Beau" ["The Beautiful"]), and the mechanical arts. He was an energetic general director and supervised the illustrations for 3,000 to 4,000 plates of exceptional quality, which are still prized by historians today. The completion of the *Encyclopédie* left Diderot without a source of income. To relieve him of financial worry, Catherine the Great of Russia first bought his library through an agent in Paris, requesting him to retain the books until she required them, and then appointed him librarian on an annual salary for the duration of his life. Diderot went to St. Petersburg in 1773 to thank her and was received with great honour and warmth. He wrote for her the *Plan d'une université pour le gouvernement de Russie* ("Plan of a University for the Government of Russia"). He stayed five months, long enough to become disillusioned with enlightened despotism as a solution to social ills and for his political ideas to harden, as evidenced in *Observations sur les instructions de Sa Majesté Impériale aux députés* ("Observations on Her Imperial Majesty's Instructions to Her Deputies"). It is now known that some of the most revolutionary passages in the writer and propagandist Guillaume Raynal's *Histoire des deux Indes* ("History of the Indies") actually came from Diderot's pen, as did arguments to legitimize revolution.

Patronage of Catherine the Great

Philosophical and scientific works. In 1751 Diderot published *Lettre sur les sourds et muets* ("Letter on the Deaf and Dumb"), which studies the function of language and deals with points of aesthetics, and in 1754 he published the *Pensées sur l'interprétation de la nature* ("Thoughts on the Interpretation of Nature"), acclaimed as the method of philosophical inquiry of the 18th century; however, he published few other works in his lifetime. His writings, in manuscript form, were known only to his friends and the privileged correspondents of the *Correspondance littéraire*, a sort of private newspaper edited by Baron Grimm (1759–81) that was circulated in manuscript form. Among his philosophical works, special mention may be made of *L'Entretien entre d'Alembert et Diderot* ("Conversation Between d'Alembert and Diderot"), *Le Rêve de d'Alembert* ("D'Alembert's Dream"), and the *Eléments de physiologie*. In all these works, he developed his materialist philosophy, foreshadowing the evolutionary doctrine of Charles Darwin and evolving the first modern theory of the cellular structure of matter. Though his speculations in the field of science are of great interest, it is the dialectical brilliance of their presentation that is exceptional. His ideas, often propounded in the form of paradox, and invariably in dialogue, stem from a sense of reality and a profound understanding of the complexities and contradictions inherent in human nature. He developed a theory of dreams that was later to impress Freud.

Materialist philosophy

Essays, novels, and plays. His essays, among them "Regrets sur ma vieille robe de chambre" ("Regrets over My Old Bathrobe") and "Entretien d'un père avec ses enfants" ("Conversation of a Father with His Children"), based on personal experience, have the qualities of form and style of his short stories and his novels: *La Religieuse* ("The Nun"), *Jacques le fataliste*, and *Le Neveu de Rameau*. *Jacques le fataliste* is in the tradition of the picaresque novel and of the "philosophical tale." In spite of his scientific determinism, Diderot's philosophical standpoint is ambivalent, as is his ethical standpoint in *Le Neveu de Rameau*, which contains a satire of contemporary society by offering a vigorous dramatic sketch of a parasite and an eccentric amoralist. In the *Supplément au voyage de Bougainville* ("Supplement to Bougainville's Voyage"), Diderot presented his idea of a free society based on tolerance and sexual liberty.

His major plays, *Le Fils naturel* (trans. 1767, *Dorval; or, The Test of Virtue*) and *Le Père de famille* ("The Father of the Family"), make tedious reading today. His theories on drama, expounded in *Entretiens sur le fils naturel* ("Discussion on the Natural Son") and *Discours sur la poésie dramatique* ("Discourse on Dramatic Poetry"), were to exercise a determining influence on the German dramatist Gotthold Lessing, whose famous *Hamburgische Dramaturgie* ("Dramatic Notes from Hamburg") appeared in 1767–69. Diderot sought greater realism on the stage through the presentation of a serious bourgeois drama and a greater moral and social impact on the spectator by showing characters in their professions and milieu. He urged modifications in stagecraft and décor and hoped to move audiences through *tableaux vivants*. In his *Paradoxe sur le comédien*, Diderot argued that great actors, like great poets, are insensitive and must remain fabulous puppets. Although he wrote literary criticism, it is as the first great art critic, covering the salons, or annual art exhibitions, for the *Correspondance littéraire*, that he is best remembered. His analysis of art, artists, and the technique of painting, together with the excellence of his taste and his style, have won him posthumous fame; his *Essai sur le peinture* ("Essay on Painting"), especially, was admired by Goethe and later by the 19th-century poet and critic Charles Baudelaire.

Art criticism

Late life and works. In 1774 Diderot, now old and ill, worked on a refutation of Helvétius' work *De l'homme* (1772; "On Man"), which was an amplification of the destroyed *De l'esprit*. He wrote *Entretien d'un philosophe avec la Maréchale* ("Conversation with the Maréchale") and published in 1778 *Essai sur les règnes de Claude et de Néron* ("Essay on the Reigns of Claudius and Nero"). Usually known as *Essai sur la vie de Sénèque* ("Essay on the Life of Seneca"), the work may be regarded as an apologia of the Roman satirist and philosopher. Diderot's intimate circle was dwindling. Mme d'Épinay and d'Alembert died before him; only Grimm and Baron d'Holbach remained. Slowly Diderot retired into the shell of his own personal and family life. The death of Sophie Volland in February 1784 was a great grief to him; he survived her by a few months, dying of coronary thrombosis on July 30, 1784, in the house in the rue de Richelieu that Catherine the Great had put at his disposal. Apocryphally, his last words were: "Le premier pas vers la philosophie, c'est l'incrédulité" ("The first step toward philosophy is incredulity"). Through the intervention of his son-in-law, he was buried in consecrated ground at Saint-Roch.

MAJOR WORKS

PHILOSOPHY AND SCHOLARSHIP: *Pensées philosophiques* (1746); *Lettre sur les aveugles* (1749); *Prospectus* (1750), to the *Encyclopédie; Lettre sur les sourds et muets* (1751); *Encyclopédie*, 17 vol. of text and 11 vol. of plates (1751–72), edited by Diderot and Jean d'Alembert and containing many articles by Diderot; *Pensées sur l'interprétation de la nature* (1754); *Le Rêve de d'Alembert* (written 1769, published 1830); *L'Entretien entre d'Alembert et Diderot* (written 1769, published 1830); *Eléments de physiologie* (1774–80).

LITERARY WORKS: (NOVELS): *Les Bijoux indiscrets* (1748); *La Religieuse* (written 1760, published 1796); *Jacques le fataliste* (written 1773, published 1796); *Le Neveu de Rameau* (written between 1761 and 1774, translated into German by Goethe 1805, French edition 1821, authentic text 1891); *Supplément au voyage de Bougainville* (written 1772, published 1796). (SHORT STORIES): "Les Deux Amis de Bourbonne" (1773); "Ceci n'est pas un conte" (written 1772, published 1798). (PLAYS): *Le Fils naturel* (published 1757; performed 1771); *Le Père de famille* (1758; performed 1761); *Est-il bon? Est-il méchant?* (performed 1781, published 1834). (ESSAYS AND TREATISES): *L'Histoire et le secret de la peinture en cire* (1755); *Entretiens sur le fils naturel* (1757); *Discours sur la poésie dramatique* (1758); *Éloge de Richardson* (1761); *Essai sur la peinture* (written 1765, published 1796); *Réflexions sur Terence* (1762); "Regrets sur ma vieille robe de chambre" (1772); *Entretien d'un père avec ses enfants* (1773); *Paradoxe sur le comédien* (written 1773–78, published 1830); *Plan d'une université pour le gouvernement de Russie* (published 1813–14), written for Catherine the Great; *Observations sur les instructions de Sa Majesté Impériale aux députés* (1774); *Réfutation de l'ouvrage d'Helvétius intitulé l'homme* (published 1875); *Essai sur les règnes de Claude et de Néron* (1778), usually called the *Essai sur la vie de Sénèque*.

TRANSLATIONS: Translations into English of a number of Diderot's individual works have been published but most of them so long ago as to be difficult to obtain. More recent translations include the following: *Denis Diderot: Selections*, ed. with introduction by E. Herriot (1953); *The Paradox of Acting*, trans. by W.H. Pollock (1957); *Diderot, Interpreter of Nature: Selected Writings*, trans. by Jean Stewart and Jonathan Kemp, and ed. with introduction by Jonathan Kemp, 2nd ed. (1963); *The Wigmaker's Art in the 18th Century: A Translation of the Section on Wigmaking in the 3rd Edition (1776) of the Encyclopédie . . .*, ed. by J. Stevens Cox (1965); Diderot's *Selected Writings*, trans. by Derek Coltman, and ed. with introduction and notes by Lester G. Crocker (1966); *The Nun*, trans. by Marianne Sinclair, with introduction and afterword by Richard Griffiths (1966); *The Encyclopédie of Diderot and d'Alembert: Selected Articles*, ed. by J. Lough (1969).

BIBLIOGRAPHY. A.M. WILSON, *Diderot: The Testing Years, 1713–1759* (1957), is the standard and most scholarly biography for the years covered. LESTER G. CROCKER, *Diderot, the Embattled Philosopher*, rev. ed. (1966), provides both a biographical and critical general study. JOHN (LORD) MORLEY, *Diderot and the Encyclopaedists*, 2 vol. (1878, reprinted 1923), is still of value. In the absence of a new critical edition (in preparation), the *Oeuvres complètes de Denis Diderot*, ed. by J. ASSEZAT and M. TOURNEUX, 20 vol. (1875–77); and the selected *Oeuvres* in five volumes in the "Classiques Garnier" are the best available.

(Ro.N.)

Diesel, Rudolf

Though best known for his invention of the pressure-ignited heat engine that bears his name, Rudolf Diesel was also an eminent thermal engineer, a connoisseur of the arts, a linguist, and a social theorist. As a one-man creative community, Diesel epitomized the confluence of the fine arts, the sciences, both "pure" and "social," and the creative pragmatics of mechanical engineering and invention. He described his artistic self as *ein Glückspilz* ("a lucky mushroom"); his career as an inventor as "enslavement to despair and ecstasy"; his crusade for peace and labour reform as "stubborn proselytizing by a chronic victim of petty and prejudiced nationalisms." During the last two decades of his life his informed admirers and critics alike came to regard his many-sided creative skills and his no less amazing ambivalences as stage props for one of the most memorable mechanical inventions of the 19th century.

By courtesy of the Deutsches Museum, Munich



Diesel, 1883.

Inventions

Diesel's inventions, which ranged from the fantasy of a "universal sun motor" to the refinement of ice ready-frozen in restaurant table bottles, have three points in common: they relate to heat transference by natural physical processes or laws; they involve markedly creative mechanical design; and they were initially motivated by the inventor's concept of sociological needs. Diesel originally conceived the diesel engine as a facility, readily adaptable in size and costs and utilizing locally available fuels, to enable independent craftsmen and artisans better to endure the powered competition of large industries that then virtually monopolized the predominant power source—the oversized, expensive, fuel-wasting steam engine. He envisaged making his rational heat engine, to adapt his own words, as big as a hut or as small as a hat, suitable for fuels from shale oils to coal or palm oils to surplus butter. He projected the engine principle as a solvency saver for all such small but useful producers as watchmakers, jewelers, dentists and dental technicians, cobblers, toy-makers, etc. His broader goal was socio-economic justice and balance, to save the engine user from being made its "tending wage slave."

Childhood and education. Diesel was born in Paris on March 18, 1858, the only son of Theodor, a Bavarian immigrant leather craftsman, and Elise Strobel Diesel, a German-born governess and language tutor. During his Paris childhood, despite his sensitivity and his being teased as a "pretty little German pig" by chauvinistic schoolmates, Rudolf won grammar school honours and, at 12, admission to the *École Primaire Supérieure*, then the most highly esteemed secondary school in Paris.

This triumph was negated by the outbreak of the Franco-Prussian War. By the French War Ministry's decree of August 28, 1870, the Diesels, as "undesirable aliens," were rounded up by police and shipped to neutral asylum in London. Rudolf was rescued from the ensuing upset and impoverishment by a teacher-cousin in Augsburg, his father's home town. The cousin, Christoph Barnickel, wrote to invite Rudolf to enter the Royal County Trade School, where Barnickel taught, as a "deserving refugee pupil and ward."

During the ensuing three years, Rudolf again suffered derision, this time as a "pretty French pig." Nevertheless, he led his class, set an all-time high scholastic record, and, despite being a "nonnational," won a scholarship to the Technische Hochschule of Munich. In his four years there he repeated his attainment of unprecedentedly high grades. With his own earnings as an undergraduate tutor and by making and selling second-degree geometrical surfaces (ellipsoids, hyperboloids, etc.), he managed to rent himself a piano, took piano and voice lessons, attended operas, concerts, art exhibitions, and lectures in Munich. He left the Hochschule as a special protégé of Carl von Linde, its most renowned professor and pioneer in mechanical refrigeration.

Life in Paris. After two years as a mechanic and parts designer at the Sulzer Machine Works of Winterthur, Switzerland, by then self-labelled as the steam engine capital of the world, Diesel returned to Paris as a thermal engineer, installer, and salesman for the Linde Refrigeration Enterprises, then operating in five countries. On the Left, or south, Bank he found friends among international artist groups and, in more sumptuous neighbourhoods, among more affluent professional men. With his exceptional personal charms, not to mention his talents as a drawing-room artiste, he became established both as a devotee of fine arts and an internationalist.

When his father, who had earlier abandoned his leather craft to open shop as a *Heilmagnetiseur*, or "Magnetic Health Builder," exhorted his son to act more like a real German, Rudolf replied that the Diesel forebears had never been real Germans. Rather, they were all Saale Valley Slavs, or Thuringians, a gentle, poetic clan who had adopted the name Diesel from earlier Slav names, including *Dössel* and *Tüssel*, on being converted as Lutherans. Since his 13th birthday, Rudolf had been a rather devout Lutheran and believed the denomination to be symbolic of international religious liberation.

At the age of 25, Diesel married Martha Flasche, a German-raised governess he had met in the home of Ernest Brandes, an international merchant friend. The following year (1884) Martha bore a son, Rudolf, who grew up to be an abstract artist and mystic; two years later a daughter, Hedy; and in 1889 a second son, Eugen, who became a writer on, and professor of, philosophy.

During 1885 the seeker of a "power provider for global justice" set up his first shop-laboratory in Paris and began his 13-year ordeal of creating his distinctive engine.

Pacifism

A dedicated pacifist, Diesel also envisaged what he privately termed his "Black Mistress" as an implement for sustaining peace. In 1888, his second year of intensive work on the engine, the inventor barely escaped alive when ammonia gas, being tested as a fuel, exploded. Almost instantly he set out to exploit the near catastrophe as a war squelcher. He proposed to fill small, readily breakable glass vials with contact-explosive gas and supply the chemical fireworks to war ministries for use in lieu of lethal bombs and bullets. Thus, all battles could be made shams, entertaining at least to the "militant imbeciles who revel in competitive uproars." When the French Patent Office turned it down, Diesel offered his proposal to Count Georg Münster, then the imperial German consul general to Paris. The rarely smiling Count is said to have laughed aloud. By that time, however, Diesel's 30th year, his stature as a thermal engineer, inventor, and internationalist was above being the subject of derision. One cause was his extraordinary scholastic brilliance. After three gruelling and impoverishing years of preliminaries in Paris, however, Diesel was obliged to take temporary employment with the Linde Enterprises, this time in Berlin. From there, late in 1892, he gained a German development patent. On the strength of this and about a dozen trunkfuls of immaculate drawings and tabulations and on promise of eventual manufacturing rights, three sponsors, the Maschinenfabrik firm at Augsburg and the firms of Sulzer and Krupp, came to the inventor's support.

Invention of the Diesel engine. At Augsburg, on August 10, 1893, Diesel's prime model, a single ten-foot iron cylinder with a flywheel at its base, ran on its own power for the first time. The pressure needle promptly shot to 80 atmospheres, at the time the highest mechanically created pressure ever recorded. But at that point the indicator plate exploded. Again the smock-clad inventor dodged barely in time to save his head. Next, a carefully revised model ran on its own power for one minute on February 17, 1894. Expert onlookers termed that an epochal minute, a proof of momentous potentials of a still imperfect engine.

Diesel spent two more years at improvements and on the last day of 1896 demonstrated another model with the spectacular if theoretical mechanical efficiency of 75.6 percent, in contrast to the then prevailing efficiency of the steam engine—10 percent or less. Although commercial manufacture was delayed another year and even then begun at snail's pace, by 1898 Diesel was a millionaire from franchise fees, in great part international.

By 1904, when Diesel first toured the United States as an elite lecturer, his appallingly bad financial management was returning him to respectable poverty. Publication of his two-volume work on social philosophy, *Solidarismus*, did not relieve his financial difficulties; it sold fewer than 200 copies. Diesel, however, was internationally recognized as the pre-eminent pioneer of the power age and a champion of the fine arts. His engines were powering pipelines, electric and water plants, automobiles and trucks, and marine craft and soon thereafter were used in mines, oil fields, factories, transoceanic shipping, and elsewhere. But by 1912, when he toured the United States, his health was worsening; he had contracted gout and was emotionally disturbed by the overtures of World War I. Even so, the oncoming year promised to be his best.

He returned to Munich to be greeted by his three grandchildren, a welcome deluge of new engine adaptations, a reopening of the Munich Opera, and a salon art exhibition that his wife had arranged in his honour. On September 27, 1913, he set out on a brief journey to London, where he would again be honoured by an engineers' convention. Two evenings later he boarded a ferry to England in company with a longtime engineer friend, with whom he dined jovially. Later that night, his stateroom and luggage still intact, Diesel disappeared at sea and presumably drowned in the English Channel.

BIBLIOGRAPHY. EUGEN DIESEL, *Diesel: Der Mensch, Das Werk, Das Schicksal* (1937), is a sensitive and revealing profile of the inventor-philosopher as recounted by his son. RUDOLF DIESEL, *Solidarismus*, 2 vol. (1903), is the inventor's own attempt to appraise his abstruse social and economic philosophy against a background of what he recognized as a paradoxical industrial society. W.R. NITSKE and C.M. WILSON, *Rudolf Diesel: Pioneer of the Age of Power* (1965), is a detailed, authoritative biography.

(C.M.W.)

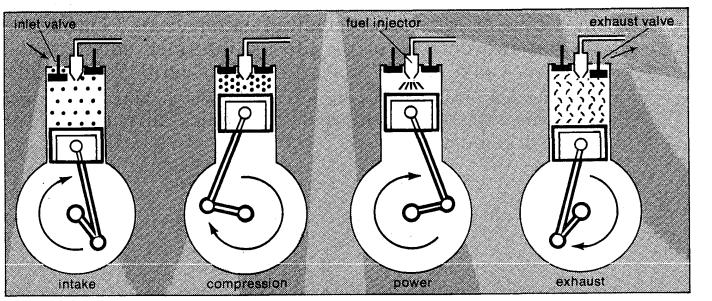
Diesel Engine

The diesel engine is an internal-combustion engine that differs from the older gasoline engine principally in that it relies on heat generated by compressing air in the cylinder to ignite the fuel, rather than on an electric spark. To generate the required heat the diesel must produce higher compression than the gasoline engine, making it bulkier, heavier, and more expensive; it also operates at slower speeds. But it can operate on cheaper, less highly refined fuel, which gives it an advantage in many transportation and construction-equipment applications: locomotives, trucks, tractors, buses, bulldozers, graders and other heavy-duty machines, and in marine propulsion.

HISTORY

Though two English engineers had patented engines that did not depend on spark ignition, Rudolf Diesel of Germany conceived his invention as an improvement on the gasoline engine that fellow-German Nikolaus Otto had developed in 1876. Seeking to increase the efficiency of the Otto engine, it occurred to Diesel that he could do away with electrical ignition if he could compress air to so small a volume that the temperature would be above the ignition point of an appropriate fuel. The cycle of operations he conceived was set forth in his patents of 1892 and 1893 (see illustration): (1) air is drawn into the cylinder as the piston moves away from the cylinder head (intake); (2) the air in the cylinder is compressed by the piston as it moves upward, toward the cylinder head (compression); (3) when the piston reaches the top of its stroke, the fuel charge is injected into the cylinder, where it is ignited by the high temperature of the compressed air. The fuel is injected at such a rate that the maximum cylinder pressure never exceeds the pressure attained by the compression of the air; after completion of the fuel injection, the piston continues to move away from the cylinder head in its downward or expansion stroke (power); (4) the burned fuel is forced from the cylinder by upward motion of the piston (exhaust). This is called four-cycle operation as four separate piston strokes are required; two up and two down.

The diesel cycle



Four-cycle diesel engine showing the sequence of cycle events.

In 1892–93 Diesel took out patents on an engine to operate on the cycle just described. Either powdered coal or liquid petroleum would be used as fuel. Powdered coal was included because of its ready availability as a waste material of the Saar coal mines. Diesel planned to use compressed air to introduce the coal dust into the engine cylinder but found it difficult to control the rate of injection so that the maximum pressure in the cylinder after ignition would not exceed a safe limit. After the

experimental engine was wrecked by an explosion in the cylinder, Diesel gave up the idea of using coal dust and devoted his efforts to the use of liquid petroleum. He continued, however, to use compressed air to introduce the liquid petroleum into the cylinder.

The first commercial engine built on Diesel's patents was installed in St. Louis, Missouri, by a brewer, Adolphus Busch, who had seen one on display at an exposition in Munich and had purchased a license from Diesel for the manufacture and sale of the engine in the U.S. and Canada. The engine operated successfully for many years and was the forerunner of the Busch-Sulzer engine that powered many submarines of the U.S. Navy in World War I. Another Diesel engine used for the same purpose was the Nelseco, built by the New London Ship and Engine Company in Groton, Connecticut.

The diesel engine became the major power plant for submarines during World War I. It was not only economical in the use of fuel but it proved itself reliable under wartime conditions. Diesel fuel, less volatile than gasoline, was more easily stored and handled.

At the end of the war many men who had operated diesels were looking for peacetime jobs. Manufacturers began to adapt diesels for the peacetime economy. One modification was the development of the so-called semidiesel that operated on a two-stroke cycle at a lower compression pressure and made use of a hot bulb or tube to ignite the fuel charge. These changes resulted in an engine less expensive to build and maintain.

Fuel injection developments. One objectionable feature of the full diesel was the necessity of a high-pressure, injection air compressor. Not only did it require energy to drive it but the sudden expansion of the air compressed to 1,000 pounds per square inch when it entered the cylinder in which the pressure was only 500–600 pounds per square inch resulted in a refrigerating effect that delayed ignition. Diesel had needed high-pressure air in order to introduce powdered coal into the cylinder; when liquid petroleum replaced the powdered coal as fuel, a pump could be made to take the place of the high-pressure air compressor.

Common rail method

There were a number of ways in which a pump could be used. In England, the Vickers Company used what was called the common rail method in which a battery of pumps maintained the fuel under pressure in a pipe running the length of the engine with leads to each cylinder. From this rail (or pipe) fuel supply line, a series of injection valves admitted the fuel charge to each cylinder at the right point in its cycle. Another method employed cam-operated jerk, or plunger-type pumps, to deliver fuel under momentarily high pressure to the injection valve of each cylinder at the right time.

The elimination of the injection air compressor was a step in the right direction, but there was yet another problem to be solved: the engine exhaust contained an excessive amount of smoke, even at outputs well within the horsepower rating of the engine, and even though there was enough air in the cylinder to burn the fuel charge without leaving a discoloured exhaust that normally indicated overload. Engineers finally realized that the problem was that the momentarily high-pressure injection air exploding into the engine cylinder had diffused the fuel charge more efficiently than the substitute mechanical fuel nozzles were able to do, with the result that without the air compressor, the fuel had to search out the oxygen atoms to complete the combustion process, and since oxygen makes up only 20 percent of the air, each atom of fuel had only one chance in five of encountering an atom of oxygen. The result was improper burning of the fuel.

The usual design of a fuel injection nozzle introduced the fuel into the cylinder in the form of a cone spray, with the vapour radiating from the nozzle, rather than in a stream or jet. Very little could be done to diffuse the fuel more thoroughly. Improved mixing had to be accomplished by imparting additional motion to the air, most commonly by induction-produced air swirls or a radial movement of the air, called squish, or both, from the outer edge of the piston toward the centre. Various methods have been employed to create this swirl and squish. Best results are apparently obtained when the air swirl bears a definite relation to the fuel injection rate. Efficient utilization of the air within the cylinder demands a rotational velocity that causes the entrapped air to move continuously from one spray to the next during the injection period, without extreme subsidence between cycles.

Swirl or squish

Price's engine. In 1914 a young engineer, William T. Price, began to experiment with an engine that would operate with a lower compression ratio than that of the diesel and at the same time would not require either hot bulbs or tubes. As soon as his experiments began to show promise, he applied for patents.

In Price's engine the selected compression pressure of 200 pounds per square inch did not provide a high enough temperature to ignite the fuel charge when starting. Ignition was accomplished by a fine wire coil in the combustion chamber. Nichrome wire was used for this because it could easily be heated to incandescence when an electric current was passed through it. The experimental engine had a single horizontal cylinder with a bore (cylinder diameter) of 17 inches and a stroke (maximum piston movement) of 19 inches and operated at 257 revolutions per minute. Because the nichrome wire required frequent replacement, the compression pressure was raised to 350 pounds per square inch, which did provide a temperature high enough for ignition when starting. Some of the fuel charge was injected before the end of the compression stroke in an effort to increase the cycle timing and to keep the nichrome wire glowing hot.

OPERATION

Cycle efficiency. The thermal efficiency of any engine cycle may be expressed by the following simple equation:

$$\text{Thermal Efficiency} = \frac{E_1 - E_2}{E_1}$$

in which E_1 represents the heat energy released from the fuel during combustion and E_2 represents the heat energy rejected during the cycle. This equation states that the thermal efficiency of an engine cycle is equal to the difference between the heat supplied by the fuel and the heat rejected during the cycle divided by the heat supplied.

In the meantime many engines of the two-stroke cycle, semidiesel type were being installed. Some were used to produce electricity for small municipalities. Some were installed in water pumping plants. Many provided power for tugs, fishing boats, trawlers, and workboats.

Diesel-electric combinations. In the early 1920s the General Electric Company suggested to the Ingersoll-Rand Company, for whom Price was working, that they cooperate in the building of a diesel-electric locomotive. At that time there were many gasoline engine driven locomotives in service. A diesel-electric locomotive with Price's engine was completed in 1924 and placed in service for switching purposes in New York City.

The success of this locomotive resulted in orders from railroads, factories, and open-pit mines. The engine used in most of these installations was a 6 cylinder, 10-inch bore, 12-inch stroke, rated 300 brake horsepower at 600 revolutions, and weighing 15,000 pounds.

Marine applications. Many diesel engines were purchased for marine propulsion. The diesels, however, normally rotated faster than was desirable for large ships' propellers because the high speeds of the huge propellers tended to create hollowed-out areas within the water around the propeller (cavitation), with resultant loss of thrust. The problem did not exist, however, with smaller propellers, and diesel engines proved especially suitable for yachts, in which speed is desired. The problem was solved by utilizing a diesel-electric installation in which the engines were connected to direct-current generators that furnished the electricity to drive an electric motor connected to the ship's propeller.

There were also many installations in which the diesel

was connected either directly or through gears, to the propeller. An unusual installation was a diesel-driven ferry that plied on the Hudson River between the towns of Hudson and Athens. In this installation the engine crankshaft was extended at both ends and coupled to shafting that carried a propeller at both ends of the ferry boat.

When diesel engines of larger horsepower and slower rotation speeds became available, they were installed in cargo and passenger carrying ships. Diesel engines were also installed in hospitals, telephone exchanges, airports, and other places to provide emergency power if regular electrical power was interrupted.

Emergency power

Dual fuel engines. The dual fuel engine is an internal combustion engine that can be operated as an oil diesel entirely on liquid petroleum or, without any mechanical alterations, can produce its horsepower rating on a combination of liquid fuel and natural, or artificially produced, gas. Liquid fuel is needed to perform the basic function of igniting the gas.

The gas diesel operates on gas as the primary fuel that is ignited by the introduction of a small pilot charge of liquid fuel into the cylinder shortly before the piston reaches top dead centre on its compression stroke. The amount of liquid fuel used in the gas diesel is never more than is needed for the ignition of the gas.

On April 30, 1901, the U.S. Patent Office issued to Rudolf Diesel patents that covered all the essential features of the dual fuel and gas diesel engines. In 1932 patents were issued to Victor Heidelburg that covered the advantages of the higher fuel economy made possible by the higher compression ratios that could be used.

Both patents lay unused until 1940, when the rising price of liquid fuel created a demand for natural gas that could be purchased at a very low rate on the basis of an interruptible contract. With such a contract the engine owner could switch from gas to oil and back again on short notice.

Diesel engine starting. A diesel engine is started by driving it from some external power source until conditions have been established under which the engine can run by its own power. The most positive starting method is by admitting air at 250 to 350 pounds per square inch to each of the cylinders in turn on their normal firing stroke. The compressed air becomes heated sufficiently to ignite the fuel. Other starting methods involve auxiliary equipment and include admitting blasts of compressed air to an air-activated motor geared to rotate the large engine's flywheel; supplying electric current to an electric starting motor, similarly geared to the engine flywheel; or by means of a small gasoline engine geared to the engine flywheel. The selection of the most suitable starting method depends upon the physical size of the engine to be started, the nature of the connected load, and whether or not the load can be disconnected during starting.

Fuel for diesels. Petroleum products normally used as fuel for diesel engines are distillates composed of heavy hydrocarbons, with at least 12 to 16 carbon atoms per molecule. These heavier distillates are taken from crude oil after the more volatile portions used in gasoline are removed. The boiling points of these heavier distillates range from 350° to 650° F. Thus, their evaporation temperature is much higher than that of gasoline that has fewer carbon atoms per molecule. Specifications for diesel fuels published in 1970 listed three grades: the first was a volatile distillate recommended for high-speed engines with frequent and wide variations in load and speed; the second, a distillate for high-speed engines in services with high loads and uniform speeds; and the third, a fuel for low- and medium-speed engines in services with sustained loads.

Water and sediment

Water and sediment in fuels can be harmful to engine operation; clean fuel is essential to efficient injection systems. Fuels with a high carbon residue can be handled best by engines of low-speed rotation. The same applies to those with high ash and sulfur content. The cetane number, which defines the ignition quality of a fuel, is ascertained by adjusting a mixture of cetane and *alpha*-methyl-napthalene until it has the same ignition quality as the fuel being tested. The percentage of cetane in this mixture is then the cetane number of the fuel under test. For the first two grades of diesel fuel described above, the minimum cetane number is 40; for the third grade, the minimum is 30, representing 30 percent cetane in the fuel.

Supercharging. The horsepower output of a diesel engine can be increased by compressing the air charge prior to its admission to the cylinder, thus increasing the weight of air available for combustion. This supercharging can be used either to increase horsepower output at sea level or to recover horsepower lost at higher altitudes due to thinner air. Supercharging can be accomplished in a number of ways: (1) by an engine-driven blower, (2) by an exhaust gas-driven compressor, or (3) by a tuned manifold that produces a phenomenon known as ramming.

In ramming, the cylinders of a multiple-pipe system are selectively connected to two or more exhaust pipes in such a way (tuned manifold) that no cylinder will be exhausted into the pipe at the same instant that another is being cleared of exhaust gases by forced fresh air. The advantage of such a precision grouping is the retention and utilization of the kinetic energy remaining within these exhaust gases and at a high level, for use as power generation when channelled to the turbine. Excessive back pressure is also prevented against the cylinder being scavenged.

THE DIESEL ENGINE TODAY

Since its success in World War II, in which it was the most prevalent power plant for armed forces equipment on the ground and at sea, the diesel engine became the conventional power source for most railroad locomotives, most heavy construction machinery, most high-powered farm tractors, and a large proportion of trucks and buses. Its weight makes it unsuitable for aircraft use, and it has had only limited (but growing) application in passenger automobiles. In general, it finds applications wherever its greater weight and cost and less smooth-running operation are offset by its lower cost of operation. An especially noteworthy application is in marine propulsion. While passenger-carrying vessels favour turbines that provide more speed and less vibration, cargo vessels, especially the largest sizes, profit from low-cost diesel power. Of 371 ships built in two recent years for the merchant fleet of the U.S.S.R., for example, 7 were propelled by steam turbines, 1 by a gas turbine, and 363 by diesel plants.

BIBLIOGRAPHY. L.V. ARMSTRONG and J.B. HARTMAN, *The Diesel Engine* (1959), a text covering the theory, basic design, and economics of the diesel engine; C.B. DICKSEE, *The High-Speed, Compression-Ignition Engine* (1940), a text covering the principles of operation and problems of compression-ignition engines running at rotative speeds required for road transport service; H.F.P. PURDAY, *Diesel Engine Design*, 5th ed. (1948), a text dealing with the functions and details of the design of engines, both large and small, and design problems such as heat flow, lubrication, and vibration stresses; A.R. ROGOWSKI, *Elements of Internal-Combustion Engines* (1953), a detailed discussion of the three stages of combustion as applied to compression-ignition engines; M.L. SMITH and K.W. STINGSON, *Fuels and Combustion* (1952), a discussion of fuel characteristics, availability, and combustion; A.J. BUCHI, *Exhaust Turbocharging of Internal Combustion Engines* (1953), a monograph on pressure fluctutation of intake air and exhaust gases, and also multiple pipe turbocharging as outlined in Dr. Buchi's patents. Other texts include: A.W. JUDGE, *Modern Smaller Diesel Engines in Theory, Construction, Operation, and Maintenance* (1969); C.C. POUNDER (ed.), *Diesel Engine Principles and Practice*, 2nd ed. (1962); and M.H. HOWARTH, *The Design of High Speed Diesel Engines* (1966).

(L.V.A.)

Dietary Laws and Food Customs

Like all other biologically and physically necessary things and acts, food and eating are always surrounded by social regulations that prescribe what may or may not be

ingested under particular social conditions. These prescriptions and proscriptions are sometimes religious; often they are secular; frequently, they are both. This article surveys the variety of laws and customs pertaining to food materials and the art of eating in human societies from earliest times to the present. It will be seen that behaviour in respect to food—whether religious, secular, or both—is institutionalized behaviour and is not separate or apart from organizations of social relations.

By an institution is meant here a stable grouping of persons whose activities are designed to meet specific challenges or problems, whose behaviour is governed by implicit or explicit rules and expectations of each other and who regularly use special paraphernalia and symbols in these activities. Social institutions are the frames within which man spends every living moment. This survey explores the institutional contexts in which dietary laws and food customs are cast in different societies; the attempt will also be made to show that customs surrounding food are among the principal means by which human groups maintain their distinctiveness and help provide their members with a sense of identity.

Theoretical approaches to food customs

Other points of view about food customs cover a wide range. What may be labelled an ecological approach suggests that food taboos among a group's members prevent over-utilization of particular foods to maintain a stable equilibrium in the habitat. Recently, investigators of such customs have been exploring the hypothesis that they provide an adaptive distribution of protein and other nutrients so that these may be evenly distributed in a group over a long period instead of being consumed at one time of the year. The ecological approach also suggests that many food taboos are directed against women to maintain a low population level; this seems to be an adaptive necessity in groups at the lowest technological levels, in which there is a precarious balance between population and available resources.

There are also psychological approaches to food customs. Psychoanalytic writers speculate that food symbolizes sexuality or identity because it is the first mode of contact between an infant and its mother. This point of view is most clearly exemplified in ideas that attitudes toward food, established early in life, tend to shape attitudes toward money and other forms of wealth and retentiveness or generosity. According to Claude Lévi-Strauss, a French anthropologist, the categories represented in food taboos enable people to order their perceptions of the world in accordance with the principle of polarities that govern the structure of the mind. Thus, they aid in maintaining such dichotomies as those between nature and culture or between man and animal.

NATURE AND SIGNIFICANCE IN WORLD CULTURES

There are no universal food customs or dietary laws. Nor are food customs and dietary laws confined to either preliterate ("primitive") or advanced cultures; such regulations are found at all stages of development. Nevertheless, different types of regulations in respect to food are characteristic of groups at different levels of cultural or socio-technological development.

Each society has attached symbolic value to different foods. These symbolizations define what may or may not be eaten and what is desirable to eat at different times and in different places. In most cases, such cultural values bear little relationship to nutritive factors. As a result, they often seem difficult to explain. Moreover, dietary customs and laws are resistant to rational argument and change. For example, experts from health and nutritional agencies find it difficult to persuade mothers to give cow's milk to children in societies in which it is looked upon as undesirable. Such customs and laws also prevent people from adopting alternative foods during periods of shortage. During and after World War II, some Indians refused to eat Western wheat and rioted and died rather than accept it.

Food as a material expression of social relationships. Cutting across dietary laws and customs is the more general association of food and drink with those social interactions that are considered important by the group. In many societies the phrase "We eat together" is used by a man to describe his friendly relationship with another from a distant village, suggesting that even though they are not neighbours or kinsmen they trust one another and refrain from practicing sorcery against each other. Among the Nyakyusa of Tanzania, "for conversation to flow merrily and discussion to be profound, there must be . . . 'the wherewithal for good fellowship,' that is, food and drink—and very great stress is laid on sharing these." In Old Testament times, almost every pact, or covenant, was sealed with a common meal; eating together made the parties as though members of the same family or clan. Conversely, refusal to eat with someone was a mark of anger and a symbol of ruptured fellowship. Eating salt with one's companions meant that one was bound to them in loyalty; references to this are found in the New Testament.

The common meal

Such sentiments, however, are not confined to tribal or ancient cultures. In Israeli kibbutzim (communal settlements), the communal dining room is a keystone institution, and commensality is one of the hallmarks of kibbutz life. The decline of communal eating and the increasing frequency of refrigerators, cooking paraphernalia, and private dining in kibbutz homes is regarded by some observers as a sign of the imminent demise of kibbutzim. In many U.S. communes there is a single facility for cooking and dining. Dinners must be taken communally; private dining is taken as a signal that one is ready to leave the group.

The provision of food and drink, if not actual feasting, is characteristic of rites of passage—*i.e.*, rites marking events such as birth, initiation ceremonies, marriage, and death—in almost all traditional cultures and in some modern nontraditional groups as well. The reason for this is that these events are regarded as being of importance not only to the individual and his family but also to the group as a whole because each event bears in one way or other on the group's continuity.

Furthermore, food and drink are almost universally associated with hospitality. In most cultures, there are explicit or implicit rules that food or drink be offered to guests, and there are usually standards prescribing which foods and drinks are appropriate. Reciprocally, these sets of rules also assert that guests are obligated to accept proffered food and drink and that failure to do so is insulting. In many societies, there are prescribed ritual exchanges of food when friends meet. Food is thus one of the most widespread material expressions of social relationships in human society.

Regulations about the quantity of food and drink consumed. It is extraordinarily rare for cultures to condone gluttony, the conventional exaggerations of the eating behaviour of the ancient Roman elite notwithstanding. Most people cannot afford to be gluttons. There are more examples of the other extreme, asceticism, though these too are infrequent.

A clear-cut example of gastronomic asceticism is provided by Indians of the U.S. Northeast, such as the Micmac, Montagnais, and Ojibwa. It was an ideal among them to eat sparingly. Preparation for this attitude began in early childhood with short fasts of a day or two, culminating in the puberty fast; the latter lasted about ten days, during which time the child was isolated in a tiny wigwam without food or water. The puberty fast also had important religious significance. During the fast, the child had to supplicate the deities for a vision (easily induced under such conditions), which came in the form of a supernatural figure, usually in animal shape; this was to become his guardian spirit.

Rules pertaining to drink are even more varied. Tribal groups throughout the world (except in Oceania and most of North America) knew alcohol; in each case, this led to the adoption of rules concerning its use.

Alcoholic beverages

Although a high intake of alcohol always has physiological effects, people's comportment is determined more by what their society tells them is the way to behave when consuming alcohol than by its toxic effects. In many societies, drinking is an established part of the total round of social activities. Robert McC. Netting, a U.S. anthro-

pologist, observed that the Kofyar of northern Nigeria "make, drink, talk, and think about beer." All social relations among them are accompanied by its consumption, and fines are levied in beer payments. Ostracism takes the form of exclusion from beer drinking; they "certainly believe that man's way to god is with beer in hand." Their beer, however, is weak in alcoholic content and is quite nutritious, and they rarely consume European beer and never distilled liquor. Among Central and South American peasants, men are allowed or required to drink themselves into a state of stupefaction during religious celebrations (fiestas); though this drinking is frequent and heavy, it does not appear to result in addiction. Representative of the other extreme are the Hopi and other Indian tribes of the U.S. Southwest who have banned all alcoholic beverages (and almost all narcotics), asserting that these substances threaten their way of life.

Most cultures, however, prescribe moderation in drinking. In ancient Mesopotamia, beer played an important role in temple services and in the economy; but the code of Hammurabi—the monument of law named after the king of Babylon—strictly regulated tavern keepers and servants (these places were supposed to be avoided by the social elite). Similar patterns obtained in ancient Egypt. The ancient Greeks sought to attribute their intellectual and material culture to the introduction of vine and olive growing. The use of wine was quite general in biblical times; it belonged to the list of indispensable provisions listed in the Old Testament in the Book of Judges (chapter 13) and the First Book of Samuel (chapters 16 and 25). Wine was no less important in New Testament times; in Revelation to John (chapter 6) it is said that only wine and oil are to be protected from the apocalyptic famine. Wine is also frequently used in biblical imagery. In both Testaments, however, wine is both praised and condemned.

Use of food in religion. The most widespread symbolic use of food is in connection with religious behaviour. In fact, eating and drinking are minimal elements in most religious behaviour and experience, whether in eating, sacrifice, or communion. According to many anthropologists, there are essentially two reasons for this. First, religion is one of the systems of thought and action by which the members of a group express their cohesiveness and identity. Implicitly or explicitly, the members of every cultural group assert that its unity and distinctiveness derive from the deity or deities associated with it. Religion is a tie that binds. But no symbolic activity in human society stands alone and without material representation. Like all other symbolizations of institutional relationships, those of religion must also have substantial form. Food and drink—and their ingestion—are among the most important substances of religion.

Reasons for symbolic use of food in religious behaviour

The second reason, closely related to the foregoing, is that one element of dogma in every religion is the definition of polluting, or supernaturally dangerous, objects or personal states. Just as there is no objective or scientific connection between the nutritive qualities of different foods and the symbolic values attached to them, there is no objective relationship between an object or a personal state and its definition as polluting. Cultures vary in the objects and states that are defined as defiling, such as saliva, sneezing, menstruation, killing an enemy in warfare, a corpse, parturition, but cutting across these is the belief held in every religion that there are foods and drinks that are polluting or defiling.

As Mary Douglas, a British anthropologist, has suggested in *Purity and Danger*, her analysis of the religiously sanctioned food taboos in Leviticus (chapter 11) and Deuteronomy (chapter 14), concepts of pollution and defilement are among the means used by preliterate or tribal societies to maintain their separateness, boundedness, and exclusivity; thus, these concepts and rules contribute strongly to the sense of identity—the social badges—that people derive from participation in the institutions of their firmly bounded or encapsulated groups. More concretely, when a person proclaims his affiliation with and allegiance to a particular group that he regards as his self-contained universe and beyond whose margins he sees danger, threat, and alienation, he simultaneously invokes—explicitly or implicitly—the many badges of his social identity; these include the totem (*i.e.*, the emblem of a family or clan) that he may not eat, the foods that are regarded as defiling, the drinks that he must avoid, the sacred meals in which he participates, and the other rituals associated with his exclusive group. He thereby asserts his separateness from people in all other groups—usually referred to in pejorative terms—and his identification with the members of his own group. Food customs are not always formalized, however; they are sometimes cast in terms of preference. Americans, for example, unless they are members of ethnic or religious groups that have their own dietary laws, often shun the "exotic" foods of alien cultures; but these avoidances are not phrased in religious or other institutional terms.

LAWS AND CUSTOMS AT DIFFERENT STAGES OF SOCIAL DEVELOPMENT

Although there are dietary laws and customs in all societies, groups differ in this regard in two important ways: in the range or extent of foods that are defined as polluting or tabooed and in conceptualizations of the consequences resulting from violations of these laws and customs. In comparing societies, however, it must be remembered that the range of variability among them is so great that it would be necessary to list hundreds of societies and their customs to get a complete and detailed picture of their food customs and laws. For purposes of both economy and conceptual coherence it is necessary to group societies into levels, or stages, of social and technological development and to compare these; in this approach, individual societies are regarded as special or particular exemplary cases of the general class of the level of development in which the groups are found or classed.

Hunter-gatherers. The earliest cultural level that anthropologists know about is generally referred to as hunting-gathering. Hunter-gatherers are always nomadic, and they live in a variety of environments. Some, as in sub-Saharan Africa and India, are beneficent environments; others, such as those of the Arctic or North American deserts, are harsh and dangerous. Encampments of hunter-gatherers are usually small (generally fewer than 60 persons) and are constantly splitting up and recombining. An important rule among almost all hunter-gatherers is that every person physically present in a camp is automatically entitled to an equal share of meat brought into the group whether or not he has participated in the hunt; this rule does not usually extend to vegetables or fruits and nuts.

It may be thought that hunter-gatherers who live in habitats of scarcity and in which hunting is dangerous would try to make maximum use of all potentially available food; they are, however, also characterized by customs and beliefs that proscribe certain foods or at least limit their consumption. Many Alaskan Eskimo groups, for instance, make a sharp distinction between land and sea products; the Eskimo believe that products of the two spheres should be kept separate, maintaining that land and sea animals are repulsive to each other and should not be brought together. Thus, for example, before hunting caribou (a spring activity), a man must clean his body of all the seal grease that has accumulated during the winter; similarly, before whaling in April, the individual's body must be washed to get rid of the scent of caribou. Weapons used for hunting caribou should not be used at sea; implements used at sea, however, may be used to hunt caribou. If these rules are violated, the hunter or whaler will be unsuccessful in his food quest; the consequences of this, of course, can be dire.

Alaskan Eskimo groups

In addition, the Eskimo observe food taboos in connection with critical periods of the individual's life and development. Among the most outstanding of these are the food taboos that a woman is subject to for four or five days after giving birth. She may not eat raw meat or blood and is restricted to those foods that are believed to have beneficial effects on the child. For example, it is felt that she should eat ducks' wings to make her child a good runner or paddler. Because the Eskimo are often beset by

food shortages, they sometimes have to eat forbidden foods. In such cases, there are several things that a person can do to neutralize the taboo. He first rubs the forbidden food over his body and then hangs the meat outside and allows it to drain. Another act that is regarded as particularly efficacious is to stuff a mitten into the collar of his parka with the hand side facing outward; it is believed that the harmful effects of the taboo food go into the mitten and travel away from him.

There are, of course, other food avoidances observed by the Eskimo, but these examples will suffice to illustrate the basic principles of dietary customs and laws among hunter-gatherers. First, the taboos are always thought to have magical consequences for the individual; observing them will assure health and strength, violating them will result in illness and weakness for the person or, in the case of a parturient mother, for her child. Second, food taboos are generally associated with critical periods during the life cycle, as in pregnancy, menses, illness, or dangerous hunts. Third—and this is true of almost all societies, not only those of hunter-gatherers—in every group's system of thought there are categories or types of foods that are regarded as dangerous, defiling, or undesirable. At first glance, these rules and customs seem arbitrary and capricious, but evidence is accumulating that there are rational elements in them. Although it would be difficult in the present stage of knowledge to apply this principle to every dietary taboo or custom in every society, it seems that prohibitions are placed on those foods that are the most difficult and dangerous to procure. Sometimes, however, these foods are also highly prized.

Corporate kin groups. With the development of corporate kin groups in social history, largely (but not exclusively) as an accompaniment of horticultural cultivation, a significant change occurred in the role of food in institutional life. Underlying the development of corporate kin groups was the development of the notion of exclusive rights to territory claimed by a group of kinsmen. This exclusive territoriality was probably designed, in large measure, to protect investments of time and effort in particular plots. The solidarity and sense of kin-group exclusiveness implicit in a corporate kin group grew out of kin-group ownership of the land and the individual's reliance on interhousehold cooperation in his productive activities. Such groups quickly evolve insignia, rules, and symbols that represent their ideals of exclusivity and inalienability of social relations; food plays an important role in this. Hence, taboos are thought to have consequences for the group as a whole rather than for the individual.

Territorial rights and initiatory rites

Another significant accompaniment of the development of corporate kin groups is the elaboration of initiatory rites, which mark an individual's transition from childhood to full membership in his community or kin group; they confer citizenship in the fullest sense of the term. Such events are celebrated by feasts, reciprocal exchanges of food, and food taboos, in addition to the ceremonial rituals themselves. Preparations for these feasts sometimes occupy the group for several months, especially when it is necessary to acquire from relatives and friends the animals that will be slaughtered and eaten, because it is rare for one family, or even one village, to own enough animals for a proper feast. They lay the groundwork for one of the basic rules of the group into which the individual is being initiated, namely, that the distribution of food (and interhousehold cooperation in its acquisition) is one of the most significant ways in which he and the members of the group are knit together.

Feasting is also an integral element of religious assemblages and ritual in these societies, as are offerings to deities, whether spirits or ancestors. Because one of the main purposes of religious activity is to symbolize the solidarity of the group, food is used as a material representation of this cohesiveness. Additionally, it is believed in almost all tribal societies, whether or not they are characterized by corporate groupings, that all plant and animal foodstuffs are made available to man through the beneficence of the gods. Man's relationship with the deities in tribal societies is always, in part, an economic one involving the deities' provision of food. A gift from the gods must be balanced by a reciprocal gift to them from their adherents. In prayer, men thank their deities for these gifts; in sacrifice and offerings, they offer gifts to their deities.

Chiefdoms. The next major social and political developments in human history are the appearance of institutions in which political and economic power is exercised by a single person (or group) over many communities. Often referred to as chiefdoms by anthropologists, this development signalled a process evident today throughout the world, namely, the steady growth of centralized power and authority at the expense of local and autonomous groupings.

Political authority in chiefdoms is inseparable from economic power, including the right by rulers to exact tribute and taxation. One of the principal economic activities of the heads of chiefdoms is to stimulate the production of economic surpluses, which they then redistribute among their subjects on different types of occasions, as during feasts in the celebration of religious ceremonies, rites of passage of members of chiefly families, and periods of famine. The accumulation of these surpluses requires conservation policies. Because techniques of food preservation were poorly developed in preliterate chiefdoms, the heads of chiefdoms often adopted the policy of placing taboos—often phrased in religious terms—on different crops or resource areas, forbidding their consumption until the prohibitions were lifted. These taboos, however, were not exclusively for conservation purposes; they were also occasionally designed to underwrite higher standards of living for the chiefs themselves. For instance, in some Polynesian societies, as in Samoa, fishermen were required to obey a taboo that a portion of their catch must be given to the chief. The penalties for violating such taboos were supernaturally produced illness or other misfortunes.

Economic powers of the chief

Complex societies. As societies became increasingly complex, heterogeneous, and divided along lines of caste, class, and ethnic affiliation, their dietary customs became correspondingly less uniform because they mirrored these divisions and inequalities. Although these distinctive customs are almost always placed in the context of religious belief and practice, according to many anthropologists, the dietary observances in everyday behaviour are primarily shaped by economic and social considerations; moreover, observances at the village level rarely correspond directly to formal prescriptions and proscriptions.

The dietary laws and customs of complex nations and of the world's major religions, which developed as institutional parts of complex nations, are always based on the prior assumption of social stratification, traditional privilege, and social, familial, and moral lines that cannot be crossed. Taboos and other regulations in connection with food are incompatible with the idea of an open society. Nevertheless, complex nations were characterized by caste organizations that, in almost all cases, religion helped to legitimate. Caste systems, in addition to their other characteristics, are supported by deeply felt fears of pollution or contamination as a result of unguarded contact of the more "pure" with those who are less "pure."

Social stratification as basis for food customs

Although there is no doubt that the development of caste is linked to some form of occupational separation in a society, which, in turn, leads to the development of ideas concerning the separation of unclean persons from the ordinary or of the ordinary from the superpure, there is considerable controversy over the origins of caste systems. Regardless of the origins, however, the separation of castes is always mirrored in rules for eating that, when breached, represent a threat to the social order and to the individual's sense of identity. There is also a question among scholars whether or not caste is unique to India. Nevertheless, in Japan as well as India, eating together implies social and ritual equality, as it does in the United States, where, unlike Japan and India, food-related caste behaviour has not been institutionalized in religion. In

India and Japan, a person who cooks for another and serves his food must be equal or superior in rank to the recipient of the food; only in this way can the latter avoid pollution. In the caste system of the United States, a Negro may cook and serve but not eat with the whites (though this is changing in many sectors). Before the U.S. civil-rights movement, violation of these eating taboos constituted defiance of caste; observance of the etiquette was evidence of the acceptance of caste. It seems that dietary rules of caste relationships were never incorporated into U.S. religion because the society as a whole was governed by an ideology—though not a reality—of free and open social mobility.

RULES AND CUSTOMS IN WORLD RELIGIONS

Judaism. Perhaps the best known illustration of the idea that the dietary laws and customs of a complex nation and its religion are based on the prior assumption of social stratification or, at least, of a sense of separateness, is provided by Judaism as spelled out in the Mosaic Law in the Old Testament books of Leviticus (chapter 11) and Deuteronomy (chapter 14). Prohibited foods may not be consumed in any form: all animals—and their products—that do not chew the cud and do not have cloven hoofs (*e.g.*, pig and horse); fish without fins and scales; blood; seafood (*e.g.*, clams, oysters, shrimp, crabs) and all other living creatures that creep; and those fowl enumerated in the Bible. All foods outside these categories may be eaten.

Prohibited foods

Interpretation of Jewish laws. Mary Douglas has offered probably the most cogent and widely accepted interpretation of these laws in her book *Purity and Danger*. She suggests that these notions of defilement are rules of separation; they symbolize and help maintain the biblical notion of the separateness of the Hebrews from other societies. A central element in her interpretation is that each of the injunctions is prefaced by the command to be holy and that it is the distinction between holiness and abomination that enables these restrictions to make sense. "Holiness means keeping distinct the categories of creation. It therefore involves correct definition, discrimination, and order." The Mosaic dietary laws exemplify holiness in this sense. The ancient Hebrews were pastoralists, and cloven-hoofed and cud-chewing hoofed animals are proper food for such people; hence, Douglas maintains, they became part of the social order and were domesticated as slaves. Pigs and camels do not meet the criteria of animals that are fit for pastoralists. As a result, they are excluded from the realm of propriety. Douglas notes that there is remarkable consistency in Mosaic dietary laws. The Bible "allots to each element its proper kind of animal life. In the firmament two-legged fowls fly with wings. In the water scaly fish swim with fins. On the earth four-legged animals hop, jump, or walk. Any class of creatures which is not equipped for the right kind of locomotion in its element is contrary to holiness." People who eat food that is "out of place," as it were, such as four-footed creatures that fly, are themselves unclean and are prohibited from approaching the Temple.

There is, however, another dimension to Old Testament food customs. In addition to expressing their separateness as a nation—membership in which was ascribed by birthright—Israelite food customs also mirrored their internal divisions, which were castelike and were inherited. Though the rules of separation referred primarily to the priests, they also affected the rest of the population. The priest's inherent separateness from ordinary men was symbolized by the prescription that he must avoid uncleanness more than anyone else. He must not drink wine or strong drink, and he must wash his hands and feet before the Temple service. Explicit in Old Testament prescriptions is that an offering sanctifies anyone who touches it; therefore, often the priests alone were permitted to consume it.

These rules symbolizing the priestly group's castelike separateness also validated a system of taxation benefitting them, couched in terms of offerings, sacrifice, first-fruit ceremonies, and tithes. The religious rationalization of taxation is illustrated in the Old Testament by the first-fruits ceremony. Fruit trees were said to live their own life, and they were to remain untrimmed for three years after they were planted. But their fruits could not be enjoyed immediately: God must be given his share in the first-fruit ceremonies. These first fruits represent the whole, and the entire power of the harvest—which is God's—is concentrated in them. Sacrifice is centred around the idea of the first-fruits offering. Its rationalization was that everything belonged to God; the central point in the sacrifice is the sanctification of the offering, surrendering it to God. Its most immediate purpose was to serve as a form of taxation to the priests; only they were considered holy enough to take possession of it.

Elaboration of the Jewish laws. After the exile of the Jews from Palestine following the conquest by Rome in the 1st century AD, a remarkable elaboration in their dietary laws occurred, probably as a result of the Jews' attempts to maintain their separateness from nations into whose midst they were thrust. Many customs evolved that have taken on the force of law for those Jews who have sought to maintain a traditional way of life. For example, the Bible does not prescribe ritual slaughter of animals, yet this practice has taken on the same compulsion as the taboo on pigs and camels; a permitted food (*e.g.*, cattle, chicken) that has not been ritually slaughtered is now regarded to be as defiling as pork. Similarly, one of the hallmarks of the Passover holiday in Judaism is the eschewal of all foods containing leaven, the consumption only of foods that have been designated as "kosher for Passover," and the use of special sets of utensils that have not been used during the rest of the year. But these, too, are postbiblical customs that have been given the force of law; the Bible prescribes nothing more than eating unleavened bread during the Passover season.

Shtetl communities

Further elaborations on the Mosaic Law in regard to food can be observed in the dietary customs of certain groups of modern Jews in their daily lives. In the pre-World War II eastern European Jewish community (or *shtetl*), behaviour in regard to food not only included the biblical prescriptions and proscriptions but, in many ways, resembled the behaviour of people in the corporate communities of tribal societies. The major life crises were celebrated by feasts or other uses of food. Wine and other foods were integral parts of circumcision ceremonies and of a boys' attainment of ritual majority (Bar Mitzwa). Weddings were also celebrated with huge feasts that required weeks, if not months, of preparation, and guests were seated at the wedding feast according to their social rank. Following the wedding celebration, grain was sprinkled on the couple's heads, apparently to promote fertility. Those who visited mourners were to eat hard-boiled eggs or other circular food because roundness symbolizes mourning.

Aside from the daily requirements of following the Mosaic dietary laws, which apply to everyone, the heaviest burden for maintaining these observances falls on the women; their ritual and secular statuses are always inferior to the men's. It is the task of the housewife to be sure that meat and dairy foods are not mixed, that ritually slaughtered meat is not blemished, and that cooking equipment and dishes and utensils for meat and dairy are rigidly separated. The only personal states of ritual pollution relating to food in *shtetl* culture also refer only to women. For instance, a woman who has not been ritually cleansed after her menses must not make or touch pickles, wine, or beet soup. If she violates this customary rule, it is believed that these foods will spoil.

"Ḥasidic" Jews

A further illustration of the idea that dietary rules and customs are inextricably associated with the maintenance of group separateness is provided by one sect of Jews in the United States, those who refer to themselves as Ḥasidim (Pious Ones). These people live in self-contained enclaves; most of them are immigrants from the *shtetl*. In addition to preserving their distinctiveness from surrounding non-Jewish communities, they are equally devoted to preserving their distinctiveness vis-à-vis other Jews; no matter what their degrees of piety, the latter are regarded by Ḥasidim as nonreligious.

This is clearly reflected in their behaviour in regard to

food. The Ḥasidim assert that the larger Jewish community (and its rabbis) do not meet Ḥasidic standards and qualifications in the manufacture, preparation, handling, and sale of food; even non-Ḥasidic ritual slaughterers are classed with assimilated Jews who do not observe dietary laws at all. Hence, their food products are regarded as forbidden, and Ḥasidim consider only their own products as permissible for consumption. Even neutral foods, such as vegetables, are defined as nonkosher if handled by a non-Ḥasid since there is always the suspicion that it may come into contact with nonkosher—and thus contaminating—matter. Thus, for instance, only milk that they designate as "Jewish" can be used; only noodles prepared by someone from the Ḥasidic community may be consumed because there is the suspicion that eggs with a drop of blood (which are forbidden) may have been used in the noodles' preparation; only approved sugar may be used; and even paper bags that hold food come under these restrictions because only a member of the community is above the suspicion that forbidden matter has been included in the glue that is used in manufacturing the bags.

The extremity of Ḥasidic strictures with regard to food has to be viewed in the context of their setting in the United States and not only in the light of their Jewish sources. The Ḥasidim regard the growing secularization of U.S. life as the greatest threat to the perpetuation of the ancient tradition of Judaism; their extremism is the wall they have erected to stave off this danger of threatened assimilation.

Black Muslim movement. Until relatively recently, the separatism of U.S. Negroes was underwritten by an intricate combination of law and custom. The attempt of the United States government to achieve an integration of blacks and whites in daily social, economic, and political life was viewed by some Negroes as a threat to their social identity. Ideologies designed to legitimate the maintenance of their social identity began to develop, especially after the desegregation decision of the Supreme Court in 1954, the most notable of which is known as the Nation of Islam (the Black Muslims). In their attempt to separate themselves from the larger aggregate of U.S. Negroes, as well as from the rest of U.S. society, the Black Muslims sought to develop a separate social identity by adopting a set of symbols to which they attached particular meanings. A person's membership in the group not only depended on assuming a Muslim name but also on eating certain foods and avoiding others, including alcohol and tobacco. Forbidden foods include meats and fish proscribed by the Bible and Qur'ān and also more than a dozen vegetables that were staples in the slave diet.

Islām. Islāmic dietary laws—as spelled out in the Qur'ān—also illustrate their relationship to the establishment of a sense of social identity and separateness. Muḥammad, the founder of Islām, was among other things a political leader who welded a nation out of the mutually warring tribes of Arabia. His religious ideology legitimated the unification of these autonomous tribes and his own paramount rule over them. The main religious tenets of Islām were derived from Judaism and early Christianity, and it is clear from the Qur'ān that Islām was intended to encompass all aspects of life.

Relationship with Judaism and Christianity

Muḥammad apparently knew more about Judaism than about Christianity, and many of his strictures in the Qur'ān were explicit in establishing distinctions between Arabs and Jews. This is evident in his dietary regulations, which borrow heavily from Mosaic Law. Specifically, Muḥammad proscribed for Muslims the flesh of animals that are found dead, blood, swine's flesh, and food that had been offered or sacrificed to idols. The most radical departure of Qur'ānic from Mosaic dietary laws was in connection with intoxicating beverages. Though Jews frown upon alcoholic beverages, they do not forbid them, and wine is an important element in many rituals and feasts; Muḥammad, however, absolutely forbade any such beverages.

Specific departures from Mosaic and Christian dietary rules notwithstanding, Islām represents a more fundamental removal from all other major religions: what is polluting, forbidden, and enjoined for one person in Islām applies equally to all. Islām's sharpest contrast in this regard is to the religions of India. This difference is highlighted by the fact that Muslims of all social statuses in an Indian village eat freely with each other, worship in the same mosques, and participate in ceremonies together.

Christianity. Christianity did not develop elaborate dietary rules and customs. This probably grew out of the controversy between the Judaizing and Hellenizing branches of the church during the earliest years of Christianity over whether or not to observe Mosaic food laws. The Council of Jerusalem settled on the formula that meat offered to idols, blood, and things strangled must be abstained from, thus freeing the Gentiles in all other respects from the law. The apostle Paul's position on the matter, however, was that "nothing is unclean in itself"; and it was thus that the New Testament repudiated the entire body of laws of purity, especially those pertaining to food. Jesus is said to have declared that defilement could not be caused by any external agent. The apostle Peter's vision of the sheet lowered from heaven and containing all types of animals that the divine voice pronounced clean and fit for food provided the church with a mandate to abandon the Old Testament food laws.

Food, however, in terms of the Last Supper and the Eucharist, plays an important role in Christianity. As told by the early Christians, Jesus foresaw his death and performed a simple ceremony during a last meal to bring home the significance of his death to the Twelve: he broke a loaf into pieces and gave it to them saying, "Take this, it is my body." After they had eaten, he took the cup of wine and said, "This is my blood."

Last Supper and Eucharist

During the 1st century AD, Christian communities developed into self-contained units with an organized life of their own. When they were beginning to see themselves as a church, they held two separate kinds of services: (1) meetings on the model of the synagogue that were open to inquirers and believers and consisted of readings from the Jewish scriptures and (2) *agapē*, or "love feasts," for believers only. The latter was an evening meal in which the participants shared and during which a brief ceremony, recalling the Last Supper, commemorated the Crucifixion. This was also a thanksgiving ceremony; the Greek name for it was *eucharist*, meaning "the giving of thanks." This common meal gradually became impracticable as the Christian communities grew larger, and the Lord's Supper was thereafter observed at the conclusion of the public portion of the scripture service; the unbaptized withdrew so that the baptized could celebrate together.

Thus, from the very inception of Christianity, food and beverage has symbolized that religious experience is not purely personal but also communal. Moreover, differences in interpretation of the Lord's Supper have provided some of the contrasts among the major Christian churches. The opposing views of Roman Catholics and Protestants over whether the Eucharist bread is changed in substance or is a symbol of the flesh of Christ is an example of the role of food as a representation of religious differences within Christianity.

The rituals of the Eucharist provide the clearest examples in the Christian churches or confessions of the relationship between social stratification and food behaviour. Christianity, unlike Judaism or Hinduism and other Asian religions, was never tied to a caste system; correspondingly, it repudiated the entire body of purity–pollution laws of the Old Testament. Christianity was, however, part of the early European social system that was based on clear-cut separations of social classes. Religious food customs in Christianity, most notably in the Eucharist, reflects this.

The first Christian churches developed alongside the most rigid social stratification in European history, with elaborate notions of class authority and superiority and subordination. The separation of those in authority from the masses of ordinary people is mirrored in the Roman eucharistic ritual in which the sacrament's celebrant—the officiating priest—partook of the bread and wine first and then served only the bread to those of the faithful who wished it.

With the Reformation during the 16th century, which was (among other things) an overthrow of the traditional social order, a slight but important change in the eucharistic ritual was introduced, reflecting the weakening—but not the abandonment—of stratification and its attendant hierarchies of authority. In many Protestant confessions the officiating minister also partook of the bread and wine first, then served it to the congregation. In the Presbyterian ritual, the minister partook first and then served it to the elders who then served the people. Although this continued to reflect a system of stratification, it was a radical departure from the Roman rule that only the officiating priest could serve everyone. These rules for both Roman Catholics and Protestants are gradually changing in the 20th century.

Fast and abstinence

Until relatively recently, the most notable dietary law in Christianity was the Roman Catholic prescription to abstain from eating meat on Friday. This ban was lifted as part of the modernization of Roman Catholicism that was begun during the reign of Pope John XXIII. In Roman Catholic abstinence meat is forbidden, but there is no restriction on the amount of food eaten; fasting means that the quantity of food is also restricted. Historically, there have been several categories of fasts. The 40 days of Lent have traditionally been a period of mortification, including practices of fast and abstinence; the rules, however, have been greatly modified in recent years. Ember Days—a Wednesday, Friday, and Saturday at each of the four seasons—seem to be survivals of full weekly fasts formerly practiced four times a year. Vigils are single fast days that have been observed before certain feast days and other festivals. Rogation Days are the three days before Ascension Day and are marked by a fast preparatory to that festival; they seem to have been introduced after an earthquake about 470 as penitential rogations, or processions, for supplication.

Also important in the Christian complex of fasting is that associated with monastic life. Mortification is seen as essential to the practice of asceticism, and, in many rules of monastic life, fasting is regarded as one of the most efficient exercises of mortification.

Religions of India. It is in the religions of India that one can most clearly observe the principles outlined above concerning the relationship between dietary laws and customs and the existence of social stratification, traditional privilege, and social, familial, and moral lines that cannot be crossed. Hinduism provides the best example, although the same principles also obtain in the religions of Jainism and Sikhism.

Relationship between dietary customs and caste systems

Food observances help to define caste ranking: Brahmins are the highest caste because they eat only those foods prepared in the finest manner (*pakkā*); everyone else takes inferior (*kaccā*) food. *Pakkā* food is the only kind that can be offered in feasts to gods, to guests of high status, and to persons who provide honorific services. Food is regarded as *pakkā* if it contains ghee (clarified butter), which is a very costly fat and which is believed to promote health and virility. *Kaccā* is defined as inferior because it contains no ghee; it is used as ordinary family fare or as daily payment for servants and artisans. When food serves as payment for services (*e.g.*, barbering), the quality of the food depends on the relative ranks of the parties to the transaction; the person making the payment gives inferior food, such as coarser bread, to a lower ranking person performing the service. Performance of a service denotes that a person is ready to accept some kind of food, and giving food denotes an expectation that a service will be performed. Members of subordinate castes pick up the dirty plates of members of superior castes, as at village feasts. Food left on plates after eating is defined as garbage (*jūṭhā*); it is felt to have been polluted by the eater's saliva. This garbage may be handled in the family by a person whose status is lower than the eater's, such as a wife. Such food may be fed to domestic animals; among humans outside the family it can only be given to members of the lowest castes, such as sweepers. The highest Brahmins do not accept any cooked food from members of any other caste, but uncooked food may be received from or handled by members of any caste. Nor will such Brahmins accept water across caste lines. Cow's milk is ritually pure and cannot be defiled, but a Brahmin will not accept milk from an untouchable—a member of the lowest caste groups—lest it has been diluted with water.

Water is easily defiled, but, if it is running in a stream or standing in a reservoir, it is not polluted even by an untouchable in it. Water in a well or container, however, is defiled by direct or indirect contact with a person of low caste. Thus, a ritually observant Brahmin will not allow a low-caste person to draw water from his well, although this rule is lapsing, possibly because of the introduction of plumbing and the removal of water from the list of scarce resources.

In the general Hindu system of purity–pollution, meats are graded as to their relative amount of pollution. Eggs are the least and beef the most defiling; but the highest caste Brahmins avoid all meat products absolutely. Also, certain strong foods (*e.g.*, onions and garlic) are thought to be inappropriate to Brahminical status. Alcohol too is prohibited; it is not considered polluting in itself, but the prohibition seems related to the Brahminical value of self-control. Alcohol's manufacture and trade is confined to members of lower castes.

People who eat at each other's feasts hold equal rank. People who eat at every house in a village occupy a very low status, and refusal to take food from another constitutes a claim to higher caste rank. More generally, givers of food outrank receivers. This, however, is a definition of collective, not of individual, rank. If a member of one caste gives food to a member of a second, all members of the first caste are regarded as higher than a third, even if there is no direct transaction between the first and third castes. Thus, the behaviour of every person in a village has consequences for the entire village.

Local variability of dietary restrictions

In actual practice, however, there is not an automatic enactment of these formal rules in village life; instead, they vary considerably according to local conditions. For instance, one of the formal rules of Hindu religious caste organization is that vegetarians outrank meat eaters, because contact with killed animals is regarded as polluting. Nevertheless, McKim Marriott, a U.S. anthropologist who has investigated village caste relationships, has found instances in which meat eaters outrank vegetarians. He concludes from his observations that it is caste rank—mostly in terms of the kinds of work that people in different castes do—that determines purity and pollution. In daily social relations this sometimes means that a caste of sufficiently high status may not be demeaned by receiving food from a lower caste if the latter is not too far below and if the proper food and vessels are used.

Status is rarely immutable over long stretches of time. In most societies, people who occupy low status try to exploit every opportunity to improve their position, and, Marriott found, Indian villagers are no exception. Because food in this culture is one of the principal indices of rank, it is used as a pawn in manoeuvres for social mobility. Specifically, members of a low caste will try to gain dominance over persons in another by feeding them, although the latter cannot be too far above the upwardly mobile group. There is no direct way of forcing a higher group to accept food; one of the techniques most often used is for the lower caste to threaten to withhold services unless a heretofore slightly higher caste receives food from the former. Such mobility, as noted earlier, affects not only the two castes concerned but also all other groups in the village, and the manoeuvring involves everyone in the community.

Marriott's emphasis on occupation (and, therefore, rank) as the determinant of food customs has not been accepted by all students of Indian society. He continues to leave some aspects of caste behaviour unexplained, such as the extreme statuses of Brahmins and untouchables, to say nothing of the existence of the total caste system itself and the mechanisms by which it is maintained. These problems have yet to be worked out. In any case, there can be no doubt that concepts of pollution and purity in regard to food in India, as everywhere else, are governed by a systematic set of rules analogous to a lan-

guage's grammar and that applications of the rules are logical and consistent within the grammatical framework. Observations of daily village life do not contradict this concept of the codification of food rules; they only suggest that earlier "grammars" may have been too narrowly conceived.

Lack of unity in Buddhism

Buddhism. Buddhism is, perhaps, the most difficult religion to discuss in terms of dietary laws and customs because it does not have any unity; its tradition has a complex history, and individual believers are characterized by varied faiths. Though Buddhism originated in India, it also diffused to—and had a great impact on—Ceylon, Tibet, China, and Japan. In each case, it was reshaped to conform with local conditions, especially those of social stratification. For example, most of the countries of Southeast Asia have caste systems in which there are outcastes or untouchables; Buddhism has been important in supporting such systems. Specifically, untouchability and the occupation of butchering animals tend to go together both in Buddhism and in many of the countries of Southeast Asia. But Burma, where Buddhism is the dominant religion, is an exception; having no caste system, Burmese society has not made butchering a basis of untouchability.

Buddhism developed its own class distinctions, most notably between the monastic elite and the lay devotees. The social and political ethic of the laity was based on a merit-making ethic that was geared primarily to the urban mercantile and artisan classes. Thus, Buddhism claimed from its inception to be a Middle View (Mādhymika), opposed equally to the extremes of sensuousness and indulgence and of self-mortification. This Middle View was exemplified in the "five precepts": no killing, stealing, lying, adultery, or drinking of alcoholic beverages. These precepts were translated into an ethic of moderation in diet. A person must allay his hunger so that he may practice the religious life. Buddhism holds that man is weak and helpless by himself; thus it sees the purpose of religious action as bringing a return from the deities. Deriving from this is the practice of holding ritual vegetarian feasts for large numbers of monks, a noble patron, or for the benefit of a departed soul to promote health and longevity. Another Buddhist custom is the issuing of a prohibition against killing animals to end a drought or to speed the recovery of a sick emperor. According to the Vedic treatise the *Śatapatha Brāhmaṇa*, food, when enclosed in the body, is linked to the body by means of the vital airs. The essence of food is invisible. Food is the highest of all things that can be swallowed, and food and breath are both gods.

The prohibition of killing animals is more stringent in Buddhism than the injunction against eating them. Buddhism allows pure flesh to be eaten if it has not been procured for eating purposes or if the eater has not supposed it to be. The sin is upon the slayer, not the eater. This notion has been used in India and Japan to justify the outcasting or untouchability of butchers.

Religions of China. China is an example of the proposition offered above that religion alone does not give rise to eating rules; instead, religion serves to legitimate customary patterns of behaviour and social relations that emerge out of economic (especially occupational) and political relationships. Although China was under strong Buddhist influence, the Chinese never developed the institution of untouchability or outcaste. Indeed, Buddhism did not really penetrate China until after the beginning of the 2nd century AD; during the previous century, Buddhism was confined to foreigners in the northern commercial cities.

Chinese culture before the Han dynasty

Before 200 BC (the approximate beginning of the Han dynasty), Chinese culture was based on a rather elaborate system of social stratification in which mobility was rare and difficult. It was, in other words, a relatively closed social system, if not feudal. During this time, there were restrictions on the consumption of food: beef, mutton, and pork were to be eaten by an emperor; beef by feudal lords; mutton by high-ranking state ministers; pork by lower ministers; fish by generals; and vegetables by commoners (who probably could not afford meat or fish anyway). Officials, in fact, were known as "meat eaters," and it was generally only the aged commoners who were allowed to eat meat.

During this time, military affairs and sacrifices were considered the two most important things in the state. Sacrifice was inseparable from veneration of the ancestors, and almost no ceremonies were conducted without sacrifice and offerings. These ceremonies were integral features of daily life and, as a result, foodstuffs became associated with the moral code that was based on maintaining fixed social and political relationships. God and ancestors were often referred to as those "who are sacrificed to," and disobedience to them was believed to result automatically in catastrophe. Ceremonies marking important personal transitions (*e.g.*, initiation to adulthood and marriage) were held in the ancestral temple and were accompanied by feasts and sacrifices to the ancestors.

Ancestral veneration and the ethos of religiously validated legitimate authority remained as integral features of Chinese culture until the most recent years. Religious belief and observance notwithstanding, however, Chinese culture underwent a drastic change with the establishment of the Han dynasty. Most notably, the social class system was opened up—at least ideally—by the adoption of the principle of recruitment for public office; in later dynasties, this was expanded into the well-known system of written examinations, of grading officeholders by merit, and other features of the famous Chinese civil service. Correlated with the removal of the barriers to social mobility and establishment of the principle of ideally open recruitment to the civil service, the pre-Han food restrictions disappeared. This was also the time of Buddhism's greatest thrusts into Chinese thought and life.

Food continued to occupy an important social and religious place in villages, at least until the establishment of the People's Republic of China. For instance, marriage ceremonies traditionally last four days; the highlight of each day's celebration is a feast or sacrifice to the ancestors, sometimes both. Feasts and sacrifices are also important features of funerals, some of which are marked by two feasts in one day. These ceremonial occasions often work considerable economic hardship on families, forcing many of them into debt.

Religions of Japan and Korea. Japan and Korea exhibit many of the same characteristics with respect to food customs as India, though with much less elaboration, and thereby the same relationships to Buddhism, though in an opposite direction. These relationships to Buddhism are also highlighted by contrasting Japan and Korea with China. Whereas post-Han China placed emphasis on achieved status and on personal superiority rather than on considerations of race or blood as a basis of social position, Japan and Korea (and also Tibet) established and continued a system of hereditary status and outcasting. As in India, therefore, the Japanese and Koreans considered pollution to be a hereditary taint; Buddhism played a major role in the legitimation of this ideology.

Outcaste groups

Outcastes in Japan are often referred to pejoratively as Eta (*eta* meaning "very dirty" or "defiling"). The accepted usage among students of Japanese society for outcastes is *burakumin*, meaning "people of a special village." They are crudely discriminated against in employment and intermarriage, live rurally or in slum conditions, have the lowest educational levels in the nation, and often suffer from malnutrition. They are required to wear special clothing, slippers, and hair styles; to stay away from other households; to remain in their own hovels at night; and to prostrate themselves before higher-caste people.

The history of the Japanese caste system in respect to food customs gives important clues to its origin. Among the ancient Japanese, meat was included in the diet, and the flesh of animals, fishes, and birds was offered to the gods as sacrifice. The flesh of ox, horse, dog, monkey, and fowl was prohibited, but deer, rabbit, and pig were not. During the 8th century AD, the Japanese began to depend

mostly upon plant rather than animal foods. In Japan's limited territory, it is understandable that cattle were raised for plowing and other agricultural work rather than for meat and milk. In 741 a law was passed forbidding the killing of cattle and horses, the latter being necessary for military as well as productive purposes. This provided a conducive atmosphere for Buddhist influences in the 6th and 7th centuries (primarily from China and Korea) that stressed the abhorrence and ritual impurity of blood and death.

Shintō

Buddhism, however, was only one of several sources of outcasting slaughterers and butchers. During the 8th century, Shintō—the only indigenous religion of Japan—began to stress concepts of uncleanness as things that are displeasing to the gods: wounds, disease, death, menstruation, and childbirth; and this too contributed strongly to the development of Eta status. It was apparently about this time that the belief developed in Japan that a person's association with blood and death changed his nature; this contamination not only carried over to a man's descendants but was thought to be communicable. It was apparently also at this time that Japanese cuisine began to favour fish (especially raw fish) as a staple source of protein.

Important in this connection is that occupational specialization began to flourish in Japan during the 9th and 10th centuries; by this time, Buddhism was widespread in Japan. Traditional occupational roles became spheres of monopoly; in the face of competition from economically specialized groups who forced them out, people dealing with slaughtering, butchering, and tanning began to form guilds. This was rationalized by Buddhist and Shintō ideas that occupations associated with animal slaughter and processing (confined to Eta) should be separated from the general body of commoner and slave occupations.

During the Heian period (794–1185), communities whose members were engaged in occupations related to death and animal products were forced outside the normal society, and they thus came to form the main body of outcastes in Japan. Increasingly, the latter were outcasted and considered untouchable, a pattern that reached its heights in the Tokugawa period (1603–1867). By the 17th century, the idea developed—supported by Shintō and Buddhism—that eating the flesh of all animals caused pollution for 100 days.

Outcasting dies slowly. Though the egalitarian ideologies of modern industrialization are incompatible with caste, outcasting tends to remain in Japan and, alongside it, many of the food customs associated with the caste system. As in India, eating together (along with marriage and social visiting) between untouchables and members of normal society is disdained. In many parts of Japan, especially in traditional villages, the diet tends to be largely vegetarian, though fish (and occasionally chicken) accompany banquets. Pigs are raised for sale to Eta, and pork is eaten in cities. Horses and cows are still used only as beasts of burden. Milk continues to be considered dirty and tends to be drunk only on a doctor's prescription. Also, as in all traditionalist groups, corporate groups of households (*kumi*), especially in villages, periodically hold festivals at their village shrines; each household takes a turn in providing food and beverage. The core of the festivity lies in drinking sake, an alcoholic beverage, and eating a grand meal.

BIBLIOGRAPHY

General works: DONALD E. CARR, *The Deadly Feast of Life* (1971), a popular account of food habits and nutritional behaviour; MARY DOUGLAS, *Purity and Danger: An Analysis of Concepts of Pollution and Taboo* (1966), widely regarded as the definitive analysis of the subject; CRAIG MCANDREW and ROBERT B. EDGERTON, *Drunken Comportment: A Social Explanation* (1969), an exploration of the ways people are expected to behave under the influence of alcohol in different cultures.

Tribal societies: The following are some accounts of tribal societies that also provide good information on food customs and dietary laws. RAYMOND FIRTH, *We, the Tikopia: A Sociological Study of Kinship in Primitive Polynesia* (1936); MEYER FORTES, "Pietas in Ancestor Worship," *Jl. R. Anthrop. Inst.*, 91:166–191 (1961), reprinted in *Man in Adaptation*, vol. 3, *The Institutional Framework*, ed. by YEHUDI A. COHEN, pp. 207–226 (1971); MARGARET MEAD, *The Mountain Arapesh*, vol. 2 (1970).

Judaism and Christianity: The basic sources in connection with both these major religions are, of course, the Old Testament (especially Leviticus II, Deuteronomy 14, and the prophets) and the New Testament (especially Acts, Luke, Mark, and Romans). In addition, the following sources may be consulted: JOHANNES PEDERSEN, *Israel: Its Life and Culture*, 4 vol. (1926–40); MARK ZBOROWSKI and ELIZABETH HERZOG, *Life Is with People: The Jewish Little-Town of Eastern Europe* (1952; reprinted as *Life Is with People: The Culture of the Shtetl*, 1962).

Islām: In addition to the Qu'rān, the following sources may be consulted: AMEER ALI, *Mohammedan Law*, 5th ed., 2 vol. (1929); CHARLES CUTLER TORREY, *The Jewish Foundation of Islam* (1933).

India: LOUIS DUMONT, *Homo hierarchicus, essai sur le système des castes* (1967; Eng. trans., *Homo Hierarchicus: The Caste System and Its Implications*, 1970); EDWARD B. HARPER (ed.), *Religion in South Asia* (1964), especially Harper's "Ritual Pollution As an Integrator of Caste and Religion," pp. 151–196; EDMUND R. LEACH (ed.), *Aspects of Caste in South India, Ceylon, and North-west Pakistan* (1960); DAVID G. MANDELBAUM, *Society in India*, 2 vol. (1970); MCKIM MARRIOTT, "Caste Ranking and Food Transactions: A Matrix Analysis," in *Structure and Change in Indian Society*, ed. by MILTON B. SINGER and BERNARD S. COHN, pp. 133–171 (1968).

Buddhism: KENNETH K.S. CH'EN, *Buddhism: The Light of Asia* (1968); CHARLES NORTON ELIOT, *Hinduism and Buddhism: An Historical Sketch*, 3 vol. (1921).

Japan: ROBERT N. BELLAH, *Tokugawa Religion: The Values of Pre-Industrial Japan* (1957); GEORGE DE VOS and HIROSHI WAGATSUMA (eds.), *Japan's Invisible Race: Caste in Culture and Personality* (1966).

China: KENNETH K.S. CH'EN, *Buddhism in China* (1964); ARTHUR F. WRIGHT, *Buddhism in Chinese History* (1959).

(Y.A.C.)

Differential Equations

Differential equations are equations that relate a function f to its derivatives. This type of equation finds widespread application in science and technology because it can be used to express natural laws that describe the behaviour of rates of change of quantities. A well-known example in physics concerns radioactivity. A radioactive element changes spontaneously into a stable element, a process that is known as radioactive decay or disintegration. The law describing this behaviour states that the rate of decay with time—that is, the amount of a substance changing per second, say—is proportional to the amount of substance present. Initially, when the material is a pure radioactive element, this rate of decay is high. As the radioactive element changes into a stable element, however, the rate of change falls because there is less radioactive material. Therefore, the rate of decay decreases continuously with time.

In this process the quantity of interest is the amount of substance (n) remaining after a given time; *i.e.*, the number of atoms or the mass. Clearly this quantity will depend in some way on the time that has elapsed; in other words, it is a mathematical function of time and can be denoted by $n(t)$.

A formulation of the law is that $\frac{dn}{dt}$ equals the product of $-\lambda$ and n (see Box, law 1). In this case n represents the amount of radioactive material, λ is a constant, and $\frac{dn}{dt}$ is a notation for the rate at which n increases with time. The minus sign is required because n is decreasing with time and not increasing. $\frac{dn}{dt}$ is known as the derivative of n with respect to t. An idea of the nature of a derivative can be obtained by considering the behaviour of a specific function of x, say x^2. If x has the value a (a real number) this function has the value a^2. If this value now changes to $(a + h)$, the value of the function becomes $(a + h)^2$. The change in the value of the function is $(a + h)^2 - a^2$, and this change has occurred as a result of x changing by h. Thus the rate at which x^2 has

changed with respect to the change in x is given by the quotient: $\frac{(a+h)^2-a^2}{h}$. This gives an average rate of change over the value h. If h now becomes smaller, and tends to zero, the quotient tends to the value $2a$, because when the $(a+h)^2$-term is expanded and the a^2-term subtracted from it, the division yields $2a+h$ (see 2). $2a$ is the rate of change of x^2 with x at the point a and is known as the derivative of the function. If a is allowed to vary—*i.e.*, to take any values—a new function $2x$ is obtained that for a given value of x expresses the rate of growth of x^2 with x at that value. In general a function of a single real variable x can be denoted by $u(x)$ and its derivative by $\frac{du(x)}{dx}$ or by $u'(x)$.

This process can be repeated: the derivative of $u'(x)$ can be formed and denoted by $u''(x)$ or by $\frac{d^2u(x)}{dx^2}$, and similarly for higher derivatives $u^{(n)}(x)$. When the independent variable is time and the dependent variable distance, the first derivative is velocity, the rate at which distance changes with time. The second derivative is acceleration, the rate at which velocity changes.

Returning to the example of radioactive decay, the equation involves a derivative and is one example of a differential equation. The solution of this differential equation is the function $n(t)$, which shows how n depends on t. The problem of solving it is that of finding the function of t the derivative of which is $-\lambda n(t)$.

A common example of a differential equation that involves a second derivative is the equation describing the motion of a simple pendulum under the influence of gravity. The second derivative of the angular displacement from the vertical θ is proportional to the sine of the angular displacement (see 3). The equation involves the length a of the pendulum and g, the acceleration of a falling body under the influence of gravity, called the acceleration of free fall. The equation may be expressed in words as follows: the rate of increase with time of the rate of increase of angular displacement with time (the angular acceleration) is directed toward the vertical position and is inversely proportional to the length of the pendulum and directly proportional to the sine of the angular displacement. This rather complicated verbalization illustrates the economy and precision of the mathematical formulation.

Most physical situations are too complicated to be expressed simply by one variable depending on another. For example, the position of a body in a vertical plane is given by two coordinates x and y, indicating the distances from a vertical and horizontal axis respectively. The motion of a body moving in this plane can be expressed by an equation involving these two coordinates (x, y), each of which depends on the time t.

In particular, if air resistance is neglected, a particle moving under the influence of gravity is described by a pair of differential equations expressed in terms of second derivatives (see 4), in which g denotes the acceleration caused by gravity. There is no acceleration in the horizontal direction because there is no force in this direction.

A further example is provided by geometry. It is shown in calculus that the curvature k of a curve the equation of which in Cartesian coordinates is $y = y(x)$ is defined in terms of the second derivative of y (see 5). A curve with the property that the rate of change with respect to x of its curvature is a prescribed function $\psi(x)$ therefore has an equation that satisfies the equation $d\kappa/dx = \psi(x)$, which is equivalent to an equation involving a third derivative (see 6). By using the alternative notation of primes denoting differentiation (see 7), it is also possible to write this equation in a more simple form (see 8). For example, a circle has constant curvature so that it has $\psi(x) \equiv 0$; hence, all circles are symbolized by the same differential equation with the term involving $\psi(x)$ set equal to zero (see 9).

The preceding examples are all equations that involve functions of a single variable. Derivatives of such functions are called ordinary derivatives, and, therefore, equations of this type are called ordinary differential equations. The pair of equations involve simultaneously the derivatives of two independent variables x and y and are an example of a pair of ordinary simultaneous differential equations.

The transverse displacement of a vibrating string is determined by an equation of a totally different kind. If the string is at rest stretched to a tension T between points $x = 0$ and $x = a$ of the x-axis, then if the string is set vibrating the displacement, y, of a point that originally had a coordinate x will depend not only on x but also on the time t. In other words, y is a function of two independent variables, x and t. Its variation is governed by the fact that the second partial derivative with respect to x is proportional to the second partial derivative with respect to t (see 10), the proportionality constant $c = \sqrt{T/\sigma}$, σ being the mass per unit length of the string. The derivatives in this equation are partial derivatives, and the equation is an example of a partial differential equation. The theory of such equations is markedly different from that of ordinary differential equations and is considered later.

This article is divided into the following sections:

Ordinary differential equations
- Historical background
- Types of problems solvable with this discipline
- Types of solutions
- Existence and uniqueness of solutions
- Forms of solutions
- First-order equations
- Linear differential equations
- Nonlinear differential equations
- Boundary value problems
- Systems of differential equations
- Perturbations
- Stability

Partial differential equations
- Partial differential equations of special interest
- Classification of partial differential equations
- Initial value and boundary value problems
- Systems of partial differential equations
- Techniques of solution of partial differential equations

(1) $$\frac{dn}{dt} = -\lambda n$$

(2) $$\frac{(a+h)^2-a^2}{h} = \frac{a^2+2ah+h^2-a^2}{h} = 2a+h$$

(3) $$a\frac{d^2\theta}{dt^2} = -g\sin\theta$$

(4) $$\frac{d^2x}{dt^2} = 0, \qquad \frac{d^2y}{dt^2} = -g$$

(5) $$\kappa = \left[1+\left(\frac{dy}{dx}\right)^2\right]^{-3/2}\frac{d^2y}{dx^2}$$

(6) $$\left[1+\left(\frac{dy}{dx}\right)^2\right]\frac{d^3y}{dx^3} - 3\frac{dy}{dx}\left(\frac{d^2y}{dx^2}\right)^2 = \left[1+\left(\frac{dy}{dx}\right)^2\right]^{5/2}\psi(x)$$

(7) $$y' = \frac{dy}{dx}, \qquad y'' = \frac{d^2y}{dx^2}, \qquad y''' = \frac{d^3y}{dx^3}, \cdots$$

(8) $$[1+(y')^2]y''' - 3y'(y'')^2 = [1+(y')^2]^{5/2}\psi(x)$$

(9) $$[1+(y')^2]y''' = 3y'(y'')^2$$

(10) $$c^2\frac{\partial^2y}{\partial x^2} = \frac{\partial^2y}{\partial t^2}$$

Existence and uniqueness of solutions to partial differential equations
Generalized theory of partial differential equations
Special functions
Dynamical systems on manifolds
Examples of state spaces
Manifolds
Dynamical systems as vector fields on manifolds
Examples of dynamical systems
Gradient dynamical systems
Hamiltonian mechanics

Ordinary differential equations

Equations of the first order and first degree

An ordinary differential equation is a relation connecting the function y of an independent variable x and its derivatives $y', y'', \cdots, y^{(n)}$ (see 11). If the derivative of the highest order occurring in the equation is $y^{(n)} = d^n y/dx^n$, the equation is said to be of order n. For example, equation (1) is a first-order equation, and equations (3) and (4) have an order of two. The power to which the highest derivative occurs is called the degree of the equation. Equations (1), (3), and (4) are of the first degree; a second-order differential equation involving $(y'')^2$ and no higher of y' (see 12), however, is of the second degree because the highest derivative occurs to the power 2. If the equation is of a form (see 13) in which the function y and its derivatives occur linearly and in which the coefficients $a_n(x), \cdots, a_1(x), a_0(x)$ depend only on x and not on y or any of its derivatives, the equation is said to be linear. If $\psi(x) \equiv 0$ (equal for all values of x), the equation is said to be homogeneous; and if $\psi(x)$ is not equal to zero for all values of x, it is said to be nonhomogeneous. Equation (1) is linear and homogeneous, but equation (3) is nonlinear.

HISTORICAL BACKGROUND

The study of differential equations is almost as old as that of the calculus itself. Sir Isaac Newton discovered a method of infinite series and the calculus in 1665–66. In 1671 he wrote an account of his theory of "fluxions," a fluxion being a derivative of a "fluent," the name Newton gave to his dependent variables. Newton discussed "fluxional equations," or as they are now called, differential equations. These he divided into three categories. In modern notation, the first category is that in which dy/dx is a function of x alone or of y alone; the second consists of ordinary differential equations of the first order of the form $dy/dx = f(x, y)$; and the third is made up of partial differential equations of the first order. Newton derived solutions of differential equations by a method using power series with indeterminate coefficients, which he claimed to be universally effective. Though written in 1671 in Latin, his work, entitled *The Method of Fluxions and Infinite Series*, was not published until 1736. As a result it had little or no influence on the development of the theory of differential equations.

A greater impact was made by the work of the German mathematician Gottfried Wilhelm Leibniz. Although his investigations were not begun until 1673, Leibniz became known through the publication of his results in 1684 in the *Acta Eruditorum* ("Deeds of Distinguished Men"), a scientific journal that had been established only two years earlier.

Foremost among the devoted followers of Leibniz were the Swiss mathematicians, the brothers Jakob and Johann Bernoulli. With others of their family, they played a notable part in the development of the theory of differential equations and of the use of such equations in the solution of physical problems.

In May 1690, in a paper in the *Acta Eruditorum* Jakob Bernoulli showed that the problem of determining the isochrone—*i.e.*, the curve in a vertical plane such that a particle will slide from any point to the bottom in exactly the same time, no matter what the starting point—is equivalent to that of solving a first-order nonlinear differential equation (see 14). He then solved the equation by what is now called the method of separation of variables; the general method was enunciated by Leibniz in the following year. Jakob Bernoulli's paper of 1690 is a milestone in the history of calculus, for in it the term integral occurs in the literature for the first time.

In 1692 Leibniz discovered the method of solving first-order equations of homogeneous type and still later that of solving linear equations of the first order. The problem of finding the general solution of what is now called Bernoulli's equation was proposed by Bernoulli in 1695 and solved by Leibniz and Johann Bernoulli by different methods. Thus within a few years of the birth of the infinitesimal calculus, most of the known methods of solving first-order ordinary differential equations had been developed.

Early applications

Numerous applications of the use of differential equations in deriving the solutions of geometrical problems were made before 1720. Among these may be mentioned the problem of determining a plane curve the curvature of which is a prescribed function of position and that of finding the orthogonal trajectories of a given family of plane curves. Some of the differential equations formulated in this way were of the second or higher order. Second-order equations of the form $F(y, y', y'') = 0$, in which the independent variable x does not appear explicitly, were discussed as early as 1712 by an Italian mathematician, Iacopo Francesco Riccati, who gave his name to Riccati's equation (see below) although, in fact, Jakob Bernoulli had earlier studied the special case $y' = x^2 + y^2$. Some of the younger members of the Bernoulli family, particularly Daniel, contributed to studies of Riccati's equation. Important work on first-order equations of exact type (see below) and of first-order equations of degree higher than the first was done in the 1730s by Alexis-Claude Clairaut, a Frenchman who was one of the most precocious mathematicians of all time.

$$(11) \qquad f(x, y, y', \cdots y^{(n)}) = 0$$

$$(12) \qquad (y'')^2 + (y')^3 + y^4 = 0$$

$$(13) \qquad a_n(x)y^{(n)} + \cdots + a_1(x)y' + a_0(x)y = \psi(x)$$

$$(14) \qquad (b^2y - a^3)^{1/2}\, y' = a$$

$$(15) \qquad (1 - x^4)^{1/2}\, y' + (1 - y^4)^{1/2} = 0$$

The second period in the history of differential equations was dominated by the Swiss mathematician Leonhard Euler, who made many contributions to the theory, starting in 1728. He introduced several methods of lowering the order of an equation, the concept of an integrating factor, the theory of linear equations of arbitrary order, the development of the use of series solutions, and the discovery that a first-order nonlinear differential equation with square roots of quartics as coefficients (see 15) has an algebraic solution. This last result, which is a special case of Abel's theorem, led to the theory of elliptic functions created in the 1820s by Niels Henrik Abel of Norway and Karl Gustav Jacob Jacobi of Germany.

Much of the early work on differential equations was concerned with the discussion of basic questions concerning what is meant by a function or by the solution of a differential equation. Such questions continued to be raised because the answers found satisfactory by one generation of mathematicians were challenged by their successors. The first attempt to establish a rigorous theory of functions was made by Abel and the French mathematician Augustin-Louis Cauchy, in the 1820s. Cauchy's work in real analysis gave an entirely new direction to the theory of ordinary differential equations, steering it away from the investigating of techniques of solution into the asking of general questions about the existence and uniqueness of solutions. Cauchy himself proved the first existence theorem and gave methods of

deriving solutions through limiting processes. The theory of analytic functions, also due to Cauchy, led to the creation of the theory of differential equations in the complex domain and, in turn, to the study of functions of several complex variables.

Recent work

Much recent work in ordinary differential equations is of a basic nature, concerned with the conditions that guarantee the existence of a solution of a given equation; the theory is more concerned with establishing that a solution exists than with trying to derive a closed form for it. Such an attitude is essential when many of the practical problems involving differential equations are solved by the use of electronic computers; the validity of numerical processes must be thoroughly investigated. The study of differential equations continues to contribute to the solution of practical problems in control theory, in orbital mechanics, and in many other branches of science and technology, and also to ask challenging questions of pure mathematicians working in such apparently abstract subjects as functional analysis and the theory of differentiable manifolds.

TYPES OF PROBLEMS SOLVABLE WITH THIS DISCIPLINE

Some simple examples of how differential equations may arise in the analysis of problems in physics and geometry have already been given. In such problems the unknown quantity has to satisfy not only a differential equation but also some other conditions. For example, the quantity $n(t)$ may have to satisfy the condition that at $t = 0$ it takes a prescribed value N. Some other problems that can be solved with the help of differential equations are listed below.

Problems in mechanics. *The catenary.* If a cable of uniformly distributed mass m per unit length is suspended between two points A and B, the curve formed by the cable is called a catenary. If a coordinate system is chosen with an origin vertically below the lowest point of the curve, and with a y-axis that is vertically upward, the shape of the curve is determined by a second-order nonlinear differential equation (see 16), involving a constant a (with the dimensions of length). The constant is expressed in terms of T, the vertical tension acting on the cable at its lowest point, and g, the acceleration caused by gravity, by the formula $a = T/(mg)$. If the distance of the origin below the lowest point of the cable is taken to be a, the required solution must satisfy the conditions $y(0) = a$, $y'(0) = 0$.

Motion in a line under known forces. The motion of a particle of mass m projected vertically upward in a straight line with velocity v is described by a differential equation relating the time derivative of mv to g (see 17) and the resistance of the medium $R(v)$, the function R being known. If the mass m is constant, the left side of the equation reduces to the time derivative of the velocity (see 18). If m is a prescribed function of v, the left side can be attacked by the rule for differentiating a product (see 19).

Escape velocity from the Earth. If the acceleration caused by gravity at the Earth's surface is g, then, at time t, the distance r from the Earth's centre of a particle moving vertically upward is determined by a differential equation which states that the second derivative of v is proportional to $(R/r)^2$ (see 20), in which R is the radius of the Earth. If the initial velocity of projection is V, then the initial conditions are the position and velocity at $t = 0$ (see 21).

The motion at any point in the upward trajectory is determined by the solution of this initial value problem. If the particle is to escape from the Earth's gravitational field, $(dr/dt)^2$ must exceed zero for all values of r. It turns out that this condition is satisfied only if V^2 is greater than or equal to $2gR$. For this reason $\sqrt{(2gR)}$ is called the escape velocity from the Earth.

Vertical motion of a rocket. A problem of interest is that of a rocket ascending in a straight line under the influence of gravity with a thrust that is constant both in magnitude and direction.

The thrust of the rocket is produced by the ejection of mass at a constant rate f with exhaust velocity V_e, which when measured relative to the rocket is constant. The height x above the Earth's surface at a time t while the rocket is still firing is determined by the non-linear differential equation (see 22) relating the second derivative of x to an expression of the form $(1 + x/R)^2$ and involving g, the acceleration caused by gravity at the Earth's surface; R, the radius of the Earth; and m_0, the initial mass of the rocket. The initial conditions in this case are that $x = 0$, $dx/dt = V$ when $t = 0$.

Motion of a planet under an inverse square law. If a planet is moving in a plane under an inverse square law, then its motion can be described by the plane polar coordinates r, θ, the origin of which is situated at the centre of force.

The inverse square law relates the radial force on the planet to the inverse of the square of the distance from the origin. The motion is described by a pair of second-order differential equations derived from Newton's second law (see 23) involving a constant c.

(16) $$a\frac{d^2y}{dx^2}=\left[1+\left(\frac{dy}{dx}\right)^2\right]^{1/2}$$

(17) $$\frac{1}{m}\frac{d}{dt}(mv)=-g-R(v)$$

(18) $$\frac{dv}{dt}=-g-R(v)$$

(19) $$\left[1+\frac{vm'(v)}{m(v)}\right]\frac{dv}{dt}=-g-R(v)$$

(20) $$\frac{d^2r}{dt^2}=-g\left(\frac{R}{r}\right)^2$$

(21) $$r=R, \qquad \frac{dr}{dt}=V, \quad \text{when } t=0$$

(22) $$\frac{d^2x}{dt^2}+g\left(1+\frac{x}{R}\right)^{-2}=\frac{V_e f}{m_0-ft}$$

(23) $$\begin{cases}\dfrac{d^2\rho}{dt^2}-\rho\left(\dfrac{d\phi}{dt}\right)^2=-c\rho^2\\[2ex] \dfrac{d}{dt}\left(\rho^2\dfrac{d\phi}{dt}\right)=0\end{cases}$$

(24) $$\begin{cases}L\dfrac{dx}{dt}+Rx+\dfrac{q}{C}=E(t)\\[2ex] L\dfrac{d^2q}{dt^2}+R\dfrac{dq}{dt}+\dfrac{q}{C}=E(t)\end{cases}$$

(25) $$\begin{cases}L_1\dfrac{dx}{dt}+R_1x+\dfrac{q}{C}+M\dfrac{dy}{dt}=E(t)\\[2ex] M\dfrac{dx}{dt}+L_2\dfrac{dy}{dt}+R_2y=0, \quad \text{with } x=dq/dt\end{cases}$$

Problems in other areas of physics. *Electrical circuits.* The electrical current x in a simple circuit containing a resistance R, a self-inductance L, and a capacitance C when a voltage $E(t)$ is applied is determined by a simple differential equation (cf. Figure 1A). The drops in voltage across the ends of the resistance and the inductance are, respectively, Rx and $L(dx/dt)$. The drop across the capacitance is q/C, in which q is the charge on the condenser. The current is given by $x = dq/dt$, and the voltage $E(t)$ must equal the sum of the voltage drops. The charge is therefore determined by the differential equation obtained by substituting $x = dq/dt$ (see 24).

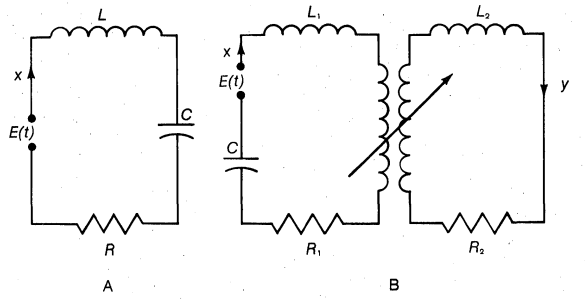


Figure 1: *Electrical circuits.*
(A) Simple circuit containing a resistance R, a self-inductance L, and a capacitance C. (B) Coupled circuits in which the currents x and y are determined by a set of differential equations.

Similarly, for the coupled circuits shown in Figure 1B, the currents x and y are determined by a pair of similar linear second-order differential equations (see 25).

Self-oscillatory problems. Certain electrical circuits, such as feedback circuits controlled by thermionic values, have the property that a source of power increases with the amplitude of the oscillation. To exhibit the main features of the behaviour of such systems, the 20th-century Dutch mathematician Balthasar van der Pol took as his mathematical model a second-order equation (see 26), which is now called van der Pol's equation. This equation is typical of the models describing a series of self-oscillatory problems. For example, the equation for the oscillations of a valve generator with a cubic valve characteristic can be reduced to van der Pol's equation; van der Pol himself used this equation in the theory of the oscillations in a symmetric multivibrator, in which there is an inductance.

The equations of van der Pol and Laplace

Potential theory. In many branches of physics, such as electrostatics, magnetostatics, irrotational motion of perfect fluids, and others, an equation of interest is the partial differential equation named after Pierre-Simon Laplace, an 18th-century French mathematician. This equation (see 27) relates second partial derivatives of the potential u. The equation can be expressed in cylindrical coordinates (ρ, ϕ, z) (see 28). If there are solutions of this equation in the form of a product of $\exp\{-\xi z\}$, $\cos(\nu\phi)$ and some function of ρ, $R(\rho)$ (see 29), then R must be a solution of a second-order linear ordinary differential equation with variable coefficients so that $R(\rho) = \omega(\xi\rho)$ in which $\omega(t)$ is a solution of a second-order linear equation with variable coefficients (see 30) known as Bessel's equation of order ν, so named after Friedrich Wilhelm Bessel, a 19th-century German mathematician.

Similarly, Laplace's equation can be expressed in spherical coordinates (see 31). For axisymmetric solutions of the form $v(r, \theta) = r^n\Theta(\theta)$—*i.e.*, solutions that do not depend on the azimuthal angle ϕ—Θ must be a solution of another linear ordinary differential equation with variable coefficients (see 32). Changing the independent variable in this equation from θ to $\mu = \cos\theta$ gives solutions of the desired form if $\Theta(\theta) = w(\cos\theta)$, in which $w(\mu)$ is any solution of the ordinary differential equation (see 32) called Legendre's equation of order n, named after Adrien-Marie Legendre, an 18th–19th-century French mathematician.

Stellar structure. An interesting ordinary differential equation arises in the discussion of the theory of stellar structure. If the gravitational equilibrium of a mass of gas is considered it is found that its pressure p varies with the distance r from the centre of the mass according to a second-order differential equation involving G, the gravitational constant, and ρ, the density (see 33). The density is related to the pressure through the physical law $\log p = \log k + (1 + 1/\mu)\log\rho$ in which K and μ are constants. If $\log\rho = \log\rho_c + \mu\log\theta$, in which ρ_c is the central density, and a new independent variable $\xi = kr$ is introduced, in which k is defined in terms of ρ_c and constants (see 34), the Lane-Emden equation and initial conditions for it are obtained (see 35). Unless μ is 0 or 1, the equation is nonlinear.

The harmonic oscillator in wave mechanics. One problem in wave mechanics is that of determining the wave function ψ of a particle of mass m moving on the x-axis under the action of a force $-kx$. The wave function ψ for a particle has the physical significance that $|\psi(x, y, z)|^2\, d\tau$ is the probability of finding the particle in a small element of volume $d\tau$ centred at the point with coordinates (x, y, z). In this problem the equation describing the particle is a particular case of a more general second-order differential equation named after a 20th-century Austrian physicist, Erwin Schrödinger (see 36), involving E, the total energy of the particle, and h, Planck's constant. The problem here is not so much to solve this ordinary differential equation as to find the possible values of E that ensure $|\psi| \to 0$ as $|x| \to \infty$. (For an understanding of convergence to the real number "infinity," see ANALYSIS, REAL.) If the variables are changed to $x = (h^2/4\pi^2mk)^{\frac{1}{4}}z$, $a = \frac{1}{4}$, and $\lambda = (4\pi E/h)(m/k)^{\frac{1}{2}}$, an equation relating $\dfrac{d^2\psi}{dz^2}$ and $(\lambda - z^2)\psi$ (see 37) is obtained. The problem is to find the set of allowed values of λ ensuring the existence of solutions such that $|\psi| \to 0$ as $|z| \to \infty$. Such a problem is called a Sturm-Liouville problem (named after the 19th-century French mathematicians Charles-François Sturm and Joseph Liouville).

(26) $$\frac{d^2x}{dt^2} - \mu(1-x^2)\frac{dx}{dt} + w^2x = 0$$

(27) $$\frac{\partial^2u}{\partial x^2} + \frac{\partial^2u}{\partial y^2} + \frac{\partial^2u}{\partial z^2} = 0$$

(28) $$\frac{\partial^2u}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial u}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2u}{\partial\phi^2} + \frac{\partial^2u}{\partial z^2} = 0$$

(29) $$u(\rho, \phi, z) = e^{-\xi z}\cos(\nu\phi)\, R(\rho)$$

(30) $$\begin{cases} \dfrac{d^2R}{d\rho^2} + \dfrac{1}{\rho}\dfrac{dR}{d\rho} + \left(\xi^2 - \dfrac{\nu^2}{\rho^2}\right)R = 0 \\ \dfrac{d^2w}{dt^2} + \dfrac{1}{t}\dfrac{dw}{dt} + \left(1 - \dfrac{\nu^2}{t^2}\right)w = 0 \end{cases}$$

(31) $$\frac{\partial^2v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial v}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2v}{\partial\phi^2} = 0$$

(32) $$\begin{cases} \dfrac{1}{\sin\theta}\left[\dfrac{d}{d\theta}\sin\theta\dfrac{d\Theta}{d\theta}\right] + n(n+1)\Theta = 0 \\ (1-\mu^2)\dfrac{d^2w}{d\mu^2} - 2\mu\dfrac{dw}{d\mu} + n(n+1)w = 0 \end{cases}$$

(33) $$\frac{1}{r^2}\frac{d}{dr}\frac{r^2}{\rho}\frac{d\rho}{dr} + 4\pi G\rho = 0$$

(34) $$k^2 = 4\pi G\rho_c^{1-1/\mu}/[K(1+\mu)]$$

(35) $$\begin{cases} \dfrac{d^2\theta}{d\xi^2} + \dfrac{2}{\xi}\cdot\dfrac{d\theta}{d\xi} + \theta^\mu = 0 \\ \theta = 1, \quad \dfrac{d\theta}{d\xi} = 0, \quad \xi = 0 \end{cases}$$

(36) $$\frac{d^2\psi}{dx^2} + \frac{8\pi^2m}{h^2}(E - \tfrac{1}{2}kx^2)\psi = 0$$

(37) $$\frac{d^2\psi}{dz^2} + (\lambda - z^2)\psi = 0$$

(38) $$m\frac{d^2x}{dt^2} = f$$

Control theory. A different kind of problem involving differential equations arises in optimal control theory. In control theory attention is first of all focussed on a process; that is, some action changing with time and usually described by a differential equation or a system of differential equations. In parallel with the concept of process, controls are considered for influencing the particular process being discussed, and the objective is the result achieved by the process through a properly applied control strategy. If with respect to some criterion of performance, the strategy that is best is sought, the problem is called an optimal control problem.

Typical control theory problem

A simple problem of this kind is that of driving a railroad locomotive from one station to another on the assumption that the same force is available for both starting and stopping; *e.g.*, two rocket motors pointed in opposite directions and each capable of a maximum thrust P. If the locomotive is treated as a particle of mass m and the decreases in m caused by loss of fuel and all resistance phenomena are neglected, the product of m and the second derivative of x equals f (see 38), in which x is the distance travelled by the locomotive from its starting point and f is the force applied by the rocket motors. The control is defined by the inequality $|f| \leqslant P$. As an objective, one of the following conditions could be taken:

(i) The journey must be accomplished in the minimum time.

(ii) To avoid excessive wear on components the journey must be accomplished with the minimum expenditure of energy.

(iii) The amount of rocket fuel used should be kept to a minimum.

Once one of these criteria has been decided upon, the optimal control problem consists of designing a scheme for applying f in such a way as to achieve the desired result; *i.e.*, of selecting from the set of solutions of the differential equations corresponding to all permissible forms of f the one that is best.

Chemical kinetics. Simple chemical reactions can be described by first-order ordinary differential equations. A simple first-order reaction such as the decomposition of nitrogen pentoxide may be considered first. If c is the concentration of nitrogen pentoxide at time t, then the derivative of c is equal to the product of a constant k, which is called the rate constant for the reaction, and $-c$ (see 39).

Similarly, in a second-order reaction of the type $A + B \rightarrow A' + B'$ in which a and b are the initial concentrations of A and B, respectively, and x is the fraction of the concentration of the reaction that has been consumed at time t, the derivative of x is proportional to the product of $(a - x)$ and $(b - x)$ (see 40).

Reversible reactions can also be described by differential equations. For example, consider a simple reaction of the type $A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} B$, such as the interconversion between α-d-glucose and β-d-glucose in which a and b are the initial concentrations of A and B and $a - x$ and $a + x$ their values at time t. (If $\frac{dx}{dt} = 0$ the reaction is at equilibrium.) The derivative of x is proportional to the difference of $k_1(a - x)$ and $k_2(b + x)$, in which k_1 and k_2 are constants (see 41).

A catalytic reaction is one in which some substance, known as a catalyst, changes the reaction rate without itself suffering any permanent change. In an autocatalytic reaction a substance A is transformed to another substance B, which then acts as a catalyst for the reaction. If x is the concentration of B at time t and N is its final value, then such a process can be described by a differential equation in which dx/dt is proportional to the product of x and $(N - x)$ (see 42).

Growth of populations. To illustrate how ordinary differential equations arise in biology, an equation arising in a simple study of the growth of populations is considered. $p(t)$ denotes the number of inhabitants of a given area at time t, and it is assumed that in the time interval t to $t + h$

(39) $$\frac{dc}{dt} = -kc$$

(40) $$\frac{dx}{dt} = k(a-x)(b-x)$$

(41) $$\frac{dx}{dt} = k_1(a-x) - k_2(b+x)$$

(42) $$\frac{dx}{dt} = kx(N-x)$$

(43) $$\frac{dp}{dt} = (N-M)p + (I-E)$$

(44) $$\frac{dp}{dt} = \varepsilon p - kp^2 + (I-E)$$

(45) $$\frac{dp}{dt} = p(\varepsilon - kp)$$

(46) $$\begin{cases} \dfrac{dC_A}{dt} = \dfrac{1}{V_A}[a_1C_T - a_2C_A - a_3 + \\ \qquad + (b_1C_T - b_2)(C_T - C_A)] \\ \dfrac{dC_T}{dt} = \dfrac{1}{V_T}[R - a_1C_T + a_2C_A + a_3] \end{cases}$$

(47) $$\begin{cases} \dfrac{dQ_1}{dt} = k_2Q_2 - (k_1 + k_3)Q_1 \\ \dfrac{dQ_2}{dt} = k_1Q_1 - k_2Q_2 \end{cases}$$

(48) $$f\{x, \phi(x), \phi'(x), \cdots, \phi^{(n)}(x)\} \equiv 0, \qquad a \leqslant x \leqslant b$$

(i) the number of individuals born is Nph,
(ii) the number of individuals dying is Mph,
(iii) the number of individuals entering the area is Ih,
(iv) the number of individuals leaving the area is Eh,

in which N, M, I, and E will, in general, be functions of both p and t. The function $p(t)$ is then determined by the fact that its first derivative is proportional to the obvious combination of these quantities (see 43).

In one population model it is assumed that I and E are constants and that $N = n - \nu p$, $M = m + \mu p$ in which m, n, μ, and ν are constants. If these forms are substituted into the above population equation and $n - m = \varepsilon$ and $\mu + \nu = k$, a nonlinear differential equation known as Verhulst's equation is obtained (see 44), the constant ε being called the coefficient of increase and k being called the limiting coefficient.

If the population is isolated (*i.e.*, if there is neither immigration nor emigration), $I = E = 0$, and the equation reduces to a simpler form in which dp/dt equals the product of p and $\varepsilon - kp$ (see 45).

Physiology. *Regulation of carbon dioxide in the body.* An example of the occurrence of a pair of ordinary simultaneous equations in physiology is provided by a simple model of the regulation of carbon dioxide (CO_2) in the human body. In this model it is assumed that the carbon dioxide is distributed between two compartments corresponding to the lungs and the remaining tissue and that respiration is a function of CO_2 only. With further simplifying assumptions it can be shown that during respiration C_A, the alveolar (lung) CO_2 concentration, and C_T, the concentration of CO_2 in the tissue, are determined by the pair of simultaneous equations that relate $\frac{dC_A}{dt}$ and $\frac{dC_T}{dt}$ each to C_A and C_T (see 46), in which V_A and V_T are the respective volumes of the alveolar

$$(49)\quad \begin{cases} x^3+y^3+3xy=1 \\ (x+y^2)y'+x^2+y=0 \end{cases}$$

$$(50)\quad (y')^2+xy'-y=0$$

$$(51)\quad y=cx+c^2$$

$$(52)\quad y=x+1$$

$$(53)\quad y(x)=-\tfrac{1}{4}x^2$$

$$(54)\quad 2(y')^3=3y$$

$$(55)\quad 9y^2=4(x-c)^3$$

$$(56)\quad \frac{d^2y}{dx^2}+\omega^2 y=0$$

$$(57)\quad y(x)=c_1\cos(\omega x)+c_2\sin(\omega x)$$

$$(58)\quad y(a)=y_0, \qquad y'(a)=v_0$$

and tissue compartments; C_I is the concentration of inspired carbon dioxide; a_1, a_2, a_3, b_1, b_2, are constants; and R is the rate at which carbon dioxide is formed by metabolism.

Distribution of creatinine in the human body. Another example taken from physiology arises in the analysis of a two-compartment model for the distribution of creatinine in the human body. The first compartment is the blood plasma; the second is not clearly identified, but there is evidence that such a compartment exists. If Q_1, Q_2 denote, respectively, the quantities of creatinine in the first and second compartments, their variation is determined by a pair of linear first-order differential equations relating dQ_1/dt and dQ_2/dt to Q_1 and Q_2 (see 47), with rate constants k_1, k_2, and k_3.

TYPES OF SOLUTIONS

A relation $y=\phi(x)$ is said to be a solution or integral of the differential equation $f(x, y, y', \cdots, y^{(n)})=0$ in the range $a \leqslant x \leqslant b$ if, when substituted into the equation, it gives a result that is identically zero in that range (see 48). It is frequently difficult or undesirable to express a solution y explicitly as a function of x, but instead to have an implicit relation $F(x, y)=0$ between the solution y and the independent variable x. Such a relation is a solution if, when solved explicitly for y in terms of x, it yields a solution in the way described above. For example, if both sides of a polynomial equation in x and y are differentiated with respect to x, then the function $y(x)$ determined by this implicit relation will be a solution of the differential equation thereby obtained (see 49). The simplest of all ordinary differential equations is $y'=g(x)$, in which g is an elementary function. An example is if $g(x)=2x$. The problem of solving this differential equation is equivalent to finding a function $y(x)$ the derivative of which is $2x$; this leads to the solution $y=x^2+c$ in which c is an arbitrary constant. The process of integration has led to the occurrence of an arbitrary constant in the solution. It can be intuitively seen (and it can be proved rigorously) that finding the solution of a differential equation of the nth order will somehow be equivalent to performing n integrations and that the final solution $y(x)$ will contain n arbitrary constants of integration.

Use of arbitrary constants

Any solution of an nth order equation of the form $y=\phi(x, c_1, c_2, \cdots, c_n)$ in which $c_1, \cdots, c_n$ are arbitrary constants is called a general solution of the equation. Any solution that may be obtained from the general solution of an equation by assigning particular values to the constants is called a particular solution of that equation. It sometimes happens that a nonlinear differential equation has a solution that cannot be obtained by assigning specific values to the arbitrary constants in the general solution. Such a solution is called a singular solution of the differential equation.

For example, a particular nonlinear differential equation (see 50) has a general solution that is linear in the independent variable (see 51) with an arbitrary constant c. A particular solution may be obtained from (51) by taking $c=1$ (see 52). On the other hand, a function proportional to x^2 (see 53) also satisfies the differential equation. Because this solution cannot be obtained from the general solution by assigning a particular value to c, it is a singular solution of the differential equation (50). In this instance the graphs of the functions defined by (51) are straight lines and the graph of the function defined by (53) is a parabola that is the envelope of that family of straight lines (cf. Figure 2).

Similarly, a nonlinear equation with the cube of the derivative proportional to y (see 54) has a general solution with y^2 proportional to $(x-c)^3$ (see 55), and the singular solution $y=0$. Here the general solution is represented by a family of semicubical parabolas, and the singular solution is the line that passes through the cusp of each member of the family (cf. Figure 3). Such a line is called a cusp locus.

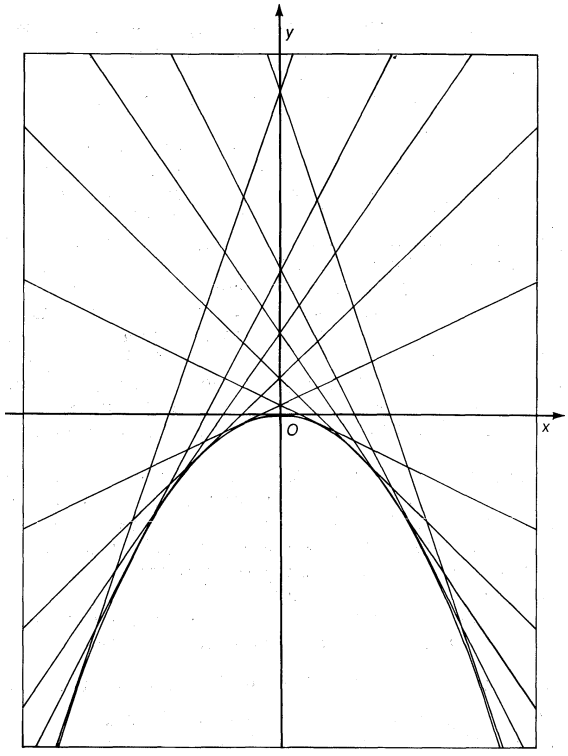


Figure 2: Graphs of functions defined by straight lines and a parabola that is the *envelope* of that family of straight lines (see text).

Boundary-value problems

It has been pointed out above that the solution of problems in physics or engineering is equivalent to that of finding solutions of differential equations satisfying certain additional conditions. To illustrate the kind of problem arising in this way, the following equation may be considered: with the product of a constant, ω^2, and the function y being added to the second derivative of y to give a result of zero (see 56). Its general solution is found to be the product of a constant, c_1, and a cosine, added to the product of another constant, c_2, and a sine (see 57). The solution corresponds to the conditions of y and its derivative at some point a being set equal to y_0 and v_0, respectively (see 58), in which a, y_0, v_0 are prescribed real numbers, is obtained by choosing c_1 and c_2 to satisfy the linear algebraic equations obtained by setting the general solution at a equal to y_0 while its derivative is set

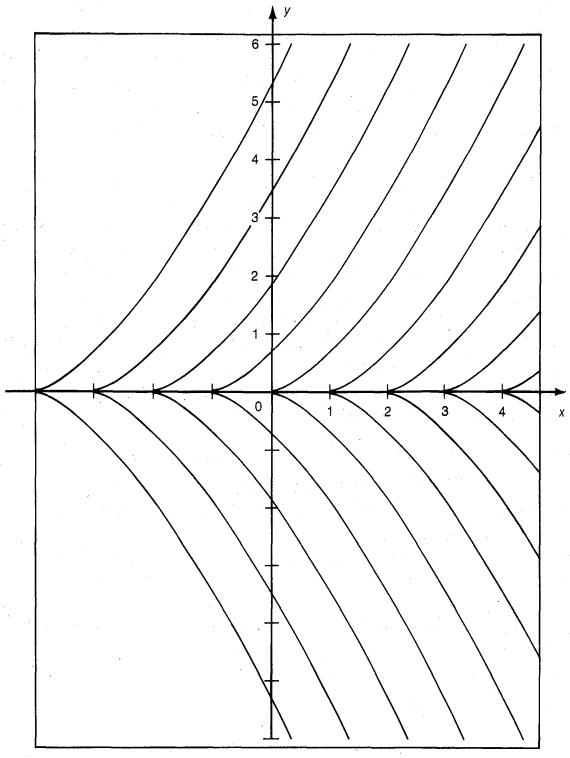


Figure 3: A family of semicubical parabolas, representing a general solution, and a singular solution, which is the line that passes through the cusp of each member of the family (see text).

equal to v_0 (see 59). The solution corresponding to the conditions of y at two different points a and b being set equal to constants y_0 and y_1, respectively (see 60), is obtained by choosing c_1 and c_2 to satisfy the linear equations obtained by setting the general solution at a and b equal to y_0 and y_1, respectively (see 61).

Problems of this kind are called boundary value problems; the problem posed by equations (56) and (58) is called a one-point boundary value problem or an initial value problem, while that posed by equations (56) and (60) is called a two-point boundary value problem.

Another kind of problem can be exemplified by reference to equation (56). It may be stated as a question: Do there exist real values of ω such that the equation (56) has nonzero solutions satisfying the boundary conditions $y(0) = y(a) = 0$? That there are real values of ω follows immediately from the fact that, if n is an integer, then the function $y = \sin(n\pi x/a)$ satisfies these boundary conditions and the differential equation considered above, with $\omega = n\pi/a$ (see 62). Hence, $\pm n\pi/a$, $n = 1, 2, 3, \cdots$, are possible values of ω. A problem of this kind is called an eigenvalue problem or a Sturm-Liouville problem.

EXISTENCE AND UNIQUENESS OF SOLUTIONS

From the geometrical interpretation of the derivative of a function, it is tempting to assume that the initial value problem consisting of a differential equation and the value y_0 of the function y at some point a (see 63) has a unique solution through each point (a, y_0) of the xy-plane. It is a simple matter, however, to construct differential equations that do not have a solution at all at a specific point and others that have more than one solution—often an infinite number—at a point. For this reason it is necessary to study sufficient conditions for the existence and uniqueness of solutions of the initial value problem (see 63). The proof of the classic theorem in this area depends on the fact that any solution of the initial value problem (63) satisfies the integral equation given by formally integrating both sides of the differential equation, the constant of integration being y_0 (see 64). Conversely, any solution of this integral equation is a solution of the initial value problem (63).

The Lipschitz condition

To state the fundamental result, the following concept is required: a function $f(x, y)$ is said to satisfy a Lipschitz condition with respect to y in a region D of the xy-plane if there exists a constant K such that the magnitude of the difference between $f(x, y_1)$ and $f(x, y_2)$ is less than or equal to K times the magnitude of the difference between y_1 and y_2 (see 65), for every pair of points (x, y_1), (x, y_2) belonging to D. The condition is named after a 19th-century German mathematician, Rudolf Lipschitz.

The basic theorem states: If $f(x, y)$ is continuous and satisfies a Lipschitz condition with respect to y in some region D of the xy-plane and if (a, y_0) is any point in this region, then the initial value problem (63) has a unique solution.

It turns out that continuity of f is sufficient to guarantee existence and that it is the Lipschitz condition on f that ensures uniqueness.

By generalizing the Lipschitz condition (see 65) to the case in which both f and y are n-vectors and D is a region of the xy-space (of dimension $n + 1$), the theorem can be established for a vector differential equation (see 63). Now any differential equation of order n can be written as a first-order equation for an n-vector. For example, the initial value problem posed by equations (56)

(59) $$\begin{cases} c_1\cos(\omega a) + c_2\sin(\omega a) = b \\ -c_1\sin(\omega a) + c_2\cos(\omega a) = v_0 \end{cases}$$

(60) $$y(a) = y_0, \qquad y(b) = y_1$$

(61) $$\begin{cases} c_1\cos(\omega a) + c_2\sin(\omega a) = y_0 \\ c_1\cos(\omega b) + c_2\sin(\omega b) = y_1 \end{cases}$$

(62) $$\frac{d^2y}{dx^2} + \frac{n^2\pi^2}{a^2}y = 0$$

(63) $$\frac{dy}{dx} = f(x, y), \qquad y(a) = y_0$$

(64) $$y(x) = y_0 + \int_a^x f\{t, y(t)\}\, dt$$

(65) $$|f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2|$$

(66) $$\frac{dz}{dx} = Az, \qquad z(a) = z_0$$

(67) $$z = \begin{bmatrix} y \\ y' \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}, \qquad z_0 = \begin{bmatrix} y_0 \\ v_0 \end{bmatrix}$$

(68) $$n = n_0e^{-\lambda t}$$

(69) $$x = c_1 + c_2t, \qquad y = c_3 + c_4t - \tfrac{1}{2}gt^2$$

(70) $$y = a\cosh(x/a)$$

(71) $$\sum_{r=0}^{\infty} \frac{(-1)^r(\frac{1}{2}x)^{2r+\nu}}{r!(\nu+1)_r}$$

(72) $$\sum_{r=0}^{\infty} \frac{(\frac{1}{2}n+\frac{1}{2})_r(\frac{1}{2}n+1)_r}{r!(n+\frac{3}{2})}x^{-2r-n-1}$$

(73) $$\sum_{r=0}^{[\frac{1}{2}n]} \frac{(\frac{1}{2}-\frac{1}{2}n)_r(-\frac{1}{2}n)_r}{r!(\frac{1}{2}-n)_r}x^{n-2r}$$

(74) $$\sum_{r=0}^{\infty} \frac{(-n)_r(n+1)_r}{r!r!}\left(\frac{1-x}{2}\right)^r$$

and (58) can be written in the form of an initial value problem for a first-order equation (see 66), but in this case z, A, z_0 denote matrices (see 67). The existence theorem for an nth-order equation can therefore be deduced from that for a first-order equation in which the independent variable is an n-vector.

FORMS OF SOLUTIONS

Closed form solutions. In many cases it is possible to express the solution of a differential equation by a simple formula. For example, if the derivative of n is proportional to $-n$, then the general solution is an arbitrary constant n_0 times a negative exponential (see 68). The pair of equations describing motion in a gravitational field has as the general solution for x a linear function and for y a quadratic (see 69), with arbitrary constants c_1, c_2, c_3, and c_4. Similarly, the solution of the equation for a catenary satisfying the initial conditions $y = a$, $dy/dx = 0$ at $x = 0$ is a hyperbolic cosine (see 70).

Series solutions. Often when it is not possible to find a closed form solution, it is a simple matter to find a solution in the form of an infinite series. This form of solution is particularly appropriate to the solution of second-order linear equations. For example, Bessel's equation has a series solution in powers of $x/2$ (see 71), in which the symbol $(\nu + 1)_r$ denotes the product $(\nu + 1)(\nu + 2) \cdots (\nu + r)$. Similarly, Legendre's equation has a solution in the form of descending powers of x that is valid only if $|x| > 1$ (see 72). The same equation has a power series solution valid if $|x| < 1$ (see 73), and a series solution in powers of $\frac{1-x}{2}$ that is valid for $-1 < x < 3$ (see 74). It will be noted that in the case in which n is a positive integer, each of the last two series reduces to a polynomial of degree n in x. Power series

Integral forms. In a similar way, the solution of a differential equation can be expressed in the form of an integral that cannot be expressed in terms of elementary functions. For example, the solution of equation (3) satisfying the initial conditions $\theta = \alpha$, $d\theta/dt = 0$ when $t = 0$ may be expressed by an equation involving an integral of a linear combination of a cosine to the $-\frac{1}{2}$ power (see 75). The integral on the left side of this equation cannot be expressed in terms of elementary functions. Similarly, Bessel's equation of order 0 has the solution as the integral of an expression involving exponentials and square roots (see 76).

Approximations. In many cases it is not possible to derive the solution of a differential equation in any of the above forms and recourse has to be made to approximate methods.

Iteration method. In dealing with initial value problems one of the simplest ways of obtaining an approximate solution is to solve the integral equation discussed earlier (see 64) by iteration; *i.e.*, to construct a sequence of approximations $y_1(x), y_2(x), \cdots, y_n(x)$ by the scheme of substituting an approximation for y into the integral equation and obtaining as the result a better approximation and repeating the process (see 77). It can be proved that, if the conditions of the existence-uniqueness theorem are satisfied, this sequence will converge to the correct solution. By taking the nth iterate, an approximate solution of the initial value problem is obtained.

For example, for the initial value problem for a certain first-order nonlinear differential equation (see 78), a sequence of approximations that are partial sums of a power series (see 79) are given.

Taylor series method. In theory this method is applicable to an equation of any order. As an example, the initial value problem for a second-order differential equation (see 80) is considered. First of all $y''(a) = f(a, y_0, y_1)$ is calculated. The given equation is differentiated with respect to x (see 81). This is evaluated at a to yield the value of $y'''(a)$. By successive differentiation $y^n(a)$ can be found for $n = 3, 4, 5, \cdots$ and an approximation to the value of $y(a + h)$ can be found from the truncated Taylor series (see 82). The process is then repeated at the point $x = a + h$, and so on.

The Runge-Kutta methods. To indicate the methods due to the German mathematicians Carl David Tolmé Runge (1895) and William Martin Kutta (1901), the initial value problem (63) is considered. In these methods the Taylor expansion is used indirectly: $y(a + h)$ is calculated (see 83) in terms of $y(a)$, the function $f(x, y)$, and constants $a_0, \cdots, a_p, b_0, \cdots, b_p, c_0, \cdots, c_p$, which are chosen in such a way that, when the expression is expanded for $y(a + h)$ in ascending powers of h, the coefficients agree with those of the Taylor expansion for $y(a + h)$.

For example, if $p = 1$ is taken, a formula involving two values of $f(x, y)$ is obtained (see 84). In this formula

(75) $$\int_\theta^\alpha \frac{d\phi}{\sqrt{(\cos\phi - \cos\alpha)}} = \frac{2g}{a}t$$

(76) $$\int_1^\infty \frac{e^{-zt}\,dt}{\sqrt{(t^2-1)}}$$

(77) $$y_{r+1}(x) = y_0 + \int_a^x f\{t, y_r(t)\}\,dt, \qquad y_0(x) = y_0$$

(78) $$\frac{dy}{dx} = 1 + x^2y^2, \qquad y(0) = 0$$

(79) $$x, \qquad x + \frac{1}{5}x^5, \qquad x + \frac{1}{5}x^5 + \frac{2}{45}x^9 + \frac{1}{325}x^{13}, \cdots$$

(80) $$y'' = f(x, y, y'), \qquad y(a) = y_0, \qquad y'(a) = y_1$$

(81) $$y''' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y''$$

(82) $$y(a+h) = y_0 + y_1 + \tfrac{1}{2}y''(a)h + \cdots + \frac{1}{n!}y^{(n)}(a)h^n$$

(83) $$y(a+h) = y(a) + h\sum_{k=0}^{p} a_k f(a + b_k h, y_0 + c_k h)$$

(84) $$y(a+h) = y(a) + (1-c)hf(a, y_0) + chf\left(a + \frac{h}{2c}, y_0 + \frac{k_0}{2c}\right)$$

(85) $$y_{r+1} = y_r + (1-c)k_r + cm_r$$

(86) $$\frac{dy}{dx} = f(x, y)$$

(87) $$\frac{dy}{dx} = f(x)g(y)$$

(88) $$y\frac{dy}{dx} = xy + x$$

(89) $$\int \frac{dy}{g(y)} = \int f(x)\,dx + c$$

(90) $$\int \frac{y\,dy}{y+1} = \int x\,dx + c$$

(91) $$\frac{dy}{dx} = g\left(\frac{y}{x}\right)$$

(92) $$x\frac{dv}{dx} + v = g(v)$$

$k_0 = hf(a, y_0)$, and c is any nonzero real number. The numerical solution of the initial value problem (63) can be generated by repeated applications of this relationship. If $x_r = a + rh$, $y_r = y(a + rh)$, $k_r = hf(x_r, y_r)$, and $m_r = hf(x_r + h/2c, y_r + k_r/2c)$, $(c \neq 0)$, then y_{r+1} can be calculated from y_r, k_r, m_r, and c (see 85). This is known as the Runge-Kutta second-order process. More elaborate schemes can, of course, be derived by taking a higher value of p.

FIRST-ORDER EQUATIONS

When first-order equations of the first degree are solved for the derivative, they are of a form in which the derivative y' equals $f(x, y)$, an explicit function of x and y (see 86). There are certain distinguishable types of such equations that may be easily solved and that are now discussed briefly.

Equations with separable variables. If the derivative can be expressed as the product of $f(x)$ and $g(y)$ (see 87), in which $f(x)$ is a function of x alone and $g(y)$ is a function of y alone, it is said to have its variables separable. For example, the nonlinear equation in which the product of y and its derivative equals the sum of xy and x (see 88) has its variables separable. The general solution of such an equation is obtained by moving the expressions involving x to one side of the equation and those involving y to the other, and then integrating both sides (see 89), involving c, as arbitrary constant. For example, the equation just considered (see 88) has general solution in terms of an integral of y divided by $y + 1$ and an integral of x (see 90).

These integrals can be evaluated, giving the result that $y - \log(y + 1) = x^2/2 + c$.

Homogeneous equations. A differential equation in which the derivative dy/dx equals $g(y/x)$, in which g is a function of y/x alone (see 91), is said to be homogeneous. Such an equation is reduced to separable form (see 92) by the substitution $y = xv$. For example, the equation in which dy/dx equals $2y^2 + x^2$ divided by xy (see 93) is homogeneous; the substitution $y = xv$ reduces it to the separable form as a differential equation with v as the dependent variable (see 94). It is possible to obtain the solution of this differential equation by certain integrations (see 95), in which k is an arbitrary constant. Performing the integrations, the general solution is obtained in terms of logarithms (see 96), in which c is an arbitrary constant; $(2k = \log c)$—*i.e.*, the sum of v^2 and one equals cx^2 (see 97). Returning to the original variables, the sum of x^2 and y^2 equals cx^4 (see 98), which is the general solution of equation (93).

Exact equations. If there exists a function $\psi(x, y)$ such that the partial derivative of ψ with respect to x equals $P(x, y)$, and the partial derivative of ψ with respect to y equals $Q(x, y)$ (see 99), the differential equation formed by the sum of $P(x, y)$ and the product of $Q(x, y)$ and y' being equal to zero (see 100) is said to be exact. The reason for this is that the function $y(x)$ defined by $\psi(x, y)$ being equal to an arbitrary constant c (see 101) has a derivative y' that satisfies the equation obtained by differentiating $\psi(x, y)$ with respect to x by the chain rule (see 102).

Hence, because of these relations (see 99), the differential equation (100) follows. It is obvious from equations (99) that a necessary condition for the equation (100) to be exact is that the partial derivative of $P(x, y)$ with respect to y equals the partial derivative $Q(x, y)$ with respect to x (see 103); it can also be shown that this condition is sufficient. For example, a nonlinear differential equation with exponential coefficients (see 104) is exact because the partial derivative with respect to x of the coefficient of y' equals the partial derivative with respect to y of the rest of the equation (see 105). It has the general solution in implicit form with an arbitrary constant (see 106). Instead of saying that the equation (100) is exact, $P(x, y)dx + Q(x, y)dy$ is sometimes termed an exact differential.

Integrating factors. Although a differential equation may not be exact in the form in which it is encountered, it has been proved that it may be made exact by multipli-

(93) $$\frac{dy}{dx} = \frac{2y^2 + x^2}{xy} = \frac{2(y/x)^2 + 1}{(y/x)}$$

(94) $$x\frac{dv}{dx} + v = \frac{2v^2 + 1}{v}$$

(95) $$\int \frac{v\,dv}{v^2 + 1} = \int \frac{dx}{x} + k$$

(96) $$\log(v^2 + 1) = 2\log x + \log c$$

(97) $$v^2 + 1 = cx^2$$

(98) $$x^2 + y^2 = cx^4$$

(99) $$\frac{\partial \psi}{\partial x} = P(x, y), \qquad \frac{\partial \psi}{\partial y} = Q(x, y)$$

(100) $$P(x, y) + Q(x, y)y' = 0$$

(101) $$\psi(x, y) = c$$

(102) $$\frac{\partial \psi}{\partial x} + \frac{\partial \psi}{\partial y}y' = 0$$

(103) $$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

(104) $$2e^x yy' + y^2e^x - 1 = 0$$

(105) $$\frac{\partial}{\partial y}(y^2e^x - 1) = \frac{\partial}{\partial x}(2e^x y)$$

(106) $$y^2e^x - x = c$$

(107) $$2xy' + 3y = 0$$

(108) $$2yy'x^3 + y^2 \cdot 3x^2 = 0$$

(109) $$\frac{\partial}{\partial y}(\mu P) = \frac{\partial}{\partial x}(\mu Q)$$

(110) $$\frac{1}{Q}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right) = g(x)$$

(111) $$\frac{\mu'(x)}{\mu(x)} = g(x)$$

(112) $$\mu(x) = \exp\{\int g(x)\,dx\}$$

(113) $$y' + a(x)y = b(x)$$

(114) $$\mu(x) = \exp\{\int a(x)\,dx\}$$

(115) $$\begin{cases} xy' + (x^2 + 1)y = x \\ a(x) = x + x^{-1} \end{cases}$$

(116) $$\begin{cases} \mu = \exp[\int (x + x^{-1})dx] \\ \quad = \exp[\frac{1}{2}x^2 + \log x] = xe^{\frac{1}{2}x^2} \end{cases}$$

(117) $$\frac{d}{dx}[xe^{\frac{1}{2}x^2}y] = xe^{\frac{1}{2}x^2}$$

(118) $$y = x^{-1} + cx^{-1}e^{-\frac{1}{2}x^2}$$

(119) $y' + a(x)y = b(x)y^n$

(120) $z' - (n-1)\,a(x)z = -(n-1)\,b(x)$

(121) $z = y^{1-n}$

(122) $y' = a(x)y^2 + b(x)y + c(x)$

(123) $z' + (2af + b)z + a = 0$

(124) $y = c + \int \frac{p\,dp}{(1+p^2)^{\frac{3}{2}}} = c - (1+p^2)^{-\frac{1}{2}}$

(125) $x^2 + (y-c)^2 = 1$

(126) $y = xy' + f(y')$

(127) $[x + f'(p)]\frac{dp}{dx} = 0$

(128) $y = cx + f(c)$

(129) $x = -f'(p), \qquad y = f(p) - pf'(p)$

(130) $P_n(x)y^{(n)} + P_{n-1}(x)y^{(n-1)} + \cdots + P_1(x)y' + P_0(x)y = f(x)$

(131) $P_n(x)D^n + P_{n-1}(x)D^{n-1} + \cdots + P_1(x)D + P_0(x)$

(132) $a_1y_1(x) + a_2y_2(x) + \cdots + a_ny_n(x) = 0$

(133) $y(x) = c_1y_1(x) + \cdots + c_ny_n(x) + Y(x)$

(134) $$W(y_1, \cdots, y_n) = \begin{vmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{vmatrix}$$

(135) $\Pi_n(D)y = f(x)$

(136) $\Pi_n(t) = p_nt^n + p_{n-1}t^{n-1} + \cdots + p_1t + p_0, \qquad (p_n \neq 0)$

(137) $c_1e^{\lambda_1x} + c_2e^{\lambda_2x} + \cdots + c_ne^{\lambda_nx}$

cation by a suitable factor. Such a multiplier is called an integrating factor. For example, the equation formed by the sum of $2xy'$ and $3y$ being equal to zero (see 107) is not exact, but multiplying throughout by x^2y gives an equation (see 108) that is exact and has general solution $y^2x^3 = c$.

If both sides of equation (100) are multiplied by $\mu(x, y)$ and the exactness condition (103) is used, then μ will be an integrating factor of the equation if, and only if, it satisfies the condition that the partial derivative of μP with respect to y equals the partial derivative of μQ with respect to x (see 109). For example, the general first-order differential equation being considered (100) will possess an integrating factor $\mu(x)$ if, and only if, the difference of the partial derivative of P with respect to y and that of Q with respect to x, divided by Q, equals $g(x)$, a function of x alone (see 110), and then μ may be taken to be any solution of the equation given by $g(x)$ being equal to $\mu'(x)$ divided by $\mu(x)$ (see 111); *i.e.*, $\mu(x)$ is an exponential (see 112).

Linear equations. If the above procedure is applied to the linear first-order equation with the coefficient of y being $a(x)$ and the inhomogeneous term being $b(x)$ (see 113), it is found to have an integrating factor $\mu(x)$ equal to an exponential with the integral of $a(x)$ in the exponent (see 114).

For example, if the sum of xy' and $(x^2 + 1)y$ equals x (see 115), then $\mu(x)$ is equal to the product of x and an exponential with exponent $x^2/2$ (see 116). The equation can be written as the derivative of the product of $\mu(x)$ and y being equal to $\mu(x)$ (see 117). Integrating and dividing by $xe^{x^2/2}$ yields the general solution in terms of x^{-1}, an exponential, and an arbitrary constant c (see 118).

Bernoulli's equation. Bernoulli's equation, in which the sum of y' and $a(x)y$ equals $b(x)y^n$ (see 119), may be reduced to a linear equation with variable coefficients (see 120) in the new independent variable z. This independent variable, z, is defined as being equal to y to the $(1 - n)$ power (see 121).

Riccati's equation. If any solution $f(x)$ can be found for Riccati's equation, in which y' equals a quadratic function of y, with coefficients $a(x)$, $b(x)$, and $c(x)$ of the quadratic, linear, and constant terms, respectively (see 122), then by the substitution $y = f(x) + z^{-1}$ the problem of finding the general solution of Riccati's equation can be reduced to that of finding the general solution of the first-order linear equation with the coefficient of z being $2af + b$ and the constant term being a (see 123). This applies no matter how trivial $f(x)$ is.

Equations of higher degree. So far the treatment has been limited to equations that are of the first order as well as the first degree. If p is written for y', the general first-order equation will be of the form $F(x, y, p) = 0$ in which F is not linear in p. When an equation of either of the forms $F(x, p) = 0$ or $G(y, p) = 0$ is solvable for p, $p = f(x)$ or $p = g(y)$, integration is immediate.

On the other hand, if $x = \phi(p)$, y can be found by the formula $y = c + \int pdx = c + \int p\phi'(p)dp$; and the general solution may be obtained by eliminating p from this last equation and $x = \phi(p)$. A similar procedure also holds if $y = \psi(p)$.

For example, if $x^2 = p^2(1 - x^2)$, then $x = ap(1 + p^2)^{-\frac{1}{2}}$ so that y can be written as the sum of an arbitrary constant c and the integral of a function of p (see 124). Eliminating p from these two relations gives the general solution, in which the sum of x^2 and $(y - c)^2$ equals 1 (see 125).

Clairaut's equation. The differential equation in which y equals the sum of the product of x and y' and $f(y')$ (see 126), in which f is a prescribed function of y' alone, is called Clairaut's equation. Writing $p = y'$ and differentiating both sides of this equation with respect to x leads to the product of dp/dx and $x + f'(p)$ being equal to zero (see 127), so that either $dp/dx = 0$ or else $x + f'(p) = 0$.

The first alternative yields the general solution (see 128) in which y is a linear function of x, and the second alternative yields the solution obtained by eliminating p from a pair of equations (see 129); if $f(p)$ is not a linear function of p, the result will be a singular solution of the given differential equation.

LINEAR DIFFERENTIAL EQUATIONS

Equations of nth order. A linear differential equation of order n is an equation involving y and its derivatives linearly, the coefficient of y^j being P_j, and the inhomogeneous term being $f(x)$ (see 130), in which $P_n, \cdots, P_0, f$ are given functions of x, and $P_n(x)$ does not vanish identically. The linear differential equation obtained from equation (130) by replacing $f(x)$ by 0 is called the homogeneous equation corresponding to the nonhomogeneous equation (130). For example, Bessel's equation and Legendre's equation are linear equations of the second order and of homogeneous type.

Homogeneous and nonhomogeneous equations

If the differential operator d/dx is denoted by D and if $\mathbf{L}$ is written for the operator formed by the sum of terms of the form P_jD^j (see 131), then equation (130) may be written in the form $\mathbf{L}y = f(x)$ and the corresponding homogeneous equation in the form $\mathbf{L}y = 0$.

Functions $y_1(x), \cdots, y_n(x)$ are said to be linearly independent solutions of the homogeneous equation $\mathbf{L}y = 0$ if $\mathbf{L}y_r(x) = 0$, $r = 1, 2, \cdots, n$, and if the functions are

such that there does not exist a set of constants a_1, a_2, $\cdots$, a_n, not all of which are zero, such that, for all values of x, the sum of the terms $a_j y_j(x)$ is equal to zero (see 132).

The basic result in the theory of linear differential equations is that if y_1, y_2, $\cdots$, y_n are n linearly independent solutions of $\mathbf{L}y = 0$ and if $Y(x)$ is any solution of equation (130), then the general solution of (130) is $Y(x)$ added to a sum of terms of the form $c_j y_j(x)$ (see 133), in which $c_1, \cdots, c_n$ are arbitrary constants.

The general solution of a linear differential equation is therefore the sum of two parts: $c_1 y_1(x) + \cdots + c_n y_n(x)$, which is called the complementary function, and $Y(x)$, which is called a particular integral. To determine whether or not a set of functions is linearly independent, the following criterion is used: a necessary and sufficient condition that $y_1(x), \cdots, y_n(x)$ are linearly independent is that the Wronskian W, defined as the determinant of the matrix formed by the n linearly independent solutions and their first $n-1$ derivatives (see 134), is not identically zero.

Equations with constant coefficients. There is a simple theory of linear equations of the type (130) in which $P_r(x) = p_r$, a constant, $(r = 0, 1, \cdots, n)$. Such equations can be written in the form of an expression Π_n, $(D)y$ set equal to $f(x)$ (see 135), in which $\Pi_n(t)$ denotes the polynomial defined as the sum of terms of the form $p_j t^j$ (see 136).

The form of the complementary function depends on the zeros of the polynomial Π_n. If $\Pi_n(t)$ has n distinct zeros $\lambda_1, \lambda_2, \cdots, \lambda_n$, the complementary function of the equation (135) is in the form of a sum of terms $c_j \exp(\lambda_j x)$ (see 137) in which, as before, $c_1, c_2, \cdots, c_n$ denote arbitrary constants. Some of the zeros may, of course, be complex; but, if the coefficients $p_n, \cdots, p_0$ are real, then complex zeros when they do occur will occur in complex pairs.

For example, if $\lambda_1 = \mu_1 + i\omega_1$, then there will be another zero, which may be labelled λ_2, of the form $\lambda_2 = \mu_1 - i\omega_1$; the linearly independent solutions $\exp(\lambda_1 x)$, $\exp(\lambda_2 x)$ may be replaced by $\exp(\mu_1 x)\cos(\omega_1 x)$, $\exp(\mu_1 x)\sin(\omega_1 x)$. Putting this another way, it can be said that, if $\Pi_n(t)$ contains the simple quadratic factor $(t-\mu_1)^2 + \omega_1^2$, then the complementary function contains the corresponding terms that are products of the exponential $\exp(\mu_1 x)$ and a linear combination of the sine and cosine of $\omega_1 x$ (see 138).

On the other hand, if λ_1 is a multiple root, of order m, say—*i.e.*, if $\Pi_n(t)$ has a factor $(t-\lambda_1)^m$—the complementary function contains the terms $(c_1 x + c_2 x^2 + \cdots + c_m x^{m-1})\exp(\lambda_1 x)$. Similarly, if $\Pi_n(t)$ has a factor of the form $[(t-\mu_1)^2 + \omega_1^2]^m$, there are $2m$ terms of the analogous form involving $\exp(\mu_1 x)\cos(\omega_1 x)$ and $\exp(\mu_1 x)\sin(\omega_1 x)$ (see 139) in the complementary function.

Use of the Laplace transform

In many practical cases it is possible to obtain a particular integral $Y(x)$ by inspection. When this cannot be done, the Laplace transform is used. Because for $Y(x)$ any solution (see 135) may be taken, it is only necessary to find the solution $Y(x)$ of that equation satisfying the conditions of $Y(x)$ and its first $n-1$ derivatives at $x = 0$ being equal to zero (see 140). If a function $\phi(x)$ is defined for all positive real values of x, its Laplace transform is defined as the result of multiplying the function by e^{-px} and integrating from $x = 0$ to infinity (see 141). When $\mathfrak{L}[\phi(p)]$ is the Laplace transform of $\phi(x)$, $\phi(x)$ is said to be the inverse Laplace transform of $\mathfrak{L}[\phi(p)]$ and $\phi(x) = \mathfrak{L}^{-1}[\phi(p); x]$.

To apply the Laplace transform to the solution of equations of the type (135), a formula is used expressing the Laplace transform of the derivative of a function in terms of the Laplace transform of the function itself. If the function $Y(x)$ satisfies the conditions (140), it is found that if $m = 1, 2, \cdots, n$ and $\mathfrak{L}[Y(p)]$ denotes the Laplace transform of $Y(x)$, then the Laplace transform of the mth derivative of $Y(x)$ is the product of p^m and $\mathfrak{L}[Y(p)]$ (see 142). Taking the Laplace transform of both sides of equation (135), it is found that $\mathfrak{L}[Y(p)]$ satisfies the simple algebraic equation $\Pi_n(p)\mathfrak{L}[Y(p)] = \mathfrak{L}[f(p)]$ in which $\mathfrak{L}[f(p)]$ denotes the Laplace transform of $f(x)$. For the particular integral, this equation may be solved for $Y(p)$ and the inverse Laplace transform (see 143) taken.

Linear equations of the second order. The general theory of linear differential equations will be illustrated by examples of equations of the second order $\mathbf{L}y = f(x)$, in which $\mathbf{L}$ denotes the operator $p(x)D^2 + q(x)D + r(x)$.

A homogeneous equation $\mathbf{L}_y = 0$ is said to be exact if it can be written in the form $D\mathbf{M}_y = 0$ in which $\mathbf{M}$ is a first-order linear differential operator of the form $p(x)D + s(x)$. The equation $\mathbf{L}_y = 0$ can be integrated once to give a first-order linear equation in which $\mathbf{M}_y$ is equal to an arbitrary constant c_1 (see 144), which can in turn be integrated by the method described above for the solution of equation (113); this will, of course, involve a second arbitrary constant c_2. This equation (144) is called a first integral of the original second-order equation. The equation $py'' + qy' + ry = 0$ is exact if, and only if, $p'' - q' + r = 0$ for all values of x, in which case $s(x) = q(x) - p'(x)$.

If the equation $\mathbf{L}_y = 0$ is not exact, but there exists a function $z(x)$ such that $z\mathbf{L}_y = 0$ is exact, z is said to be an integrating factor of $\mathbf{L}_y = 0$. From the condition for a second-order equation to be exact it is easily deduced that z is an integrating factor if it is a solution of the equation $\mathbf{L}^* z = 0$ in which the operator $\mathbf{L}^*$ is defined in terms of differentiation (see 145) and is called the adjoint of L. If $\mathbf{L}^* = \mathbf{L}$, the operator L is said to be self-adjoint; the condition for this is that $q(x) = p'(x)$.

The two standard methods of solving a second-order equation are based on writing the solution in the form $y = uv$ in which u is a prescribed function of x and v is the new dependent variable.

The two standard methods of solution

If u is chosen to be any solution of the first-order equation $2pu' + qu = 0$, the resulting equation for v is of the form $v'' + P(x)v = 0$, and the general solution of this equation may be known. This procedure is known as reducing the equation to normal form, or as removing the first derivative.

(138) $$e^{\mu_1 x}(c_1 \cos \omega_1 x + c_2 \sin \omega_1 x)$$

(139) $$(c_1 + \cdots + c_m x^{m-1})e^{\mu_1 x}\cos(\omega_1 x) + (c_{m+1} + \cdots + c_{2m}x^{m-1})\, e^{\mu_1 x}\sin(\omega_1 x)$$

(140) $$Y(0) = Y'(0) = \cdots = Y^{(n-1)}(0) = 0$$

(141) $$\overline{\phi}(p) \equiv \mathfrak{L}[\phi(x); p] = \int_0^\infty \phi(x)e^{-px}\,dx$$

(142) $$\mathfrak{L}[Y^{(m)}(x); p] = p^m \overline{Y}(p)$$

(143) $$Y(x) = \mathfrak{L}^{-1}\left[\frac{\overline{f}(p)}{\Pi_n(p)}; x\right]$$

(144) $$p(x)y' + s(x)y = c_1$$

(145) $$\mathbf{L}^* z = (pz)'' - (qz)' + r$$

(146) $$y_2(x) = y_1(x)\int_a^x \frac{dt}{p(t)[y_1(t)]^2}$$

(147) $$\sum_{r=0}^{\infty} a_r x^{r+\rho}, \qquad a_0 \neq 0$$

(148) $$y''(x) + k^2 \sin y = 0$$

(149) $$y'' + ay + by^2 = 0$$

(150) $$y'' - \varepsilon(1 - y^2)y' + y = 0$$

(151) $y'(x)=g(x, c_1)$

(152) $$\int \frac{dp}{\sqrt{(1+p^2)}}=\int \frac{dx}{a}+c_1$$

(153) $$y=a\cosh\left(\frac{x}{a}+c_1\right)+c_2$$

(154) $$\int \frac{dy}{h(y, c_1)}=x+c_2$$

(155) $$p^2+ay^2+\tfrac{2}{3}by^3=c_1$$

(156) $$p^2-2k^2\cos y=c_1$$

(157) $$\frac{d}{dx}H(x, y, y')=0$$

(158) $$\begin{cases} xyy''+2yy'+xy'^2=0 \\ \dfrac{d}{dx}(2xyy'+y^2)=0 \end{cases}$$

(159) $$\left[\frac{d}{dx}+p(x)\right]M(x, y, y')=0$$

(160) $$M(x, y, y')=c_1\exp[-\textstyle\int p(x)\,dx]$$

(161) $$y''+\alpha(x)y'+\beta(x)y=f(x), \qquad a<x<b$$

(162) $$y(x)=c_1y_1(x)+c_2y_2(x)+Y(x)$$

(163) $$\begin{cases} c_1y_1(a)+c_2y_2(a)=y_0-Y(a) \\ c_1y_1'(a)+c_2y_2'(a)=y_1-Y'(a) \end{cases}$$

(164) $$c_1=\frac{y_0y_2'(a)-y_1y_2(a)}{W(y_1, y_2; a)}, \qquad c_2=\frac{y_1y_1(a)-y_0y_1'(a)}{W(y_1, y_2; a)}$$

(165) $$\begin{cases} \dfrac{\partial^2G}{\partial x^2}+\alpha(x)\dfrac{\partial G}{\partial x}+\beta(x)G=0, \qquad G(a, \xi)=0 \\ \dfrac{\partial G(a, \xi)}{\partial x}=0, \qquad G(\xi, \xi)=0, \qquad \dfrac{\partial G(\xi, \xi)}{\partial x}=1 \\ Y(x)=\displaystyle\int_a^x G(x, \xi)f(\xi)\,d\xi \end{cases}$$

The second method consists in finding a particular solution of the equation by inspection and taking this solution to be u. The resulting second-order equation for v is, in effect, a first-order linear equation for $w=v'$. This can easily be solved and v obtained by a further integration. This method is particularly useful in the case of a self-adjoint equation $(py')'+ry=0$; if $y_1(x)$ is any solution of this equation the general solution is $c_1y_1(x)+c_2y_2(x)$ in which $y_2(x)$ is defined in terms of $y_1(x)$ and an integral (see 146), the lower limit of integration a being chosen to ensure that the integral exists.

Another useful method of solving a linear second-order equation is to derive a solution in the form of a power series (see 147), by substituting a series in the equation. If the coefficient of the lowest power of x is equated to zero, a quadratic equation is obtained; either of the roots of this quadratic equation gives a possible value of ρ. If the coefficient of x to the power $r+\rho$ is equated to zero, a recurrence relation is obtained for the coefficients a_r of this series.

NONLINEAR DIFFERENTIAL EQUATIONS

Some examples of nonlinear differential equations. Nonlinear equations of the second order may now be considered.

One of the simplest nonlinear second-order differential equations has the sum of y'' and $k^2\sin y$ equal to zero (see 148). It describes the motion of a simple pendulum, and also the deflection of a thin strut.

The equation (see 16) defining the catenary is a nonlinear equation.

The motion of an undamped mass-spring system in which the spring control takes the asymmetrical form $ay+by^2$ may be described by the nonlinear equation in which the sum of y'' and the spring control term equals zero (see 149). If there is damping, the motion is described by the slightly more complicated equation $y''+ky'+ay+by^2=0$, $k>0$.

Further, as was noted above, certain properties of thermionic tubes can be predicted from considerations of a differential equation involving an expression of the form $(1-y^2)y'$ (see 150).

Methods of solution of nonlinear differential equations. It is much more difficult to derive the solutions of nonlinear equations than those of linear equations. When a given differential equation is one of a simple type, however, its solution can be determined by elementary means.

Equations not containing y explicitly. A nonlinear second-order equation of the type $y''=f(x, y')$ can be regarded as a first-order equation $p'=f(x, p)$ for the derivative $p=y'$. When the solution $p(x)=g(x, c_1)$ of this equation has been found, the solution of the original equation is found by integrating the resulting equation (see 151). The catenary equation is of this form and so has a first integral (see 152), which is equivalent to $\sinh^{-1}(p)=x/a+c_1$; *i.e.*, to $p=\sinh(x/a+c_1)$. Integration of this equation in turn leads to the general solution in terms of a hyperbolic cosine (see 153).

Equations not containing x explicitly. In a similar way a nonlinear second-order equation of the type $y''=f(y, y')$ can be reduced to the first-order form $p(dp/dy)=f(y, p)$ because if $p=y'$, then $y''=p(dp/dy)$. If the solution of this first-order equation is $p(y)=h(y, c_1)$, the solution of the original equation is obtained by an integration (see 154).

For example, the undamped mass-spring equation (149) has a first integral involving p^2 and y^3 (see 155); and the pendulum equation (148) has a first integral involving p^2 and $\cos y$ (see 156).

Exact equations. If a nonlinear equation $F(x, y, y', y'')=0$ can be written as the derivative of some function $H(x, y, y')$ being equal to zero (see 157), it has the first integral $H(x, y, y')=c_1$ and hence can be solved by the methods available for the solution of first-order equations.

For example, a second-order equation involving x, y, y', $(y')^2$, and y'' can be written in the equivalent form in which the derivative of an expression involving a first derivative of y equals zero (see 158), and therefore it can be solved by this method.

A slight generalization of this method arises when the equation can be factorized in the form of the derivative of an expression involving x, y, and y' added to the product of that expression and a function of x being equal to zero (see 159), which can be integrated to give a first-order equation involving an exponential (see 160).

BOUNDARY VALUE PROBLEMS

When differential equations occur in the discussion of a physical problem, it usually happens that the solution of interest is not the general solution but the particular solution that satisfies certain additional conditions. For instance, to solve the problem of the catenary does not require the general solution of the second-order differential equation (16) but the particular solution, $y(x)$, that satisfies the additional geometrical conditions $y(0)=a$, $y'(0)=0$.

To illustrate the ideas involved, equations of the second order will be discussed; they are easily generalized to equations of the nth order.

Initial value problems. The problem of finding a function $y(x)$ in the interval $a<x<b$ satisfying a second-order differential equation $F(x, y, y', y'')=0$ and the conditions $y(a)=y_0$, $y'(a)=y_1$, in which y_0, y_1 are pre-

scribed real numbers, is called the initial value problem for the differential equation. The values y_0, y_1 are called the initial values of the required solution. The adjective initial is appropriate because in a dynamical problem (as in the problem of determining the escape velocity from the Earth, discussed above) the conditions of this type specify the initial state and initial rate of change of the system the subsequent history of which is described by the differential equation itself. Initial value problems are also known as one-point boundary value problems.

Linear equations. The linear second-order differential equation with x between a and b, in which $\alpha(x)$, the coefficient of y', $\beta(x)$, the coefficient of y, and $f(x)$, the inhomogeneous term, are prescribed functions (see 161), may now be considered with the initial values prescribed as above.

If $y_1(x)$ and $y_2(x)$ are linearly dependent solutions of the homogeneous form of the equation and $Y(x)$ is a particular solution of the nonhomogeneous equation (161), then the solution of the initial value problem for the differential equation is the sum of $Y(x)$ and a linear combination, with constants c_1 and c_2, of $y_1(x)$ and $y_2(x)$ (see 162).

The constants c_1 and c_2 are chosen to satisfy the simultaneous linear equation arising from setting the solution and its derivative equal to y_0 and y_1, respectively (see 163). If the particular solution $Y(x)$ is chosen in such a way that $Y(a) = Y'(a) = 0$, the equations for c_1 and c_2 reduce to a simple pair the solution of which can be found by Cramer's rule (see 164). This solution involves $W(y_1, y_2; a)$, the Wronskian $y_1(a)y_2'(a) - y_2(a)y_1'(a)$, which is nonzero because the solutions y_1 and y_2 are linearly independent. The solution of the initial value problem is given by equations (162) and (164).

The Green's function for a linear initial value problem. There is a simple formula for the determination of the particular integral $Y(x)$ of the linear differential equation of second order (see 161) satisfying the initial condition $Y(a) = Y'(a) = 0$. If $G(x, \xi)$ is a function of the two variables x and ξ such that G and its partial derivatives with respect to x satisfy the homogeneous equation, conditions at $x = \xi$ and at $x = a$, then $Y(x)$ can be expressed as an integral of the product $G(x, \xi)f(\xi)$ (see 165). The function $G(x, \xi)$ is called the Green's function of the problem (named after George Green, a 19th-century English mathematician). It can be shown that the Green's function can be expressed in terms of y_1, y_2, and the Wronskian $W(y_1; y_2; \xi)$ (see 166) in which $y_1(x)$ and $y_2(x)$ have the same meanings as before.

For example, for the initial value problem in which the sum of the second derivative y'' and $\omega^2 y$ equals $f(x)$ with y and its first derivative equalling y_0 and y_1, respectively, at $x = 0$ (see 167), the functions $y_1(x) = \cos(\omega x)$ and $y_2(x) = \sin(\omega x)$ may be taken to obtain the formula $G(x, \xi) = \omega^{-1} \sin\{\omega(x - \xi)\}$ from equation (166). Therefore, it is possible to express the solution as the sum of a linear combination of $\cos(\omega x)$ and $\sin(\omega x)$ and the integral of the product of $f(\xi)$ and the Green's function (see 168).

Nonlinear equations. The solution of the initial value problem for a nonlinear equation is a much more difficult problem except in those rare cases in which it is possible to find a general solution of the differential equation. The particular values of the arbitrary constants c_1 and c_2 are then found by fitting the values of $y(a)$ and $y'(a)$.

For equations of the type $y'' = g(x, y, y')$, a solution can be generated in the following manner. Differentiation of both sides of the equation by the chain rule produces an expression for the third derivative y''' (see 169), and it is possible to obtain higher derivatives by repeated differentiation.

Substitution of the initial values, the expressions for $y''(a)$, $y'''(a)$, and higher derivatives at $x = a$ (see 170) in the Taylor expansion about $x = a$ (see 171) leads to a series expansion for the solution of the initial value problem.

Two-point boundary value problems. The problem of finding a function $y(x)$ in the interval $a < x < b$ satisfying a second-order differential equation $F(x, y, y', y'') = 0$ and the conditions $y(a) = y_0$, $y(b) = y_1$ in which y_0, y_1 are prescribed real numbers is called the two-point boundary value problem for the differential equation. This problem can be generalized to cover the boundary conditions expressed by setting linear combinations of y and y' at each point equal to prescribed constants (see 172); only the simple case is treated here.

Linear equations. The two-point boundary value problem for a second-order linear equation (see 161) is con-

(166) $$G(x, \xi) = \frac{y_1(\xi)\, y_2(x) - y_1(x)\, y_2(\xi)}{W(y_1, y_2; \xi)}$$

(167) $$\begin{cases} \dfrac{d^2y}{dx^2} + \omega^2 y - f(x), & x > 0 \\ y(0) = y_0, \qquad y'(0) = y_1 \end{cases}$$

(168) $$y_0(x) = y_0 \cos(\omega x) + \frac{y_1}{\omega} \sin(\omega x) + \frac{1}{\omega} \int_0^x f(\xi) \sin\{\omega(x - \xi)\}\, d\xi$$

(169) $$y''' = \frac{\partial g(x, y, y')}{\partial x} + \frac{\partial g(x, y, y')}{\partial y} y' + \frac{\partial g(x, y, y')}{\partial y'} y''$$

(170) $$\begin{cases} y(a) = y_0, \qquad y'(a) = y_1, \qquad y''(a) = g(a, y_0, y_1) \\ y'''(a) = g_x(a, y_0, y_1) + y_1 g_y(a, y_0, y_1) + g(a, y_0, y_1)\, g_{y'}(a, y_0, y_1) \end{cases}$$

(171) $$y(x) = y(a) + \frac{(x - a)}{1!} y'(a) + \frac{(x - a)^2}{2!} y''(a) + \frac{(x - a)^3}{3!} y'''(a) + \cdots$$

(172) $$\lambda_1 y(a) + \mu_1 y'(a) = y_0 \quad \lambda_2 y(b) + \mu_2 y'(b) = y_1$$

(173) $$y(x) = y_0 \frac{y_2(x)}{y_2(a)} + y_1 \frac{y_1(x)}{y_1(b)} + \eta(x)$$

(174) $$\eta(x) = y_2(x) \int_a^x \frac{y_1(\xi) f(\xi)\, d\xi}{W(y_1, y_2; \xi)} + y_1(x) \int_x^b \frac{y_2(\xi) f(\xi)\, d\xi}{W(y_1, y_2; \xi)}, \qquad a < x < b$$

(175) $$G(x, \xi) = \begin{cases} \dfrac{y_1(\xi)\, y_2(x)}{W(y_1, y_2; \xi)}, & a < \xi < x \\ \dfrac{y_1(x)\, y_2(\xi)}{W(y_1, y_2; \xi)}, & x < \xi < b \end{cases}$$

(176) $$y(x) = \frac{y_0 y_2(x)}{y_2(a)} + \frac{y_1 y_1(x)}{y_1(b)} + \int_a^b G(x, \xi) f(\xi)\, d\xi$$

(177) $$y_1(a) = 0, \qquad y_2(b) = 0$$

(178) $$\begin{cases} \dot{x} = a_{11} x + a_{12} y \\ \dot{y} = a_{21} x + a_{22} y \end{cases}$$

(179) $$\begin{cases} (a_{11} - \lambda) A + a_{12} B = 0 \\ a_{21} A + (a_{22} - \lambda) B = 0 \end{cases}$$

(180) $$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$$

sidered: $y_1(x)$ and $y_2(x)$ are taken to be linearly independent solutions satisfying the conditions $y_1(a) = 0$, $y_2(b) = 0$, and $\eta(x)$ is taken as the particular solution of the equation (161) satisfying the conditions $\eta(a) = \eta(b) = 0$. The solution of the two-point boundary value problem is then the sum of a linear combination of $y_1(x)$ and $y_2(x)$ and a function $\eta(x)$ (see 173).

The Green's function for a linear two-point boundary value problem. It can be shown that the function $\eta(x)$ can be expressed in terms of y_1x, y_2x, and integrals involving $f(\xi)$ (see 174), which gives the required particular integral. The function $G(x, \xi)$ defined in terms of the functions and their Wronskian (see 175) is called the Green's function of the two-point boundary value problem. In terms of it the solution of the problem may be written as the sum of an integral of $G(x, \xi)f(\xi)$ and a linear combination of $y_1(x)$ and $y_2(x)$ (see 176). In the use of this formula it is necessary to remember that $y_1(x)$, $y_2(x)$ are not completely arbitrary linearly independent solutions of the homogeneous form of (161) but must satisfy the boundary conditions (see 177).

Nonlinear equations. When the general solution of a second-order nonlinear equation is not known, the solution of a two-point boundary value problem is obtained by solving the initial value problem for $y(a) = y_0$, $y'(a) = \gamma$, and (usually by a method of trial and error) to find the value of γ that leads to a solution for which $y(b) = y_1$.

SYSTEMS OF DIFFERENTIAL EQUATIONS

Two simultaneous differential equations in two variables. Two simultaneous equations in two variables will now be considered. They can be solved if they are linear and of first order.

Linear systems. A pair of homogeneous linear equations of the first order with dependent variables x and y and independent variable t can be solved for dx/dt and dy/dt formally written as dotted letters. In other words, the system is reduced to a pair of equations expressing the derivative with respect to t each in terms of a linear combination of x and y (see 178). The functions $x(t) = Ae^{\lambda t}$, $y(t) = Be^{\lambda t}$ satisfy these equations if the linear algebraic equations obtained by substituting these expressions for $x(t)$ and $y(t)$ into the original system are satisfied (see 179).

Regarded as equations for A and B, these equations have a nontrivial solution only if λ is a root of the equation that arises from the requirement that the determinant of the matrix of a system of homogeneous linear algebraic equations be zero (see 180), in which case A and B may be expressed in terms of an arbitrary constant c (see 181).

If the roots λ_1, λ_2 of the quadratic equation (180) are distinct, there are two linearly independent solutions, and the general solution of the pair of equations (178) may be written as linear combinations of these two solutions involving the a_{ij} and two arbitrary constants, c_1 and c_2 (see 182). For example, if the sum of the derivative of x with respect to t, ax, and y equals zero and the derivative of y with respect to t equals the sum of $6x$ and y, then λ_1 must equal -1, λ_2 must equal -2, and the a_{ij} are determined (see 183) so that the general solution is a linear combination of e^{-t} and e^{-2t} involving arbitrary constants c_1 and c_2 (see 184). In order to obtain the solution for which $x(0) = x_0$, $y(0) = y_0$, it is necessary to choose the constants c_1 and c_2 in such a way that the solution satisfies these conditions at $t = 0$ (see 185). These equations are solved for c_1 and c_2 to give the particular solution desired (see 186).

Use of the Laplace transform. To obtain the solution of the pair of first-order linear differential equations (178), the formulas expressing the result of using the Laplace transform (often formally written as a bar over the expression transformed) on a derivative (see 187) are used to obtain simultaneous linear algebraic equations (see 188) for the Laplace transforms $\mathfrak{L}[x(p)]$, $\mathfrak{L}[y(p)]$ of the unknown functions $x(t)$, $y(t)$. For instance, the pair (183) considered above can be attacked in this way (see 189) with the solution giving the Laplace transforms $\mathfrak{L}[x(p)]$ and $\mathfrak{L}[y(p)]$ in terms of p, x_0, and y_0 (see 190). If these equations are appropriately rewritten (see 191) and the Laplace transforms are inverted, the solution (186) is obtained.

(181) $A = a_{12}c, \qquad B = (\lambda - a_{11})c$

(182) $\begin{cases} x(t) = a_{12}c_1e^{\lambda_1 t} + a_{12}c_2e^{\lambda_2 t} \\ y(t) = (\lambda_1 - a_{11})c_1e^{\lambda_1 t} + (\lambda_2 - a_{11})c_2e^{\lambda_2 t} \end{cases}$

(183) $\begin{cases} \dot{x} + 4x + y = 0, & y - 6x - y = 0 \\ a_{11} = -4, & a_{12} = -1, & a_{21} = 6 \\ a_{22} = 1, & \lambda_1 = -1, & \lambda_2 = -2 \end{cases}$

(184) $x(t) = -c_1e^{-t} - c_2e^{-2t}, \qquad y(t) = 3c_1e^{-t} + 2c_2e^{-2t}$

(185) $-c_1 - c_2 = x_0, \qquad 3c_1 + 2c_2 = y_0$

(186) $\begin{cases} x(t) = (3e^{-2t} - 2e^{-t})x_0 + (e^{-2t} - e^{-t})y_0 \\ y(t) = 6(e^{-t} - e^{-2t})x_0 + (3e^{-t} - 2e^{-2t})y_0 \end{cases}$

(187) $\mathfrak{L}[\dot{x}(t); p] = p\bar{x}(p) - x_0, \qquad \mathfrak{L}[\dot{y}(t); p] = p\bar{y}(p) - y_0$

(188) $\begin{cases} (a_{11} - p)\bar{x}(p) + a_{12}\bar{y}(p) + x_0 = 0 \\ a_{21}\bar{x}(p) + (a_{22} - p)\bar{y}(p) + y_0 = 0 \end{cases}$

(189) $(p + 4)\bar{x}(p) + \bar{y}(p) = x_0, \qquad -6\bar{x}(p) + (p - 1)\bar{y}(p) = y_0$

(190) $\bar{x}(p) = \dfrac{(p - 1)x_0 - y_0}{(p + 1)(p + 2)}, \qquad \bar{y}(p) = \dfrac{6x_0 + (p + 4)y_0}{(p + 1)(p + 2)}$

(191) $\begin{cases} \bar{x}(p) = \dfrac{3x_0 + y_0}{p + 2} - \dfrac{2x_0 + y_0}{p + 1} \\ \bar{y}(p) = \dfrac{6x_0 + 3y_0}{p + 1} - \dfrac{6x_0 + 2y_0}{p + 2} \end{cases}$

(192) $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

(193) $\mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}, \qquad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$

(194) $\dot{\mathbf{z}} = A\mathbf{z}$

(195) $\dfrac{d^2x}{dt^2} + p(t)\dfrac{dx}{dt} + q(t)x = f(t)$

(196) $\dfrac{dx}{dt} - y = 0, \qquad \dfrac{dy}{dt} + p(t)y + q(t)x = f(t)$

(197) $\mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}, \qquad A(t) = \begin{bmatrix} 0 & -1 \\ q(t) & p(t) \end{bmatrix}, \qquad \boldsymbol{\phi} = \begin{bmatrix} 0 \\ f(t) \end{bmatrix}$

(198) $\dot{\mathbf{z}} + A(t)\mathbf{z} = \boldsymbol{\phi}(t)$

(199) $\begin{cases} \dot{x}_1 = a_{11}x_1 + a_{12}x_2 \cdots + a_{1n}x_n \\ \dot{x}_2 = a_{21}x_1 + a_{22}x_2 \cdots + a_{2n}x_n \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_n = a_{n1}x_1 + a_{n2}x_2 \cdots + a_{nn}x_n \end{cases}$

(200) $\dot{\mathbf{z}} = A\mathbf{z}$

Matrix form of the pair of equations. The pair of equations (178) may be written in matrix form (see 192) if matrices $\mathbf{z}$, with elements x and y, and $\mathbf{A}$, with elements a_{11}, a_{12}, a_{21}, and a_{22} (see 193), are introduced. This last equation can be written in a matrix form similar to the form of a single first-order equation (see 194).

Second-order equations. The general second-order linear differential equation (see 195) can be written as a pair of simultaneous equations by defining a new variable y as equal to the first derivative dx/dt (see 196). In terms of matrices (see 197) similar to those in the previous example, this pair can, in turn, be written in matrix form (see 198).

A system of n-simultaneous equations in n variables. The system (178) can be generalized to the set of linear differential equations in n variables (see 199), which can be written in matrix form by writing the derivative of $\mathbf{z}$ with respect to t as equal to the matrix product $A\mathbf{z}$ (see 200), in which $\mathbf{z}$ is an n-component column vector and A is an $n \times n$ matrix of coefficients (see 201). If $\mathbf{z}$ equals a constant vector $\mathbf{c}$ times an exponential $e^{\lambda t}$ (see 202), and if I is the unit $n \times n$ matrix, then the matrix product $A\mathbf{c}$ must equal $\lambda I\mathbf{c}$ (see 203) for $\mathbf{z}$ to be a solution of the system.

In order for $\mathbf{c}$ to be a non-null vector, it is necessary that $\det(A - \lambda I) = 0$; *i.e.*, the vector (202) is a solution of equation (200) only if λ is an eigenvalue (see ALGEBRA, LINEAR AND MULTILINEAR) of the matrix A. This is illustrated by equation (180) in the case $n = 2$.

The main result is that the general solution of the system of equations (200) is the product of a constant vector $\mathbf{z}_0$ and an exponential e^{tA} (see 204), the exponential with matrix exponent being defined by the Maclaurin series expansion of the exponential (see 205).

PERTURBATIONS

Because it is difficult to obtain exact solutions of nonlinear differential equations, recourse is often taken to approximate methods, foremost among which are perturbation methods that are applicable to equations containing a small parameter ε. A typical such equation is one in which y'', $\omega^2 y$ and an additional term $\varepsilon f(y, y')$ all sum to zero (see 206); for example, van der Pol's equation (150) is of this form.

The two principal perturbation methods are the Lindstedt-Poincaré method and the Krylov-Bogolyubov method, which are discussed below.

Lindstedt-Poincaré method. The method used extensively by the 19th–20th-century Swedish mathematician Anders Lindstedt and the French mathematician Henri Poincaré for the solution of differential equations in celestial mechanics consists in assuming a solution that can be expanded in powers of ε, the coefficients of the expansion being $y_0(x)$, $y_1(x)$, $y_2(x)$, $\cdots$ (see 207), when it is assumed that $0 < \varepsilon < 1$, and that ω^2 has an expansion in powers of ε (see 208), the coefficients of which α_0, α_1, α_2, $\cdots$ are constants the values of which have to be determined.

For example, in order to find periodic solutions of the equation considered above (see 206) in a special case (see 209), a solution of the form given by the expansion (207) is substituted in the differential equation, and the coefficients of the powers of ε are equated to zero. This leads to a sequence of linear second-order equations with constant coefficients (see 210), which are easily solved; the unknown constants α_0, α_1, α_2, $\cdots$ are determined by the fact that each of the functions $y_r(x)$ must be periodic, with the same period.

Krylov-Bogolyubov method. If ε is zero, the equation being considered (206) clearly has the solution $y = A \sin(\omega x + \phi)$, in which A and ϕ are constants. The Krylov-Bogolyubov method consists in allowing A and ϕ to become slowly varying functions of x when ε is small but not zero.

An approximate solution is $y = A(x) \sin[\omega x + \phi(x)]$; this approximation is then substituted into the differential equation. And, because A and ϕ are slowly varying, it is thus possible to determine from the integrals first A and then ϕ (see 211).

STABILITY

There are very few differential equations the general solution of which can be derived, and so the study of the qualitative properties of solutions of differential equations without solving the equations explicitly has developed. One qualitative phenomenon of great practical interest is the concept of stability of a certain solution of a system of differential equations.

Suppose that a physical system is represented by a differential equation in which the derivative of an n-vector $\mathbf{x}$ equals a function of $\mathbf{x}$ and t, $\mathbf{f}(\mathbf{x}, t)$, which is also an n-vector (see 212), and that $\mathbf{x} = u(t)$ is a special solution of this equation describing a special regime of the system. Any physical system will be subject to small disturbances so that the regime $u(t)$ will not maintain itself, and a disturbed motion $\mathbf{x} = w(t)$ results. As time goes on, either $w(t) \to u(t)$, in which case stability occurs, or $w(t)$ deviates from $u(t)$ and instability results.

(201) $$\mathbf{z} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \qquad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix}$$

(202) $$\mathbf{z} = \mathbf{c}e^{\lambda t}$$

(203) $$(A - \lambda I)\mathbf{c} = 0$$

(204) $$\mathbf{z}(t) = e^{tA}\mathbf{z}_0$$

(205) $$e^{tA} = 1 + \frac{tA}{1!} + \frac{t^2A^2}{2!} + \cdots + \frac{t^nA^n}{n!} + \cdots$$

(206) $$y'' + \varepsilon f(y, y') + \omega^2 y = 0$$

(207) $$y(x) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \cdots$$

(208) $$\omega^2 = \alpha_0 + \varepsilon\alpha_1 + \varepsilon^2\alpha_2 + \cdots$$

(209) $$f(y, y') = -\varepsilon(1 - y^2)y', \qquad y(0) = 0, \qquad y'(0) = \omega_0 A$$

(210) $$\begin{cases} y_0'' + \alpha_0 y_0 = 0, & y_0(0) = 0, \quad y_0'(0) = \omega_0 A \\ y_1'' + \alpha_0 y_1 = -\alpha_1 y_0 + y_0'(1 - y_0^2) \\ \quad y_1(0) = y_1'(0) = 0 \\ y_2'' + \alpha_0 y_2 = -(\alpha_2 y_0 + \alpha_1 y_1) + y_1'(1 - y_0^2) - 2y_0 y_0' y_1 \\ \quad y_2(0) = y_2'(0) = 0 \end{cases}$$

(211) $$\begin{cases} \dfrac{dA}{dx} = -\dfrac{\varepsilon}{2\pi\omega}\displaystyle\int_0^{2\pi} f(A \sin t, A\omega \cos t) \cos t \, dt \\ \dfrac{d\phi}{dx} = \dfrac{\varepsilon}{2\pi\omega}\displaystyle\int_0^{2\pi} f(A \sin t, A\omega \cos t) \sin t \, dt \end{cases}$$

(212) $$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$$

(213) $$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}, t)$$

Putting $w(t) = u(t) + z(t)$ a 19th–20th-century-Soviet mathematician, Aleksandr Lyapunov, obtained an equation in which $\mathbf{z}$ and $\mathbf{g}$ are n-vectors, the derivative of $\mathbf{z}$ with respect to t equals $\mathbf{g}(\mathbf{z}, t)$ (see 213), and $\mathbf{g}(0, t) = 0$ for $t > 0$. To the regime $u(t)$ of the first equation (see 212) there corresponds the regime $\mathbf{z} = \mathbf{0}$ of the second equation (see 213). The origin $\mathbf{z} = \mathbf{0}$ is now considered as the basic motion and a solution $\mathbf{z}(t)$ of (213) is thought of as a disturbed motion. The terms stability of the undis-

turbed motion or simply the stability of the origin are used in this case.

The property $\mathbf{g}(0, t) = \mathbf{0}$ for a suitable range of t characterizes the origin as a critical point of (213). It is assumed that this critical point is isolated and also that $\mathbf{g}$ is continuous and satisfies conditions for the uniqueness of solutions of Lyapunov's equation (213) in a certain region Ω specified by $\| \mathbf{x} \| < A$, $t > 0$, in which $\| \mathbf{x} \|$ denotes the length of the vector $\mathbf{x}$ and A denotes a positive constant. Stability is characterized by Lyapunov as follows: if for any prescribed ε, $0 < \varepsilon < A$, there exists $\delta > 0$ such that every trajectory originating within the sphere $\| \mathbf{x} \| = \delta$ always remains within the sphere $\| \mathbf{x} \| = \varepsilon$, the origin is stable (cf. Figure 4 in which these spheres are denoted by $S(\delta)$ and $S(\varepsilon)$ respectively). If every trajectory originating within $S(\delta)$ tends to 0 the origin is said to be asymptotically stable. If, however, no matter what δ is chosen, some trajectory originating from a point within $S(\delta)$ eventually reaches $S(\varepsilon)$, the origin is unstable.

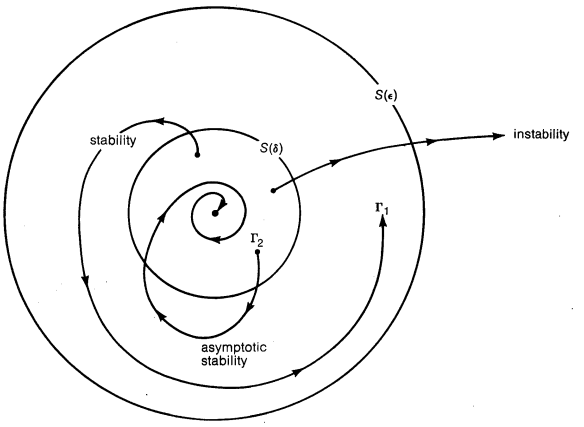


Figure 4: Three types of trajectories, stable as represented by Γ_1, asymptotically stable as represented by Γ_2, and unstable, drawn in relation to an origin and two spheres denoted by $S(\delta)$ and $S(\varepsilon)$.

Lyapunov attacked the stability problem by two distinct methods. The first, applicable only to some analytical systems, consists essentially in finding explicit series solutions convergent near the origin and considering the behaviour of this series as $t \to \infty$. The second method does not require the construction of explicit solutions, and for this reason that method has a greater range of validity than the first; it rests on the construction of a scalar function $V(\mathbf{x}, t)$.

For example, Lyapunov's stability states that if there can be found a positive definite function $V(\mathbf{x}, t)$ the time derivative of which taken along all the trajectories in Ω is negative or zero, then the origin is stable. Of four such theorems of Lyapunov, two refer to stability and two to instability; the latter may be replaced by a more comprehensive theorem of Četaev.

Partial differential equations

In most physical problems any dependent variable is likely to be a function of more than one independent variable. The example of a vibrating string has already been given. Another example is in the study of thermal effects in a solid body in which the temperature u may vary from point to point in the solid as well as from time to time. As a consequence, the function $u(x, y, z, t)$ is written to denote that u depends on (x, y, z), the coordinates of a typical point, and on the time t.

If for such a function, the difference quotient, the difference between $u(a + h, b, c, \tau)$ and $u(a, b, c, \tau)$, divided by h (see 214), tends to a finite number as h tends to zero, that number is denoted by $u_x(a, b, c, \tau)$, or by $\dfrac{\partial u(a, b, c, \tau)}{\partial x}$, and is called the partial derivative with respect to x of the function $u(x, y, z, t)$ at the point (a, b, c, τ). It measures the rate of change at that point of the function u with respect to x, the other variables remaining fixed. As the point (a, b, c, τ) varies, a new function $u_x(x, y, z, t)$ or $\partial u/\partial x$ is generated that expresses the rate of growth with x of the function u. The process can be repeated: the partial derivative of $\partial u/\partial x$ with respect to x is denoted by $\partial^2 u/\partial x^2$. Similarly the consideration of the variation of u with respect to the other variables leads to the definition of the other partial derivatives, the partial derivatives with respect to y, z, and t (see 215); and the partial derivatives with respect to x of these yield in turn the mixed derivatives of the second order in which u is differentiated twice, once with respect to each of two different variables (see 216).

When the laws of physics are applied to a situation involving a function of this kind, a relation involving the function u and its partial derivatives is often obtained. Such a relation is called a partial differential equation. For example, if $u(x, y, z, t)$ were the temperature at time t at the point (x, y, z) in a homogeneous isotropic solid, the variation of u is specified by the heat conduction equation, in which the sum of the second partial derivatives with respect to x, y, and z is proportional to the first partial derivative with respect to t (see 217). The expression on the left side of this equation arises so frequently that it is given a special name and a special symbol; it is called the Laplacian of u and is now denoted by $\Delta_3 u$.

As in the theory of ordinary differential equations, the order of a partial differential equation is defined to be the order of the derivative of highest order occurring in the equation. For example, the heat equation (see 217) is a second-order equation in four variables, an equation stating that the sum of the cube of the partial derivative of u with respect to x and the partial derivative of u with respect to y equals zero (see 218) is a first-order equation in two variables while an equation involving first partial derivatives with respect to x, y, and t with variable coefficients (see 219) is a first-order equation in three variables.

The study of partial differential equations began with

(214) $$\frac{u(a+h, b, c, \tau) - u(a, b, c, \tau)}{h}$$

(215) $$\frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}, \frac{\partial u}{\partial t}$$

(216) $$\frac{\partial^2 u}{\partial x \partial y}, \frac{\partial^2 u}{\partial x \partial z}, \frac{\partial^2 u}{\partial x \partial t}$$

(217) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{1}{\kappa}\frac{\partial u}{\partial t}$$

(218) $$\left(\frac{\partial u}{\partial x}\right)^3 + \frac{\partial u}{\partial y} = 0$$

(219) $$x\frac{\partial u}{\partial x} + y\frac{\partial u}{\partial y} + \frac{\partial u}{\partial t} = 0$$

(220) $$\frac{\partial u}{\partial \xi} = 0$$

(221) $$u = f(\eta)$$

(222) $$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0$$

(223) $$u = f(\xi) + g(\eta)$$

(224) $$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 4\pi\sigma(x, y, z)$$

Early work on partial differential equations

the study of continuum mechanics and physics. Although there were some earlier investigations, the first proper studies of these types of equations began in the 18th century with the studies of Leonhard Euler (Swiss) and Jean Le Rond d'Alembert (French) on wave motion, Pierre-Simon Laplace (French) on potential theory, Jean-Baptiste-Joseph Fourier (French) on the conduction of heat, and Carl Friedrich Gauss (German) on potential theory and electromagnetic theory. The theory of partial differential equations has always gone hand-in-hand with developments in mathematical physics and has drawn much of its strength from the association, as is seen even in the more theoretical works of pure mathematicians of distinction such as Augustin-Louis Cauchy (French), Bernhard Riemann (German), and Sofya Kovalevskaya (Russian). The growth of abstract analysis has had important repercussions on the theory of partial differential equations, and this influence has in turn suggested fruitful problems to workers in both abstract and numerical analysis.

Probably the simplest partial differential equation involving a function $u(\xi, \eta)$ of two independent variables is the one that may be verbalized as follows: the first partial derivative with respect to one of the variables equals zero (see 220), which may readily be seen to have the general solution of an arbitrary function $f(y)$ of the other variable only (see 221). A comparison with the case of ordinary differential equations in which the equation $u'(x) = 0$ has solution $u(x) = c$ illustrates the point that the role played by arbitrary constants in the theory of ordinary differential equations is taken over by arbitrary functions in the theory of partial differential equations.

Similarly, if the second mixed partial derivative of u with respect to ξ and η equals zero (see 222), then the general solution is the sum of $f(\xi)$ and $g(\eta)$, in which the functions f and g are arbitrary and are each functions of only one variable (see 223).

This statement can be extended to a general partial differential equation of order n for a function of k independent variables. Subject to a variety of restrictions that must be imposed, the general solution of such an equation depends on n arbitrary functions of $k - 1$ independent variables.

PARTIAL DIFFERENTIAL EQUATIONS OF SPECIAL INTEREST

Some partial differential equations have arisen so frequently in the study of physics that they have been studied intensively. Collectively, they are known as the equations of mathematical physics, and have throughout the development of the theory of partial differential equations provided the motivation for many studies of theoretical, as well as practical, interest.

Poisson's equation. A simple linear partial differential equation of the second order is named after a 19th-century French physicist, Siméon-Denis Poisson. It arises in the mathematical treatment of electrostatics. Gauss's law of electrostatics states that the flux of the electric vector (E_x, E_y, E_z) out of a surface is equal to 4π times the charge contained within that surface; this leads to the relation that the sum of $\partial E_x/\partial x$, $\partial E_y/\partial y$, and $\partial E_z/\partial z$ is proportional to $\sigma(x, y, z)$ (see 224), in which σ, the density of electric charge, is assumed to be known. It is also known physically that the electrostatic field is characterized by the fact that the electric vector can be derived from a potential function u; *i.e.*, that there exists a function u such that each component of the electric vector equals minus the partial derivative of u with respect to that variable (see 225). The insertion of these expressions into the previous equation leads to the relation that the sum of $\Delta_3 u$ and $4\pi\sigma$ (x, y, z) equals zero (see 226), in which Δ_3 is the Laplacian operator introduced above. Poisson's equation arises also in the theory of the gravitational potential.

Laplace's equation. In the absence of electric charges the function σ of equation (226) reduces to the simpler form in which the Laplacian of u equals zero (see 227). This is called Laplace's equation. It arises also in the analysis of the irrotational motion of a perfect fluid and in the steady-state conduction of heat (see below *The diffusion equation*).

In the cylindrical coordinates (ρ, ϕ, z) Laplace's equation assumes a form that can be derived by the chain rule (see 228) in which it is more suitable for the discussion of problems involving regions bounded by cylinders. In a similar manner, in spherical polar coordinates (r, θ, ϕ), which are used in the analysis of problems concerning regions that are bounded by spheres or cones, Laplace's equation is transformed to an expression involving partial derivatives with respect to these variables (see 229).

The wave equation. *Sound waves in space.* If, because of the passage of a sound wave, the gas at a point (x, y, z) at time t has velocity (u, v, w) and pressure p and

(225) $$E_x = -\frac{\partial u}{\partial x}, \qquad E_y = -\frac{\partial u}{\partial y}, \qquad E_z = -\frac{\partial u}{\partial z}$$

(226) $$\Delta_3 u + 4\pi\sigma(x, y, z) = 0$$

(227) $$\Delta_3 u = 0$$

(228) $$\frac{\partial^2 u}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial u}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

(229) $$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2 \sin\theta}\left(\frac{\partial}{\partial \theta}\sin\theta\frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u}{\partial \phi^2} = 0$$

(230) $$\frac{\partial u}{\partial t} = -c^2\frac{\partial s}{\partial x}, \qquad \frac{\partial v}{\partial t} = -c^2\frac{\partial s}{\partial y}, \qquad \frac{\partial w}{\partial t} = -c^2\frac{\partial s}{\partial z}$$

(231) $$\frac{\partial s}{\partial t} + \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$$

(232) $$u = -\frac{\partial \psi}{\partial x}, \qquad v = -\frac{\partial \psi}{\partial y}, \qquad w = -\frac{\partial \psi}{\partial z}$$

(233) $$\begin{cases} \dfrac{\partial}{\partial x}\left(\dfrac{\partial \psi}{\partial t} - c^2 s\right) = 0, \qquad \dfrac{\partial}{\partial y}\left(\dfrac{\partial \psi}{\partial t} - c^2 s\right) = 0 \\ \dfrac{\partial}{\partial z}\left(\dfrac{\partial \psi}{\partial t} - c^2 s\right) = 0 \end{cases}$$

(234) $$\frac{\partial s}{\partial t} = \Delta_3 \psi$$

(235) $$\frac{1}{c^2}\frac{\partial^2 \psi}{\partial t^2} = \Delta_3 \psi$$

(236) $$\Delta_2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

(237) $$\Delta_2 u = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}$$

(238) $$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}$$

(239) $$c_x = -\kappa\frac{\partial u}{\partial x}, \qquad c_y = -\kappa\frac{\partial u}{\partial y}, \qquad c_z = -\kappa\frac{\partial u}{\partial z}$$

(240) $$\frac{\partial u}{\partial t} + \frac{\partial c_x}{\partial x} + \frac{\partial c_y}{\partial y} + \frac{\partial c_z}{\partial z} = 0$$

(241) $$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\kappa\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\kappa\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(\kappa\frac{\partial u}{\partial z}\right)$$

density ρ, then for small oscillations $\rho = \rho_0(1 + s)$, $p = p_0 + c^2\rho_0 s$, in which s is the condensation of the gas and $c^2 = (dp/d\rho)_0$, and ρ_0, p_0 are the equilibrium values of p and ρ. The equations of (small) motions are that the partial derivative of each component of the velocity with respect to time equals the product of c^2 and the appropriate partial derivative of s (see 230); and the equation of continuity is to the same approximation that the sum of $\partial u/\partial x$, $\partial v/\partial y$, $\partial w/\partial z$, and $\partial s/\partial t$ equals zero (see 231). For irrotational motion there exists a function ψ with the property that each component of the velocity equals minus the appropriate partial derivative of ψ (see 232).

The above equations (see 230 and 232) are therefore equivalent to the equation that the partial derivative of $\partial\psi/\partial t - c^2 s$ with respect to each of x, y, and z is zero (see 233). The equation of continuity (231) is equivalent to the restriction that the partial derivative of s with respect to t equals the Laplacian of ψ (see 234) so that the function ψ satisfies the relation that the second partial derivative of ψ with respect to t is proportional to the Laplacian of ψ (see 235). This equation is called the three-dimensional wave equation.

It can also be shown that in the absences of charges or currents the scalar potential and each component of the vector potential of an electromagnetic field satisfy the three-dimensional wave equation with c the velocity of light.

Transverse vibrations of a membrane. If a thin elastic membrane of uniform areal density σ is stretched to a uniform tension T and if, in the equilibrium position, the membrane coincides with the xy-plane, then the small transverse vibration $u(x, y, t)$ of the point (x, y) of the membrane satisfies an equation in which the two-dimensional Laplacian of u (see 236) is proportional to the second partial derivative with respect to t (see 237), and the wave velocity c is defined by the equation $c^2 = T/\sigma$. Equation (237) is called the two-dimensional wave equation.

Transverse vibrations of a string. If a string of uniform linear density σ is stretched to a uniform tension T, and if, in the equilibrium position, the string coincides with the x-axis, then when the string is disturbed slightly from its equilibrium position, the transverse displacement $u(x, t)$ satisfies an equation in which the second partial derivative of u with respect to x is proportional to the second partial derivative with respect to t (see 238), with constant of proportionality $c^2 = T/\sigma$. This equation is known as the one-dimensional wave equation.

The diffusion equation. Another interesting partial differential equation arises in the analysis of the diffusion process in physical chemistry. This is a process leading to the equalization of concentrations within a single phase, and it is governed by laws relating the rate of flow of the diffusing substance with the concentration gradient causing the flow. If u is the concentration of the diffusing substance, then the diffusion current vector (c_x, c_y, c_z) is defined by the equations in which each component of the diffusion current vector equals a quantity $-\kappa$ times the appropriate first partial derivative of u (see 239). This is known as Fick's law, named after a 19th-century German physiologist, Adolf Fick, in which κ is the diffusion coefficient for the substance under consideration.

Fick's law

The continuity for the diffusion substance takes a form (see 240) that is similar to the usual continuity equation so that the concentration $u(x, y, z, t)$ satisfies the diffusion equation (see 241) obtained by substituting Fick's law into the equation of continuity. In the most general case, the quantity κ is a function of x, y, z and of the concentration u, but if κ does happen to be a constant, the diffusion equation (241) reduces to the form of the heat conduction equation (217).

The diffusion equation (241) is satisfied also by the temperature u in a solid conducting heat; for that reason this equation is sometimes known as the heat conduction equation. In the steady-state case in which κ is constant, it follows from equation (217) with $\partial u/\partial t \equiv 0$ that u satisfies Laplace's equation (227).

The form (217) of the diffusion equation is also satisfied by the vorticity in a viscous fluid started into motion from rest and by the electric field vector in the propagation of long waves in a good conductor.

If u depends only on one space coordinate x and the time t, the heat conduction equation (217) reduces to a form in which the second partial derivative of u with respect to x is proportional to the first partial derivative of u with respect to t (see 242); this is called the one-dimensional diffusion equation.

The Fokker-Planck equation. An equation very similar to the diffusion equation (242) is the Fokker-Planck equation linearly relating a quantity P and its first and second derivatives with respect to x with its first derivative with respect to t (see 243), which reduces in the case that the coefficient D of the second derivative is equal to zero to a first-order equation with a variable coefficient of $\partial P/\partial x$ (see 244). The physical interpretation is that P is the probability that a random variable has the value x at time t. For instance, P might be the probability distribution of the deflection x of an electrical noise trace at time t.

Birth-and-death equations. Equations similar to the Fokker-Planck equation arises in the theory of birth and

(242) $$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\kappa}\frac{\partial u}{\partial t}$$

(243) $$\frac{\partial P}{\partial t} = \beta\frac{\partial}{\partial x}(xP) + D\frac{\partial^2 P}{\partial x^2}$$

(244) $$\frac{\partial P}{\partial t} = \beta P + \beta x\frac{\partial P}{\partial x}$$

(245) $$u(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n$$

(246) $$\frac{\partial u}{\partial t} = (z-1)(\lambda z - \mu)\frac{\partial u}{\partial z}$$

(247) $$\frac{\partial E}{\partial x} + Ri + L\frac{\partial i}{\partial t} = 0, \qquad \frac{\partial i}{\partial x} + GE + C\frac{\partial E}{\partial t} = 0$$

(248) $$\frac{\partial^2 u}{\partial x^2} = LC\frac{\partial^2 u}{\partial t^2} + (RC + LG)\frac{\partial u}{\partial t} + RGu$$

(249) $$\Delta_3\psi + \frac{8\pi^2 m}{h^2}(W - V)\psi = 0$$

(250) $$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} = 0, \qquad \frac{\partial \sigma_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} = 0$$

(251) $$\sigma_{xx} = \frac{\partial^2\psi}{\partial y^2}, \qquad \sigma_{xy} = -\frac{\partial^2\psi}{\partial x\partial y}, \qquad \sigma_{yy} = \frac{\partial^2\psi}{\partial x^2}$$

(252) $$\frac{\partial^2}{\partial y^2}\{\sigma_{xx} - \eta(\sigma_{xx} + \sigma_{yy})\} + \frac{\partial^2}{\partial x^2}\{\sigma_{yy} - \eta(\sigma_{xx} + \sigma_{yy})\} = 2\frac{\partial^2 \sigma_{xy}}{\partial x\partial y}$$

(253) $$\frac{\partial^4\psi}{\partial x^4} + 2\frac{\partial^4\psi}{\partial x^2\partial y^2} + \frac{\partial^4\psi}{\partial y^4} = 0$$

(254) $$(\tfrac{1}{2}\sigma_{xx} - \tfrac{1}{2}\sigma_{yy})^2 + \sigma_{xy}^2 = k^2$$

(255) $$\left(\frac{\partial^2\psi}{\partial x^2} - \frac{\partial^2\psi}{\partial y^2}\right)^2 + 4\frac{\partial^4\psi}{\partial x^2\partial y^2} = 4k^2$$

death processes in bacterial populations. If it is assumed that the probability of the birth or death of a bacterium is proportional to the number present, and if $P_n(t)$ is the probability of there being n bacteria in the population at time t, the probability generating function $u(z, t)$ defined by a power series in z with the nth term having coefficient $P_n(t)$ (see 245) satisfies a partial differential equation in which $\partial u/\partial t$ equals a quadratic in z, with constants λ and μ, times $\partial u/\partial z$ (see 246).

Similar equations arise in birth-and-death problems in which different physical assumptions are made and in discussions of the probability distribution of telephone conversations carried on over a certain number of telephone lines.

Poincaré's telegraphy equation. An interesting second-order equation was derived by Poincaré in his study of the one-dimensional flow of electricity in a long insulated cable. The current $i(x, t)$ and the voltage $E(x, t)$ then satisfy the pair of differential equations involving (see 247) R, L, C, and G, which denote, respectively, the series resistance, the inductance, capacitance, and conductance per unit length. It can then be shown that both E and i satisfy a second-order partial differential equation linearly relating u, its second partial derivative with respect of x, and its first and second partial derivatives with respect to t (see 248), which is known as the telegraphy equation.

If the leakage to ground is small, so that G and L may be taken to be small, the telegraphy equation reduces to the one-dimensional diffusion equation with $\kappa = (RC)^{-1}$. On the other hand, if the emphasis is on high-frequency phenomena on a cable, the terms involving the time derivatives predominate. Equations (247) show that this is equivalent to taking G and R to be zero, and hence that the telegraphy equation (248) reduces to the one-dimensional wave equation (238) with $c = (LC)^{-\frac{1}{2}}$. (In this context, the equation is sometimes referred to as the radio equation.)

Schrödinger's equation. In wave mechanics the wave function ψ of a single particle such as an electron has the physical interpretation that $|\psi(x, y, z)|^2 d\tau$ is the probability that the electron will be found in a small element of volume $d\tau$ the centre of which is at the point with coordinates (x, y, z). In the steady-state case the variation of ψ throughout a field the potential energy of which is $V(x, y, z)$ is described by Schrödinger's equation involving m, the mass of the particle, h, Planck's constant, and E, the total energy of the particle and linearly relating the Laplacian of ψ and the quantity $(E - V)\psi$ (see 249).

Higher-order equations in physics. The above equations are—with the exception of one or two first-order equations—all of the second order. It is therefore important to notice that not all physical problems can be formulated in terms of partial differential equations of the second order.

For instance, a state of plane strain in a two-dimensional solid free from body forces can be specified by three stress components σ_{xx}, σ_{xy}, and σ_{yy} that satisfy equilibrium conditions that can be stated as first-order linear partial differential equations (see 250). If σ_{xx}, σ_{xy}, and σ_{yy} are equal to the appropriate second partial derivatives of any arbitrary function ψ, then they (see 251) satisfy the equilibrium conditions, and ψ is called an Airy stress function.

The Airy stress function

If the body is elastic, the components of stress satisfy compatibility equations involving the second partial derivatives of σ_{xx}, σ_{xy}, and σ_{yy} (see 252), from which it follows that $\psi(x, y)$ satisfies a fourth-order linear equation that can be written $\Delta_2\Delta_2\psi = 0$ and is called the two-dimensional biharmonic equation (see 253). (The reason for this is that Laplace's equation is sometimes referred to as the harmonic equation.)

If, instead of being elastic, the body is perfectly plastic, then the stress components satisfy the quadratic condition; this quadratic condition is known as the Hercky-Mises condition (see 254). Hence, the Airy stress function satisfies a fourth-order nonlinear partial differential equation (see 255).

CLASSIFICATION OF PARTIAL DIFFERENTIAL EQUATIONS

A partial differential equation governing a dependent variable u is said to be linear if u and its partial derivatives occur only to the first power. For example, if second-order equations in two independent variables x and y are considered, then a linear equation involves a sum of u and its first and second derivatives, with coefficients a, b, c, f, g, h, and k that are functions of x and y only (see 256).

The solution of a nonlinear equation is considerably more difficult to obtain than the solution of a linear equation; general methods of solution have been devised, however, for two special types of nonlinearity. A nonlinear equation in which the partial derivatives of u occur linearly—but in which u itself does not do so (see 257)—is said to be quasi-linear, whereas an equation that is linear in only the highest order derivatives (see 258) is said to be semilinear.

(256) $$a\frac{\partial^2 u}{\partial x^2}+2b\frac{\partial^2 u}{\partial x\partial y}+c\frac{\partial^2 u}{\partial y^2}+f\frac{\partial u}{\partial x}+g\frac{\partial u}{\partial y}+hu=k$$

(257) $$a\frac{\partial^2 u}{\partial x^2}+2b\frac{\partial^2 u}{\partial x\partial y}+c\frac{\partial^2 u}{\partial x\partial y}+f\frac{\partial u}{\partial x}+g\frac{\partial u}{\partial y}=h(x, y, u)$$

(258) $$a\frac{\partial^2 u}{\partial x^2}+2b\frac{\partial^2 u}{\partial x\partial y}+c\frac{\partial^2 u}{\partial y^2}=f\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$$

(259) $$a\frac{\partial u}{\partial x}+b\frac{\partial u}{\partial y}=c$$

(260) $$\frac{dx}{a}=\frac{dy}{b}=\frac{du}{c}$$

(261) $$\frac{dx}{x^2}=\frac{dy}{y^2}=\frac{du}{(x+y)u}$$

(262) $$x^2\frac{\partial u}{\partial x}+y^2\frac{\partial u}{\partial y}=(x+y)u$$

(263) $$u^2(1+u_x^2+u_y^2)=1$$

(264) $$\begin{cases}F(D, D')z=f(x, y)\\ F(D, D')=\sum\limits_r\sum\limits_s c_{rs}D^rD'^s\end{cases}$$

(265) $$\sum_{s=1}^{n} x^{s-1}\phi_s(ax-y)e^{-bx}$$

(266) $$F(D, D')e^{ax+by}=F(a, b)e^{ax+by}$$

(267) $$\sum_{n=0}^{\infty} c_n\cos(nx+\varepsilon_n)e^{-\kappa n^2 t}$$

First-order equations. There is a particular kind of non-linear equation of the first order in which the product a and $\partial u/\partial x$ is added to the product of b and $\partial u/\partial y$ to yield c (see 259), in which a, b, and c are functions not only of x and y but of u also. This type of equation has been investigated by the 18th-century French mathematician Joseph-Louis Lagrange.

The general solution of this equation is $F(\xi, \eta) = 0$, in which F is an arbitrary function of two variables and $\xi(x, y, u) = C_1$ and $\eta(x, y, u) = C_2$ form a solution of the simultaneous ordinary differential equations expressed by equating the ratios of dx to a, dy to b, and du to c (see 260). For example, because a system of equations of this form with x^2 for a, y^2 for b, and $(x + y)u$ for c (see 261) have solution $xy/u = c_1$, $(x - y)/u = c_2$, it follows that the general solution of the equation with these terms for

a, b, and c (see 262) is $F(xy/u, (x-y)/u) = 0$, in which the function F is arbitrary.

Classes of integrals in first-order equations

The problem of solving a general nonlinear equation of the first order $f(x, y, u, u_x, u_y) = 0$ in which the function f is not linear in u_x and u_y is more difficult. It turns out that there are three classes of integrals of an equation of this type:

(1) Two-parameter systems of surfaces $F(x, y, u, a, b) = 0$; such an integral is called a complete integral.
(2) If any one-parameter subsystem $F(x, y, u, a, \phi(a)) = 0$ is taken and its envelope is formed (by eliminating a from the equations $F = 0$, $F_a = 0$), a general integral is obtained.
(3) If the envelope of the complete integral in (1) (obtained by eliminating a, b from the equations $F = 0$, $F_a = 0$, $F_b = 0$) exists, a singular integral of the equation is obtained.

For example, there is the equation in which the sum of 1, u_x^2, and u_y^2, that, when multiplied by u^2 (see 263), has the complete integral $(x-a)^2 + (y-b)^2 + u^2 = 1$. If $b = a$ and if the envelope of the resulting subsystem of spheres is written, then the general integral $(x-y)^2 + 2u^2 = 2$ is obtained; moreover, because the envelope of the complete integral consists of the planes $u = \pm 1$, then it follows that $u = 1$ and $u = -1$ are both singular integrals of the equation.

Solutions of nonlinear first-order equations are usually obtained by a method known as Charpit's method, which is too complicated to be described here. The solutions of some special types are, however, easily derived.

If the equation is of the type $f(u_x, u_y) = 0$—*i.e.*, if it involves only the derivatives u_x, u_y—the complete integral is $u = ax + Q(a)y + b$, in which a and b are arbitrary constants and $q = Q(a)$ is the solution of the algebraic equation $f(a, q) = 0$.

If the equation is of the type $f(u, u_x, u_y) = 0$, u_x and u_y are found by solving the equations $f(u, u_x, u_y) = 0$, $u_x = au_y$ and then determining $u(x, y)$ from the equation $du = u_x dx + u_y dy$.

A separable equation $f(x, u_x) = g(y, u_y)$ is solved by determining u_x and u_{xy} from the equations $f(x, u_x) = a$ and $g(y, u_y) = a$, respectively, and finding u as in the last case.

And, finally, there are Clairaut equations, $u = xu_x + yu_y + f(u_x, u_y)$, which have complete integrals $u = ax + by + f(a, b)$.

Linear equations with constant coefficients. Linear partial differentials with constant coefficients can be solved simply. Such an equation can be written in the form of an equality between a linear combination of derivatives $F(\mathrm{D}, \mathrm{D}')z$ and $f(x, y)$, a function of x and y (see 264). Any solution of the equation is called a particular integral and the most general solution of the corresponding homogeneous equation $F(\mathrm{D}, \mathrm{D}')z = 0$ is called the complementary function of the equation; if z_1 is a particular integral and u is the complementary function, the general solution is $u + z_1$. It follows from the linearity of the operator F that if $u_1, u_2, \cdots, u_n$ are solutions of $Fz = 0$, then $\sum_{r=1}^{n} c_r u_r$, in which the c's are arbitrary constants, is also a solution.

The operators $F(\mathrm{D}, \mathrm{D}')$ are classified into two main types: $F(\mathrm{D}, \mathrm{D}')$ is reducible if it can be written as the product of linear factors of the form $\mathrm{D} + a\mathrm{D}' + b$ with a, b constants; if it cannot be written in such a form, then $F(\mathrm{D}, \mathrm{D}')$ is said to be irreducible.

The form of the complementary function in the reducible case follows from the fact that if $(\mathrm{D} + a\mathrm{D}' + b)^n$ is a factor of $(\mathrm{D}, \mathrm{D}')$ and if the functions $\phi_1, \phi_2, \cdots, \phi_n$ are arbitrary, then a sum involving these functions and an exponential (see 265) is a solution of $F(\mathrm{D}, \mathrm{D}') = 0$. When F is irreducible it is not always possible to find a solution with the full number of arbitrary functions, but it is possible to construct solutions that contain as many arbitrary constants as are required. The method of deriving such solutions depends on the result of differentiating an exponential (see 266) so that e^{ax+by} is a solution of $F(\mathrm{D}, \mathrm{D}') = 0$, provided that a and b are connected through the equation $F(a, b) = 0$. In this way it is possible to construct a solution of the homogeneous equation that contains as many arbitrary constants as may be required.

For example, the one-dimensional diffusion equation (see 242) has a solution $\exp(ax + bt)$ only if $b = a^2\kappa$, and this relation is satisfied if $a = \pm in$ and $b = -\kappa n^2$ are taken. In this way solutions of the diffusion equation can be constructed as the sum of terms each of which is the product of a cosine and an exponential and involving constants c_n and ε_n (see 267).

Semilinear equations of the second order with variable coefficients. Next in order of complexity to an equation with constant coefficients is a semilinear second-order equation (see 258). This equation is reduced to canonical form by changing the independent variables from x, y to ξ, η. The choice of ξ and η is dictated by the nature of the roots $\lambda_1(x, y)$, $\lambda_2(x, y)$ of the quadratic equation $a\lambda^2 + 2b\lambda + c = 0$. There are three cases to be considered depending on the sign of the discriminant Δ defined by the equation $\Delta = b^2 - ac$.

The three types of semilinear second-order equations

(i) $\Delta > 0$: If for every (x, y) in a region Ω of the xy-plane, $\Delta > 0$, the roots λ_1, λ_2 are real and distinct. If ξ and η are now chosen to be such that $\xi(x, y) = c_1$ and $\eta(x, y) = c_2$ are, respectively, the integral curves of the ordinary differential equations $y' + \lambda_1(x, y) = 0$ and $y' + \lambda_2(x, y) = 0$, the equation is transformed to the canonical form in which the second mixed partial derivative equals some function of ξ, η, u and the first partial derivatives of u (see 268). In this case the equation is said to be hyperbolic in the region Ω. The curves $\xi(x, y) = c_1$, $\eta(x, y) = c_2$ are called the characteristic curves of the equation, or, more simply, the characteristics of the equation. An equation that is hyperbolic in a region Ω, therefore, has two families of characteristics in that region.

For example, an equation in which the coefficient of $\partial u/\partial x$ is $1/x$ and the coefficient of $\partial^2 u/\partial y^2$ is x^2 (see 269) is hyperbolic at all points in the xy-plane, has the two families of characteristics $x^2/2 + y =$ constant and canonical form with the second mixed partial equal to zero (see 270). From past equations, it is apparent that this equation has the general solution $u = f(\xi) + g(\eta)$, and thus the original equation has the general solution $u = f(x^2/2 + y) + g(x^2/2 - y)$, in which the functions f and g are arbitrary.

(ii) $\Delta = 0$: The function ξ is defined in precisely the same way as in case (i) and η is taken to be any function of x and y that is independent of ξ. The canonical form of the equation in this case has the second partial derivative with respect to η equal to a function of ξ, η, u and the first partial derivatives of u (see 271). An equation of this type is said to be parabolic.

For example, an equation in which twice the second mixed partial derivatives is added to the other two second partial derivatives to give a result of zero (see 272) is parabolic. If $\xi = x - y$ and $\eta = x + y$, it is found that it reduces to the canonical form in which the second partial derivative with respect to η equals zero (see 273), with solution $u = \eta f(\xi) + g(\xi)$, in which f and g are arbitrary functions. The solution of the original equation is therefore obtained by substituting the appropriate quantities for η and ξ in this expression (see 274).

(iii) $\Delta < 0$: In this case the roots λ_1 and λ_2 are complex conjugates so that if ξ and η are chosen as in case (i) they will be complex-valued functions of the real variables x and y. In terms of this pair of functions ξ, η another pair α, β are defined as a linear combination of ξ and η involving complex numbers (see 275). When transformed to the new independent variables, the equation takes the canonical form in which the second partial derivative with respect to α, when added to the second partial derivative with respect to β, equals some function of α, β, u, and the first partial derivatives of u (see 276). Such an equation is said to be of elliptic type.

For example, an equation in which the second partial derivative with respect to x, when added to the product of x^2 and the second partial derivative with respect to y,

gives zero (see 277) is elliptic throughout the whole xy-plane. If the method in (i) is used, then it is found that $\xi = x^2/2 + iy$ and $\eta = x^2/2 - iy$ may be taken, and hence $\alpha = x^2/2$, $\beta = y$. Transforming to these new variables, the equation has the canonical form with the function being the product of $-1/(2\alpha)$ and the first partial derivative with respect to α (see 278).

INITIAL VALUE AND BOUNDARY VALUE PROBLEMS

In applications of partial differential equations, the concern is not so much with determining the general solution of an equation as with finding the solution that satisfies some other prescribed conditions. Problems involving the solution of a partial differential equation and the fulfillment of additional conditions have been classified into the following types. Only second-order equations in two independent variables will be considered.

(268) $$\frac{\partial^2 u}{\partial \xi \partial \eta} = h\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right)$$

(269) $$\frac{\partial^2 u}{\partial x^2} - \frac{1}{x}\frac{\partial u}{\partial x} = x^2 \frac{\partial^2 u}{\partial y^2}$$

(270) $$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0$$

(271) $$\frac{\partial^2 u}{\partial \eta^2} = h\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right)$$

(272) $$\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} = 0$$

(273) $$\frac{\partial^2 u}{\partial \eta^2} = 0$$

(274) $$u = (x+y)f(x-y) + g(x-y)$$

(275) $$\alpha = \tfrac{1}{2}(\xi + \eta), \qquad \beta = \tfrac{1}{2}i(\eta - \xi)$$

(276) $$\frac{\partial^2 u}{\partial \alpha^2} + \frac{\partial^2 u}{\partial \beta^2} = k\left(\alpha, \beta, u, \frac{\partial u}{\partial \alpha}, \frac{\partial u}{\partial \beta}\right)$$

(277) $$\frac{\partial^2 u}{\partial x^2} + x^2 \frac{\partial^2 u}{\partial y^2} = 0$$

(278) $$\frac{\partial^2 u}{\partial \alpha^2} + \frac{\partial^2 u}{\partial \beta^2} + \frac{1}{2\alpha}\frac{\partial u}{\partial \alpha} = 0$$

(279) $$u(0, t) = 0, \qquad u(a, t) = 0, \qquad t > 0$$

(280) $$u(x, 0) = f(x), \qquad u_t(x, 0) = g(x), \qquad 0 < x < a$$

(281) $$u_{xy} = f(x, y, u, u_x, u_y)$$

Initial value problems. The classic case of an initial value problem arises in the discussion of the transverse oscillations of a stretched string. The transverse displacement $u(x, t)$ satisfies the partial differential equation (238) for $0 < x < a$, $t > 0$, in which a denotes the length of the string. Because the string is assumed to be fixed at its ends, $u(0, t)$ and $u(a, t)$ must both equal zero for all positive values of t (see 279), and if the problem is that of determining the subsequent shape of the string when its initial shape and velocity are prescribed, the function u must satisfy equations in which $u(x, 0)$ and $u_t(x, 0)$ equal prescribed functions $f(x)$ and $g(x)$ (see 280).

These particular equations—that is, (238), (279), and (280)—describe a problem that is analogous to an initial value problem in the theory of differential equations. Arguing from this fact and also from physical intuition, one would expect to realize the existence of a unique solution $u(x, t)$. Because u satisfies the boundary conditions (279) as well as the initial conditions (280), the present problem is said to be a mixed initial and boundary value problem.

A different example is provided by solutions of the heat conduction equation (see 242). Here, on physical grounds, it would be expected that a unique function $u(x, t)$ would be determined by the boundary conditions (see 279) and the single initial condition $u(x) = f(x)$, $0 < x < a$.

Cauchy problems. A generalization of the initial value problems considered above is Cauchy's problem. This is best described for a hyperbolic equation in canonical form with independent variables x and y (see 281). The Cauchy problem for this equation consists of finding a solution u with the property that prescribed values $u = u(s)$, $u_x = p(s)$, $u_y = q(s)$ are assumed along a given curve with parametric equations $x = x(s)$, $y = y(s)$. Because u, p, q must satisfy the compatibility equation found by the chain rule (see 282), p and q cannot be assigned independently. The values of u and its normal derivative on the curve $\partial u/\partial n$ are in fact sufficient for the specification of u, p, and q. These quantities are referred to as the Cauchy data.

Dirichlet problems. The first boundary value problem, also called the Dirichlet problem, consists of finding a solution $u(x, y)$ of an elliptic equation in a region Ω of the xy-plane that takes on prescribed values on the boundary of Ω.

Neumann problems. The second boundary value problem for an elliptic partial differential equation in a region Ω consists in finding a solution that possesses a prescribed normal derivative $\partial u/\partial n = f$ on the boundary B of Ω. It is often called a Neumann problem. It turns out the Neumann problem cannot be solved for Laplace's equation $u_{xx} + u_{yy} = 0$ unless the prescribed function f satisfies the condition that the integral of f around the boundary equal zero (see 283).

Robin problems. The third boundary value problem, or Robin's problem, for an elliptic partial differential equation in a region Ω consists of finding a solution u of the equation satisfying the condition that the sum of αu and the normal derivative of u equal f (see 284) at all points of the boundary of Ω, α and f being prescribed.

SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS

In many applications it is necessary to consider a pair of first-order equations each involving one partial derivative with respect to x and two partial derivatives with respect to y (see 285) with coefficients of the partial derivatives with respect to y being a_{11}, a_{12}, a_{21}, a_{22}, which are functions of x and y and b_1, b_2 are functions of x, y, u. The first stage in solving equations of this type is to transform the system into one in which each equation involves differentiations in one direction only.

If a certain quadratic equation in λ formed from a_{11}, a_{22}, a_{12}, and a_{21} (see 286) has two real distinct roots $\lambda_1(x, y)$, $\lambda_2(x, y)$, the system can be transformed to a pair of equations, each involving only one differentiation operator D_j (see 287), in which D_j is the operator of differentiation consisting of differentiation with respect to x added to the product of λ_j and differentiation with respect to y (see 288) in the direction defined by an ordinary differential equation in which the derivative of y equals λ_j (see 289). It should be emphasized that this reduction is possible if, and only if, the equation (286) has two distinct real roots. When this is the case, the system is said to be of hyperbolic type. On the other hand, if the roots of (286) are complex, the system is of elliptic type. The two families of curves in the xy-plane defined by the ordinary differential equations (289) are called the characteristics of the system.

If the column vector U, with elements u and v, the column vector B, with elements b_1 and b_2, and the square matrix A, with elements a_{11}, a_{12}, a_{21}, and a_{22} (see 290), are introduced, the system may be written as a matrix equation relating the partial derivatives U_x and U_y (see 291)

and the equation in λ (see 286) as an equation stating that the determinant of $A - \lambda I$ is zero (see 292). It is now an easy matter to generalize the system being considered (see 291) by taking U, B to be column vectors with m components and A is a matrix with m rows and m columns. The determinant equation (292) now has m roots. If these roots are real and distinct the system can be transformed to m equations, each involving only one differentiation operator D_j (see 293). The only change in the interpretation is that the equations defining D_j (288 and 289) now hold for $j = 1, 2, \cdots, m$. Again the system (291) is hyperbolic.

A quasi-linear second-order partial differential equation can be replaced by a system of the type in which the partial derivative of U with respect to x equals the matrix product of A and the partial derivative of U with respect to y (see 294), in which U has 8 components and A is an 8×8 matrix.

TECHNIQUES OF SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

Separation of variables. One of the oldest techniques for the solution of a linear partial differential equation for a function u of the independent variables $x, y, \cdots, t$ is to assume a solution that is a product of $X(x), Y(y), \cdots, T(t)$ of functions of one variable (see 295) and by substituting in the original equation to obtain ordinary differential equations for the functions $X, Y, \cdots, T$.

For example, if u satisfies the one-dimensional wave equation (see 238), the equation has solutions of the type $u(x, t) = X(x)T(t)$ if X and T can be chosen such that the ratio of $X''(x)$ to $X(x)$ is proportional to the ratio of $T''(t)$ to $T(t)$ (see 296). Because the left side of this equation is a function of x alone and the right side is a function of t alone, the two can be equal for all values of x and t if, and only if, each is equal to a constant $-\lambda c^2$, say; *i.e.*, it is necessary that X and T satisfy second-order linear differential equations with constant coefficients (see 297).

The value of λ is determined by the boundary conditions. For example, if the boundary conditions are that $u(0, t) = u(a, t) = 0$ for all values of $t > 0$, they are satisfied if $X(0) = X(a) = 0$. This is possible only if $\lambda = n^2\pi^2/a^2$, $n = 1, 2, \cdots$, when $X(x) = (\sin n\pi x/a)$. If in addition $u_t(x, 0) = 0$, then $T'(0) = 0$ also; *i.e.*, $T(t) = \cos(n\pi ct/a)$; thus a solution of the equation (238) in the form of a function $u_n(x, t)$ equalling the product of a cosine and a sine, involving t and x respectively (see 298), is obtained. This technique is called separation of variables.

Quite complicated solutions can be constructed by combining this method with the principle of linear superposition. For example, an infinite linear combination of the $u_n(x, t)$, with coefficients c_n (see 299) with u_n given above (see 298), will be a solution of the wave equation (see 238) for $0 \leqslant x \leqslant a$, $t \geqslant 0$, satisfying the boundary conditions $u(0, t) = u(a, t) = 0$ and the initial conditions $u(x, 0) = f(x)$, $u_t(x, 0) = 0$ if an infinite sequence of constants $c_1, c_2, \cdots$ can be found such that $f(x)$ can be written as an infinite series with terms $cn \sin(n\pi x/a)$ (see 300). This leads to a consideration of the theory of Fourier series (see ANALYSIS, FOURIER). It should also be observed that the values $\lambda_n = n^2\pi^2/a^2$, $n = 1, 2, \cdots$, are called the eigenvalues of the stated problem.

Integral transform solutions. Closely related to the method of separation of variables is the method of integral transforms. To illustrate the method, the solution of Laplace's equation in two dimensions (see 301) is considered in the semi-infinite strip $x \geqslant 0$, $0 \leqslant y \leqslant a$, when u satisfies the boundary conditions $u(0, y) = 0$, $u(x, y) \to 0$ as $x \to \infty$, $0 \leqslant y \leqslant a$ and $u(x, 0) = 0$, $u(x, a) = f(x)$.

The method of separation of variables then gives a solution of the form $\sin(\xi x)\sinh(\xi y)$, in which ξ is any constant. By the superposition principle such a solution, when multiplied by $F(\xi)/\sinh(\xi a)$ and integrated from zero to infinity (see 302), will be the required solution if a function F can be found with the property that for all $x > 0$, $f(x)$ equals the integral from zero to infinity of $F(\xi)\sin(\xi x)$ (see 303). This leads to a consideration of the theory of Fourier integrals. The Fourier sine transform of $f(x)$ is $F(\xi)$.

Probably the best known example of an integral transform is the Laplace transform, which was considered earlier. To illustrate its use, the solution of the one-dimensional diffusion equation (see 242) is considered in the range $x \geqslant 0$, $t \geqslant 0$, and satisfying the conditions $u(x, 0) = 0$, $u(0, t) = f(t)$, $u(x, t) \to 0$ as $x \to \infty$. If the Laplace transforms $\mathfrak{L}[u(x, p)]$ of $u(x, t)$ and $\mathfrak{L}[f(p)]$ of $f(t)$ (see 304) are introduced, then it is easily shown that $\mathfrak{L}[u(x, p)] = \mathfrak{L}[f(p)] \exp[-x\sqrt{(p/\kappa)}]$; thus the required solution is the function the Laplace transform of which is $\mathfrak{L}[f(p)] \exp[-x\sqrt{(p/\kappa)}]$. The solution can be found by means of the inversion theorem, which gives a method of finding a function the Laplace transform of which is prescribed.

Variational methods. Approximate solutions of partial differential equations can often be obtained by using the fact that the function u that makes a certain integral an extremum satisfies a partial differential equation.

To illustrate the procedure, the function $u(x, y, z)$ is considered that makes the volume integral $I(u)$ of a function $F(x, y, z, u, u_x, u_y, u_z)$ of the coordinates, u, and its first partial derivatives (see 305), an extremum with respect to twice-differentiable functions that assume prescribed values at all points of the boundary S of V. This

(282) $$\frac{du}{ds} = p(s)\frac{dx}{ds} + q(s)\frac{dy}{ds}$$

(283) $$\int_B f ds = 0$$

(284) $$\frac{\partial u}{\partial n} + \alpha u = f$$

(285) $$\begin{cases} \dfrac{\partial u}{\partial x} + a_{11}\dfrac{\partial u}{\partial y} + a_{12}\dfrac{\partial v}{\partial y} = b_1 \\ \dfrac{\partial v}{\partial x} + a_{21}\dfrac{\partial u}{\partial y} + a_{22}\dfrac{\partial v}{\partial y} = b_2 \end{cases}$$

(286) $$(a_{11} - \lambda)(a_{22} - \lambda) - a_{21}a_{12} = 0$$

(287) $$\begin{cases} D_1 u + (\lambda_1 - a_{11})D_1 v = b_1 + (\lambda_1 - a_{11})b_2 \\ D_2 u + (\lambda_2 - a_{11})D_2 v = b_1 + (\lambda_2 - a_{11})b_2 \end{cases}$$

(288) $$D_j = \frac{\partial}{\partial x} + \lambda_j \frac{\partial}{\partial y}, \qquad j = 1, 2$$

(289) $$\frac{dy}{dx} = \lambda_j(x, y), \qquad j = 1, 2$$

(290) $$U = \begin{bmatrix} u \\ v \end{bmatrix}, \qquad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

(291) $$U_x + AU_y = B$$

(292) $$\det(A - \lambda I) = 0$$

(293) $$\sum_{k=1}^{m} c_{jk} D_j U_k + \sum_{k=1}^{m} c_{jk} b_k = 0, \qquad j = 1, 2, \cdots, m$$

(294) $$U_x = AU_y$$

(295) $$u(x, y, \cdots, t) = X(x)Y(y)\cdots T(t)$$

(296) $$c^2 X''(x)/X(x) = T''(t)/T(t)$$

(297) $$X''(x) + \lambda X(x) = 0, \qquad T''(t) + c^2\lambda T(t) = 0$$

must satisfy the Euler-Lagrange equation (see 306). Hence, the solution of the Dirichlet problem for the Euler-Lagrange equation (306) is the function u that takes prescribed values on S and makes $I(u)$ an extremum. The value of this observation lies in the fact that in certain cases direct methods may produce a solution of the extremal problem.

Dirichlet's principle

For example, from among the functions that have continuous second derivatives in V and on S and take on the prescribed values f on S, that function which makes the integral of the squares of the first derivatives (see 307) an extremum is the solution of the Dirichlet problem for Laplace's equation (see 308). This is known as Dirichlet's principle.

Perturbation methods. Perturbation theory was created by a British physicist Lord Rayleigh in the 19th century. He gave a formula for computing the natural frequencies and modes of a vibrating system the physical characteristics of which differ slightly from those of a simpler system the behaviour of which is completely known. In terms of the method of separation of variables, a differential equation of the form $[\mathbf{L}(\varepsilon)+\lambda]X=0$ has to be solved, subject to certain boundary conditions, $\mathbf{L}(\varepsilon)$ being a differential operator containing a small numerical parameter ε, when the eigenvalues and eigenfunctions of $[\mathbf{L}(0)+\lambda]X=0$ with the same boundary conditions, are known.

For example, the method of separation of variables applied to the one-dimensional wave equation with an additional term of the product of ε and the fourth partial derivative of u with respect to x (see 309) leads to the perturbed form of the first of the separated equations (297), a fourth-order linear ordinary differential equation (see 310).

The perturbation method consists essentially in expanding $\mathbf{L}(\varepsilon)$ in powers of ε with coefficients $\mathbf{L}^1$, $\mathbf{L}^2$, $\cdots$, which are differential operators (see 311) and assuming that X_n and λ_n can also be expanded in powers of ε, in which $X_n^{(0)}(x)$ is the eigenfunction corresponding to the eigenvalue $\lambda_n^{(0)}$ of the unperturbed problem (see 312).

Approximate solutions. One of the most frequently used methods of obtaining approximate solutions of partial differential equations is the method of finite differences, which consists essentially in replacing each partial derivative by a difference quotient. To be specific, the case of a function $u(x, t)$ of two independent variables x and t is considered. If the value of u were known at each point (rh, sk) in which r and s are integers and h and k are small positive quantities, an approximation could be made to the partial derivative $\partial u/\partial x$ at the point (rh, sk) by the difference quotient formed by taking the difference between $u_{r+1,s}$ and $u_{r,s}$ and dividing by h (see 313) in which $u_{r,s}\equiv u(rh, sk)$, and the second derivative $\partial^2 u/\partial x^2$ by applying the process again to the difference quotient just obtained (see 314). Suitable approximations for $\partial u/\partial t$ and $\partial^2 u/\partial t^2$ can be found by the same process (see 315).

One possible finite-difference approximation to the one-dimensional diffusion equation (see 242) is therefore found by replacing the derivatives by the appropriate difference quotients (see 316), and the equation can be solved for $u_{r,s+1}$, with $\rho=k/h^2$. This gives a formula by means of which the unknown $u_{r,s+1}$ might be determined at the $(r, s+1)$ mesh point in terms of the known functions along the sth row. In this way the unknown values of u can be calculated along the first row, $t=k$, in terms of known boundary values along $t=0$. The values along the second row, $t=2k$, can be obtained in terms of the calculated values along the first, and the process is repeated to obtain the values along successive rows in terms of previously calculated values.

If u is the exact solution of the one-dimensional diffusion equation (see 242) and U the exact solution of the difference equation (see 316), then if $U\to u$ as both h and k tend to zero, the finite difference equations are said to be convergent. At first sight it appears as though the mesh lengths h and k can be chosen arbitrarily, but this is not always so; for example, for the particular case of the difference equation (see 316) convergence only occurs if $\rho\leqslant\frac{1}{2}$. In the case of hyperbolic equations, it is possible to combine the use of characteristics with that of finite differences to obtain numerical solutions with a high degree of accuracy.

The Cauchy-Kovalevskaya theorem

To obtain approximate solutions of boundary value problems for elliptic equations, it is sometimes possible to reformulate the problem as an integral equation and then to obtain approximate solutions of it.

(298) $$u_n(x,t)=\cos\left(\frac{n\pi ct}{a}\right)\sin\left(\frac{n\pi x}{a}\right)$$

(299) $$u(x,t)=\sum_{n=1}^{\infty}c_n u_n(x,t)$$

(300) $$f(x)=\sum_{n=1}^{\infty}c_n\sin\left(\frac{n\pi x}{a}\right)$$

(301) $$u_{xx}+u_{yy}=0$$

(302) $$u(x,y)=\int_0^{\infty}F(\xi)\frac{\sinh(\xi y)}{\sinh(\xi a)}\sin(\xi x)\,d\xi$$

(303) $$f(x)=\int_0^{\infty}F(\xi)\sin(\xi x)\,d\xi$$

(304) $$\begin{cases}\mathfrak{L}[u(x,p)]=\int_0^{\infty}u(x,t)e^{-pt}\,dt\\ \mathfrak{L}[f(p)]=\int_0^{\infty}f(t)e^{-pt}\,dt\end{cases}$$

(305) $$I(u)=\int_V F(x,y,z,u,u_x,u_y,u_z)\,d\tau$$

(306) $$\frac{\partial F}{\partial u}=\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right)+\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right)+\frac{\partial}{\partial z}\left(\frac{\partial F}{\partial u_z}\right)$$

(307) $$I(u)=\int_V[u_x^2+u_y^2+u_z^2]\,d\tau$$

(308) $$\begin{cases}u_{xx}+u_{yy}+u_{zz}=0 & \text{within } V\\ u=f, & \text{on } S\end{cases}$$

(309) $$\frac{\partial^2 u}{\partial x^2}-\varepsilon\frac{\partial^4 u}{\partial x^4}=\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}$$

(310) $$X''(x)-\varepsilon X^{(iv)}(x)+\lambda X(x)=0$$

(311) $$\mathbf{L}(\varepsilon)=\mathbf{L}(0)+\varepsilon\mathbf{L}^{(1)}+\varepsilon^2\mathbf{L}^{(2)}+\cdots$$

(312) $$\begin{cases}X_n=X_n^{(0)}(x)+\varepsilon X_n^{(1)}(x)+\varepsilon^2 X_n^{(2)}(x)+\cdots\\ \lambda_n=\lambda_n^{(0)}+\varepsilon\lambda_n^{(1)}+\varepsilon^2\lambda_n^{(2)}+\cdots\end{cases}$$

EXISTENCE AND UNIQUENESS OF SOLUTIONS TO PARTIAL DIFFERENTIAL EQUATIONS

The basic existence and uniqueness theorem of the theory of partial differential equations is the Cauchy-Kovalevskaya theorem, which in its simplest form relates to a system in two independent variables x, y in which u_x equals the product of $A(u)$ and uy, $u(0, y)$ being equal to $h(y)$ (see 317), u and h being column vectors with m components, the vector h being prescribed, and A is an $m\times m$ matrix.

To state the result it is necessary to use the fact that a function is said to be analytic at a point if it has a Taylor expansion at that point.

The Cauchy-Kovalevskaya theorem then states that about any point at which the components of the given matrix A and of the given vector h are analytic, a neighbourhood can be found in which there exists a unique

vector u, with analytic coefficients, which is a solution of the initial value problem (see 317).

Although it has the appearance of great generality, this theorem has the great limitation that it is restricted to problems involving only analytic functions. What it does show is that within the class of analytic solutions of equations with analytic coefficients, the number of arbitrary functions required for a general solution is equal to the order of the equation. When k independent variables are involved, a generalization of the theorem shows that each such arbitrary function is a function of $k - 1$ variables.

GENERALIZED THEORY OF PARTIAL DIFFERENTIAL EQUATIONS

The generalized theory of partial differential equations depends on generalizing the concept of a derivative. In order to do this it is necessary to introduce the idea of a test function.

(313) $$\frac{u_{r+1,s}-u_{r,s}}{h}$$

(314) $$\frac{u_{r+1,s}-2u_{r,s}+u_{r-1,s}}{h^2}$$

(315) $$\begin{cases} \dfrac{u_{r,s+1}-u_{r,s}}{k} \\ \dfrac{u_{r,s+1}-2u_{r,s}+u_{r,s-1}}{k^2} \end{cases}$$

(316) $$\begin{cases} \dfrac{u_{r,s+1}-u_{r,s}}{k}=\kappa\dfrac{u_{r+1,s}-2u_{r,s}+u_{r-1,s}}{h^2} \\ u_{r,s+1}=u_{r,s}+\kappa\rho(u_{r-1,s}-2u_{r,s}+u_{r+1,s}) \end{cases}$$

(317) $$u_x=A(u)u_y \qquad u(0,y)=h(y)$$

(318) $$\int_\Phi f(x,y)\frac{\partial^{r+s}\phi}{\partial x^r\partial x^s}\,dx\,dy = (-1)^{r+s}\int_\Phi g(x,y)\phi(x,y)\,dx\,dy$$

(319) $$\begin{cases} u_{xx}+u_{yy}=f(x,y) \text{ in } \Omega \\ u=g(x,y) \text{ on } \partial\Omega \end{cases}$$

(320) $$\begin{cases} B(\phi,u)=-\displaystyle\int_\Omega (D_{1,0}\phi D_{1,0}u+D_{0,1}\phi D_{0,1}u)\,dx\,dy \\ B(\phi,u)=\displaystyle\int_\Omega \phi f\,dx\,dy \end{cases}$$

(321) $$\int_\Omega \phi(u_{xx}+u_{yy})\,dx\,dy=-\int_\Omega (\phi_x u_x+\phi_y u_y)\,dx\,dy$$

Again the treatment is restricted to two dimensions. The function $\phi(x, y)$ is said to be a test function if it vanishes outside a bounded region Φ of the xy-plane and if it is infinitely differentiable in that region. If f is locally integrable in a region Ω and if there exists a locally integrable function g such that for all test functions ϕ, the integral of $f(x, y)\partial^{r+s}\phi/\partial x^r\partial y^s$ over Φ equals the integral of $(-1)^{r+s}g(x, y)\phi(x, y)$ over Φ (see 318) g is called a generalized derivative of f and $g = D_{r,s}f$ is written.

Now the Dirichlet problem for Laplace's equation in two dimensions may be considered (see 319); in this instance, $\partial\Omega$ denotes the boundary of Ω. A generalized solution of the Dirichlet problem is defined by introducing a bilinear function $B(\phi, u)$, and $B(\phi, u)$ is defined for all test functions ϕ (see 320).

A function u is said to be a generalized solution of the stated Dirichlet problem if, and only if, $u = g$ on $\partial\Omega$ and if for every test function ϕ, $B(\phi, u)$ equals the integral of ϕf over Ω (see 320). On the other hand, a classical solution is a twice-differentiable function that assumes the correct boundary values; therefore, from Green's theorem (see 321), it follows that, if a classical solution to the stated Dirichlet problem exists, then it is also a generalized solution.

The theory of generalized solutions of partial differential equations forms what is known as the abstract theory of partial differential equations. (I.N.S.)

Special functions

Certain special functions are often used in mathematics and in physics, chemistry, engineering, and other branches of science and technology. It is possible to classify these into elementary functions and higher transcendental functions. The elementary functions include the exponential, logarithmic, trigonometric, and related functions. Frequently the term special functions is only applied to high transcendental functions. Some of the more important ones arising in differential equations will now be described.

The gamma and beta functions. The gamma function arose from attempts to extend the factorial function, $n! \equiv n(n-1)(n-2)\cdots 1$, to non-integral values of n. This function is defined by $\Gamma(x)$ being equal the integral from $t = 0$ to infinity of $e^{-t}t^{(x-1)}$ (see 322), for values of x greater than zero. The function has the property that $\Gamma(x + 1)$ equals the product of x and $\Gamma(x)$.

It follows that, if n is a positive integer, then $\Gamma(n + 1)$ equals $n!$. Furthermore, $\Gamma(\frac{1}{2})$ equals the square root of π (see 323).

The function was first defined by Euler and is widely used in mathematics, physics, engineering, and especially in probability theory and statistics. It arises in the series solutions of many differential equations.

Some useful related functions are:

(i) The *Incomplete Gamma functions*, which are defined by letting the upper or the lower limit of integration be a variable (see 324).

(ii) The *Beta function* $\beta(p, q)$, which is defined as being equal to the integral from zero to one of the product of t^{p-1} and $(1 - t)^{q-1}$ (see 325) for positive values of p and q. It is related to the gamma function because $\beta(p, q)$ equals the product of $\Gamma(p)$ and $\Gamma(q)$ divided by $\Gamma(p + q)$ (see 326).

Hypergeometric function. The hypergeometric function is the function $F(a, b; c; x)$ defined by the so-called hypergeometric series (see 327). It was known to Euler and gets its name from the fact that it is a generalization of the geometric series $(1 + x + x^2 + \cdots)$ and reduces to this series when $a = 1$ and $b = c$. It satisfies the hypergeometric equation, a second-order linear differential equation with variable coefficients (see 328). Twenty-four solutions of this equation can be expressed in terms of hypergeometric equations.

The hypergeometric equation occurs in problems in fluid flow and other branches of physics and engineering. The hypergeometric function is important in mathematics because many other special functions are related to it.

Legendre polynomials. Legendre polynomials arise as particular solutions of Laplace's equation: $\Delta_3 u = 0$. If R is the distance between a point with Cartesian coordinates (x, y, z) and the point $(0, 0, 1)$, then the reciprocal distance, $\frac{1}{R}$, is a solution of Laplace's equation. If spherical polar coordinates are used for the point (r, θ, ϕ), this solution is $1 - 2r\cos\theta + r^2$ to the $-\frac{1}{2}$ power. It is axisymmetric because it is independent of ϕ, the azimuthal angle.

If $\cos\theta$ is replaced by μ the solution is $(1 - 2r\mu - r^2)^{-\frac{1}{2}}$, and it can be expanded to give a power series in r with the coefficients being polynomials in μ. Thus it can be written in the form of an infinite series with terms of the form $r^n P_n(\mu)$ for r less than one (see 329).

The coefficients P_n are called the Legendre polynomials of degree n.

The polynomials are solutions of Legendre's equation (see 228), a particular case of Laplace's equation expressed in spherical polar coordinates for axisymmetric solutions. They can be produced from the generating function $(1 - 2r\mu - r^2)^{-1/2}$ and are explicitly given by Rodrigues' formula (see 330), which is named after a 19th-century French mathematician, Olinde Rodrigues.

Associated Legendre functions. These are functions satisfying the associated Legendre differential equation, a second-order linear equation with variable coefficients (see 331). This reduces to Legendre's equation when $m = 0$. Its solutions are associated Legendre functions (or polynomials). They are denoted by P_n^m and can be expressed in terms of the Legendre polynomials (see 332).

Spherical harmonics. The Legendre polynomials and associated Legendre functions are particular solutions of Laplace's equation. In general, functions satisfying this equation are known as harmonic functions and solutions that are homogeneous in x, y, and z are called spherical (solid) harmonics. If spherical polar coordinates (r, θ, ϕ) are used, a spherical (solid) harmonic of degree n, denoted R_n, can be written as the product of r^n and $S_n(\theta, \phi)$ (see 333), a function of θ and ϕ satisfying a second-order linear partial differential equation, which is found substituting R_n into Laplace's equation and applying the method of separation of variables (see 334). It is called a spherical surface harmonic. If $R_n(x, y, z)$ is a polynomial of degree n and n is a positive integer, then S_n is a polynomial in $\cos\theta$, $\sin\theta$, $\cos\phi$, and $\sin\phi$.

Zonal, sectoral, and tesseral harmonics. A polynomial spherical harmonic of degree n can be represented uniquely by the product $C_n r^{2n+1}$, in which C_n is a constant, times n differentiations $\partial/\partial h_1, \partial/\partial h_2, \cdots, \partial/\partial h_n$ of $1/r$, in which (see 335) $h_1, h_2, \cdots, h_n$ are n directions in space, and $\dfrac{\partial}{\partial h}$ denotes differentiation in a direction h. If the n directions coincide, then, on a unit sphere about the origin, there are n circles of latitude on which the spherical surface (or solid) harmonic vanishes. These divide the sphere into zones, and the spherical harmonic is known as a zonal harmonic. The curves on the unit sphere on which the surface harmonic vanishes are called nodal lines. If the n directions are situated on a plane and are at angles of $\dfrac{\pi}{n}$ to each other, the nodal lines are n circles of longitude. These divide the sphere into sectors, the spherical harmonic being known as a sectoral harmonic. If $(n - m)$ directions coincide in one particular direction and the remaining m directions lie in a plane at right angles to this direction and are at angles $\dfrac{\pi}{m}$, then there are m circles of longitude and $(n - m)$ circles of latitude. These divide the unit sphere into rectangles, the spherical harmonic being a tesseral harmonic.

It can be proved that spherical surface harmonics can be represented by the product $S_n^m(\theta, \phi)$ of the associated Legendre function $P_n^m(\cos\theta)$ and the complex exponential $\exp[im\phi]$ (see 336). Laplace's equation is satisfied by a linear combination of terms $R_n^m(x, y, z)$, which are products of r^n, $\exp[im\phi]$, and $P_n^m(\cos\theta)$ (see 337).

$P_n^m(\cos\theta)$ is an associated Legendre polynomial. If $m = n$, the harmonic is a sectoral harmonic. If $1 \leqslant m \leqslant n \leqslant 1$, it is a tesseral harmonic. If $m = 0$, it is a zonal harmonic; the Legendre polynomial $P_n(\cos\theta)$ is a zonal harmonic of degree n.

Spherical harmonics and surface harmonics arise in a variety of physical applications of Laplace's equation. Surface harmonics also arise in connection with Poisson's equation (see 226), the wave equation (see 235), the diffusion equation (see 241), Schrödinger's equation (see 249), and other partial differential equations.

Bessel functions. Just as the Laplace equations expressed in spherical polar coordinates lead to the Legendre equation and Legendre polynomials, its expression in cylindrical polar coordinates leads to the equation known as Bessel's equation. Bessel's equation of order ν is a second-order differential equation with variable coefficients (see 338) and is satisfied by the Bessel function (or Bessel function of the first kind), which can be expressed as a power series in power of $t/2$ (see 339). $J_\nu(t)$ is the Bessel function of order ν. Such functions were used by the German astronomer Friedrich Wilhelm Bessel in the 19th century in problems in astronomy, although they had been used earlier by Bernoulli and Euler.

A second solution of the equation, if ν is not an integer, is $J_{-\nu}(t)$ leading to the general solution $AJ_\nu(t) + BJ_{-\nu}(t)$, A and B being constants.

If ν is an integer, p, the second solution is denoted by $Y_p(t)$, and a general solution is $AJ_p(t) + BY_p(t)$. For non-integral and zero values of ν, $Y_\nu(t)$ can be expressed as the difference of $J_\nu(t)\cos(\nu\pi)$ and $J_{-\nu}(t)$, divided by $\sin(\nu\pi)$ (see 340), and is called a Bessel function of the second kind. It can be shown that if ν tends to an integer p (or to zero) in the limit, the expression gives $Y_p(t)$ (or

(322) $$\Gamma(x) = \int_0^\infty e^{-t} t^{(x-1)}\, dt$$

(323) $$\begin{cases} \Gamma(x+1) = x\Gamma(x) \\ \Gamma(n+1) = n! \\ \Gamma(\frac{1}{2}) = \sqrt{\pi} \end{cases}$$

(324) $$\begin{cases} \gamma(x, z) = \displaystyle\int_0^z e^{-t} t^{x-1}\, dt \\ \Gamma(x, z) = \displaystyle\int_z^\infty e^{-t}\, t^{x-1}\, dt \end{cases}$$

(325) $$\beta(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt$$

(326) $$\beta(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

(327) $$\begin{cases} 1 + \dfrac{abx}{c \cdot 1} + \dfrac{a(a+1)b(b+1)x^2}{c(c+1)2!} + \\ \quad + \dfrac{a(a+1)(a+2)\, b(b+1)(b+2)x^3}{c(c+1)(c+2)3!} + \cdots \end{cases}$$

(328) $$x(1-x)\frac{d^2y}{dx^2} + [c - (a+b+1)x]\frac{dy}{dx} - aby = 0$$

(329) $$\begin{cases} \dfrac{1}{\sqrt{1 - 2r\cos\theta + r^2}} = 1 + r\mu + \tfrac{1}{2}r^2(3\mu^2 + 1) + \cdots \\ \qquad\qquad = \displaystyle\sum_0^\infty r^n P_n(\mu) \end{cases}$$

(330) $$P_n(\mu) = \frac{1}{2^n n!}\frac{d^n}{d\mu^n}(\mu^2 - 1)^n$$

(331) $$(1-\mu^2)\frac{d^2\omega}{d\mu^2} - 2\mu\frac{d\omega}{d\mu} + \left\{(n+1) - \frac{m^2}{1-\mu^2}\right\}\omega = 0$$

(332) $$P_n^m(\mu) = (-1)^m(1-\mu^2)^{\frac{m}{2}}\frac{d^m}{d\mu^m}P_n(\mu)$$

(333) $$R_n(x, y, z) = r^n S_n(\theta, \phi)$$

(334) $$\frac{\partial^2 S}{\partial\theta^2} + \cot\theta\frac{\partial S}{\partial\theta} + \csc^3\theta\frac{\partial^2 S}{\partial\phi^2} + n(n+1)S = 0$$

(335) $$c_n \frac{r^{2n+1}\partial^n}{\partial h_1 \partial h_2 \cdots \partial h_n}\frac{1}{r}$$

$Y_0(t)$; see 341). Thus $AJ_\nu(t) + BY_\nu(t)$ is a general solution for all values of ν. Bessel functions of the second kind are sometimes called Neumann functions. The functions $H_\nu^1(t)$, defined as the sum of $J_\nu(t)$ and $iY_\nu(t)$, and $H_\nu^2(t)$, defined as their difference (see 342) are called Hankel functions, or Bessel functions of the third kind.

If t is replaced by it (in which $i = \sqrt{-1}$), the differential equation is the modified Bessel equation and its solutions are modified Bessel functions (of the first and second kind) of which $I_\nu(t)$, the modified Bessel function of the first kind, is the product of $(i)^{-\nu}$ and $J_\nu(it)$, and $K_\nu(t)$,

(336) $$S_n^m(\theta, \phi) = e^{im\phi} P_n^m(\cos\theta)$$

(337) $$R_n^m(x, y, z) = r^n e^{im\phi} P_n^m(\cos\theta)$$

(338) $$\frac{d^2\omega}{dt^2} \frac{1}{t} \frac{d\omega}{dt} + \left(1 - \frac{\nu^2}{t^2}\right) = 0$$

(339) $$J_\nu(t) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\Gamma(\nu + n + 1)} \left(\frac{t}{2}\right)^{\nu + 2n}$$

(340) $$Y_\nu = \frac{J_\nu(t)\cos(\nu\pi) - J_{-\nu}(t)}{\sin(\nu\pi)}$$

(341) $$Y_p(t) = \lim_{\nu \to p} \frac{J_\nu(t)\cos(\nu\pi) - J_{-\nu}(t)}{\sin(\nu\pi)}$$

(342) $$\begin{cases} H_\nu^{(1)}(t) = J_\nu(t) + iY_\nu(t) \\ H_\nu^{(2)}(t) = J_\nu(t) - iY_\nu(t) \end{cases}$$

(343) $$\begin{cases} I_\nu(t) = (i)^{-\nu} J_\nu(it) \\ K_\nu(t) = \frac{1}{2}\pi(\sin\nu\pi)^{-1} [I_{-\nu}(t) - I_\nu(t)] \end{cases}$$

(344) $$H_n(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n}$$

(345) $$\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + 2ny = 0$$

(346) $$\int_a^b P_n(x)\, P_m(x)\, W(x)\, dx = 0, \qquad m \neq n$$

(347) $$\int_a^b [P_n(x)]^2\, W(x)\, dx = 1$$

(348) $$L_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x})$$

(349) $$x\frac{d^2y}{dx^2} + (1 - x)\frac{dy}{dx} + nx = 0$$

the modified Bessel function of the second kind, is defined in terms of $I_\nu(t)$ and $(\sin\nu\pi)^{-1}$ (see 343). If ν is an integer p, $K_p(t)$ is the limit of the expression.

Bessel functions are widely used in physics, especially in problems in potential theory and diffusion. They frequently arise in problems involving cylindrical boundaries and are sometimes called cylinder functions.

Hermite polynomials. Hermite polynomials are defined in terms of the derivatives of $\exp(-x^2)$ for nonnegative integers n (see 344). They are solutions of Hermite's equation, named after the 19th-century mathematician Charles Hermite. This is a reduced form of Schrödinger's equations and is used in wave mechanics, especially in the theory of wavefunctions of the harmonic oscillator. Hermite's equation, also, is a second-order differential equation, with the coefficient of the first derivative being $-2x$ (see 345).

Orthogonal polynomials. Hermite polynomials are examples of orthogonal polynomials. A system of polynomials P_n is orthogonal on the interval (a, b) with respect to a weight function $w(x)$ if the integral over the interval of the product of two elements of the system, $P_n(x)$ and $P_m(x)$, and the weight $w(x)$, equals zero whenever m and n are not equal (see 346). The system is orthonormal if it is orthogonal and normalized by the integral over the interval of the product of the square of $P_n(x)$ and $w(x)$ being equal to unity (see 347).

Hermite polynomials are orthogonal on the interval $(-\infty, \infty)$ with a weighting function of $\exp(-x^2)$. Other systems of orthogonal polynomials exist. For example, Legendre polynomials are orthogonal in the interval $(-1, 1)$ with a weighting factor of 1. The Laguerre polynomials, the nth of which is defined as the product of e^x and the nth derivative of $x^n e^{-x}$ (see 348), are orthogonal over the interval $(0, \infty)$ with a weighting factor of e^{-x} (sometimes more general functions with a weighting factor of $x^a e^{-x}$ are called Laguerre polynomials). They are solutions of Laguerre's equation (see 349). This, like Hermite's equation, is used in wave mechanics. The associated Laguerre polynomials $L_n^m(x)$ are defined by $L_n^m(x)$ being equal to the mth derivative of the nth Laguerre polynomial (see 350). They satisfy the associated Laguerre equation (see 351). (Ed.)

Dynamical systems on manifolds

A dynamical system is a way of describing the passage in time of all points of a given space $\mathfrak{S}$. $\mathfrak{S}$ can be thought of, for example, as the space of states of some physical system. Then if x is in $\mathfrak{S}$, one unit in time later x will have moved to a point denoted by x_1. At time zero x is at x or x_0. Two units after time zero x will have moved to x_2. One unit before time zero x was at x_{-1}. If this procedure is extrapolated to fill up the real numbers R, the trajectory x_t is obtained, for all time t, a real number. Thus for each x, x_t is a curve in $\mathfrak{S}$ and represents the life history of x as t goes from $-\infty$ to ∞. If x_t is also assumed to be differentiable in (t, x), then $x \to x_t$ can be regarded as a transformation, or diffeomorphism, from $\mathfrak{S}$ onto $\mathfrak{S}$ for each t. In this case $(\mathfrak{S}, x_t)$ define a dynamical system.

In applications of dynamical systems to fields outside of mathematics, the first goal is to explicate the manifold of states $\mathfrak{S}$. A state of a system is information characterizing the situation at a given moment. The first step in this explication process is to define the state in the most obvious unrestricted way. This will give a space that is possibly too large, and natural constraints on physical laws will cut down the unrestricted states to a subset of attainable or physical states. A physical state is one that has an actual possibility of occurring. These physical states form the final manifold of states. A number of examples in the following section will explain the process given in the previous sentences.

Definition of a state of a system

EXAMPLES OF STATE SPACES

Example from economics. In the pure exchange economy of theoretical economics, a model is used with n different commodities, each measured in quantity by a positive real number (fixing a unit of measurement). Thus commodity space P will be the set of points of real Cartesian space $\mathfrak{R}^n$ with each coordinate positive. A point of P represents a bundle of commodities possessed, for example, by a certain consumer in the economy.

It is assumed that there are a finite number of consumers, say m, with the possessions of the ith consumer denoted by x_i in P. An unrestricted state of the economy is a point $x = (x_1, \cdots, x_m)$ in which each x_i is in P. The space of all these states is denoted by P^m, the Cartesian product of P with itself m times. Thus an unrestricted state of the economy simply gives the possessions of the consumers of that economy.

Now, it is supposed that the total resources in this economic model are fixed, say, at a point w in P. Thus the space of attainable states is a subset W of P^m described by the condition $(x_1, \cdots, x_m)$ is in W if the sum of the x_i is w.

Examples from classical mechanics. *Particle in Euclidean space E^3.* Ordinary 3-dimensional Euclidean space is denoted by E^3. A particle in E^3 has its state characterized by its position x, a point in E^3, and its velocity $\mathbf{v}$, a vector (or point) in E^3. Thus $\mathbf{v}$ gives the direction in which the particle is moving together with its speed, the length of $\mathbf{v}$. The vector $\mathbf{v}$ can be thought of as being based at x. The states of this system are then all pairs $(x, \mathbf{v})$, x, $\mathbf{v}$ each in E^3. This space is denoted by $E^3 \times E^3$, the Cartesian product of E^3 with E^3.

Particle on a sphere. D^3 is a unit ball in E^3, so that D^3 consists of all points the distance of which from the origin is less than or equal to one. The 2-sphere S^2 is its boundary, or points of unit distance, from the origin of E^3. The space of states for the physical system of a particle on S^2 will then be a certain subset of $E^3 \times E^3$ of the first example.

There are two constraints: the first is that x must be in S^2, and the second is that $\mathbf{v}$ must be perpendicular to x. The first constraint follows by definition, and the second constraint follows because if the velocity were not perpendicular to x, then x would have to leave the 2-sphere. The set of points $(x, \mathbf{v})$ that satisfies these two constraints is the state space $\mathfrak{S}$ for this problem. $\mathfrak{S}$ can also be thought of in the following way. If $T_x(S^2)$ is the tangent space to S^2 at x, so $T_x(S^2)$ is the set of all vectors $\mathbf{v}$ in E^3 based at x with $\mathbf{v}$ tangent to the surface S^2 (or perpendicular to x), then $\mathfrak{S}$ is the union of all these $T_x(S^2)$ as x varies over S^2. In other words, $\mathfrak{S}$ is the tangent bundle of S^2, $T(S^2)$ or the set of all $(x, \mathbf{v})$ in $S^2 \times E^3$ in which $\mathbf{v}$ is in $T_x(S^2)$.

Particle on any surface in E^3. The previous example extends immediately to the case of a particle constrained to move on any (smooth) surface M in Euclidean space E^3. The sphere S^2 is replaced by M everywhere in the discussion. Thus, for example, the space $\mathfrak{S}$ of all states for this system is $T(M)$ or the set of $(x, \mathbf{v})$ in $M \times E^3$ in which $\mathbf{v}$ is tangent to M at x. $T(M)$ is called the tangent bundle of M and may be represented as the union of all the fibres $T_x(M)$ as x ranges over the points of M. $T_x(M)$ in the space of all of the possible velocities for the particle at x.

Plane rigid body. A point x_0 and a directed line n_0 are fixed in B through x_0. Then the position of B in the plane E^2 is characterized by the position of x_0 in E^2 and the angle between n_0 and a fixed directed reference line n in E^2. (For simplicity it is assumed that B has only one possibility for its orientation in E^2.) Therefore, the space of positions or configurations of B is a three-dimensional space or manifold, the Cartesian product of E^2 and the circle S^1, $E^2 \times S^1$. In this instance, the fact that there is a correspondence between angles and points on the circle is used. $M = E^2 \times S^1$ is sometimes called the space of generalized coordinates for the configuration of B. If (x, θ) is in $E^2 \times S^1$, then (x, θ) stands for the configuration of B, in which x_0 is located at x and θ is the angle between n_0 and n.

The possible velocities for B at (x, θ) are vectors $(\mathbf{v}_1, \mathbf{v}_2)$ in $E^2 \times \mathfrak{R}$ (or E^3), in which $\mathbf{v}_1$ is the ordinary velocity of x in E^2 and $\mathbf{v}_2$ is angular velocity. Therefore, for this mechanical system, the state space $\mathfrak{S}$ is the product $(E^2 \times S^1) \times (E^2 \times \mathfrak{R})$ of configuration and velocity space.

A circuit with three electrical components. A series connection between a resistor, an inductance, and a capacitor, wired to form a closed loop, may now be considered. To form the space of states, each of these three components is oriented. A state of the circuit consists of the current through each component and the voltage across each component. If i_ρ, v_ρ stand for the current and voltage, respectively, in the resistor, then i_λ, v_λ stand for current and voltage in the inductance and i_γ, v_γ for current and voltage in the capacitor. For historical reasons the capacitor is conventionally oriented in the opposite direction from the other two components. The Kirchhoff laws imply the following relations on the currents and voltages. The Kirchhoff current law (KCL) asserts that $i_\lambda = i_\rho = -i_\gamma$ or that the current that flows into a junction between two components equals that flowing out.

The Kirchhoff laws and Ohm's law

The Kirchhoff voltage law (KVL) asserts that $v_\lambda + v_\rho = -v_\gamma$, or that the voltages around a loop of components sum to zero.

The set of states in $\mathfrak{R}^6$ satisfying these Kirchhoff laws forms a three-dimensional linear subspace K of $\mathfrak{R}^6$. Any physical state has to lie in K. To obtain the precise set of physical states, another step is needed, in order to bring in the implications of the resistors on (i_ρ, v_ρ).

The most common type of resistor is linear in that it obeys Ohm's law; *i.e.*, $v_\rho = R_\rho i_\rho$, in which R_ρ is some positive constant. For a physical state, Ohm's law must be satisfied as well as Kirchhoff's laws. This cuts down the space of physical states to a two-dimensional linear subspace $\mathfrak{S}$ for $\mathfrak{R}^6$.

More generally, any resistor, linear or nonlinear, is defined by its characteristic Λ_ρ, which is a curve in the two-dimensional (i_ρ, v_ρ)-plane. Then a generalized version of Ohm's law stipulates that for a physical state in $\mathfrak{R}^6$, the components (i_ρ, v_ρ) lie in Λ_ρ. It can be proved in the general case that these constraints force the physical states to lie in a two-dimensional manifold Σ of $\mathfrak{R}^6$.

Suppose, for example, that the resistor is current controlled so that the characteristic Λ_ρ is the graph in $\mathfrak{R}^2$ of some smooth real function f. Thus $v_\rho = f(i_\rho)$. In this case it is convenient to take as a representation of Σ, the Cartesian (i_λ, v_γ)-plane $\mathfrak{R}^2$ identifying $\mathfrak{R}^2$ and Σ under the diffeomorphism $\mathfrak{R}^2 \to \Sigma \subset \mathfrak{R}^6$, which sends (i_λ, v_γ) into a state in $\mathfrak{R}^6$ that is completely determined by i_λ and v_γ (see 352). Then the space of states for this circuit is equivalent to $\mathfrak{R}^2$ or (i_λ, v_γ)-space.

Simple electrical circuits. This example just given can be generalized to what could be called simple electrical circuits. A simple electrical circuit consists of components and nodes. The components are circuit elements that are either resistors, inductances, or capacitors, each with two terminals. In the circuit, groups of terminals meet at the different nodes. It is assumed that the components have a given orientation so that the current flowing in one direction is given by a positive real number, and the current flowing is given by a negative number if it flows in the opposite direction.

The unrestricted states of the current have as components the currents through each component and the voltages across each component. So if there are N components, the space of unrestricted states $\mathfrak{S}$ is $2N$-dimensional Cartesian space $\mathfrak{R}^{2N}$.

The laws of physics impose constraints on these states. First of all, Kirchhoff laws assert that a physical state must be in a linear subspace K of $\mathfrak{R}^{2N}$ with dimension K equal N. In formal terms Kirchhoff's current law says that for each node, the sum of the currents entering the node equals the sum of the currents issuing from the node. Kirchhoff's voltage law then requires that the voltages across each component of a closed cycle add to zero. It follows that the set of states in $\mathfrak{R}^{2N}$ satisfying Kirchhoff's current law and Kirchhoff's voltage law forms a linear subspace of dimension N.

The final physical constraint is given by a generalization of Ohm's law. A resistor component ρ is allowed to have as characteristic any non-singular curve—*i.e.*, one-dimensional submanifold—Λ_ρ in the (i_ρ, v_ρ)-plane. This means that a physical state in K must have i_ρ, v_ρ components satisfying the condition that (i_ρ, v_ρ) is in Λ_ρ. If the ρth resistor is linear, then Λ_ρ is a line through the origin with positive slope. Thus the physical states finally satisfy these further conditions, one for each resistor. The final subspace Σ of $\mathfrak{S}$ defined by these conditions and Kirchhoff conditions will be a submanifold in the general case the dimension of which is equal to the number of inductor components plus the number of capacitor components.

Miscellaneous examples. The 19th–20th-century Italian mathematician Vito Volterra has studied the growth and decline of two interacting species; for example, rabbits and foxes. The population of each species can be represented by a real positive number, subject to a slight approximation. Then the corresponding state space is the full positive quadrant in the plane—*i.e.*, the set of pairs (r, f) of positive real numbers r and f.

In Turing's models in biology, the states appear as the concentrations of a number of chemicals (or morphogens) in each cell of some structure. Sometimes, as in hydrodynamics, the space of states appears as an ∞-dimensional space (a manifold or linear space).

MANIFOLDS

In each case of the examples, the state space is a manifold. A manifold can often be thought of as a domain (or open set) in Euclidean space of some dimension. More generally a manifold is defined by glueing together such open sets; thus, for example, while the 2-sphere is a 2-dimensional manifold, it is not equivalent to an open set of the plane. It can be obtained though as the union of two, 2-dimensional disks that overlap in a ring. Every k-dimensional manifold can be represented by a k-dimensional surface in Euclidean space of some higher dimension. It is most reasonable to take a manifold—*i.e.*, differentiable manifold—as the basic space of a dynamical system.

If U is an open set in Euclidean space E and x is in U, then a tangent vector of U at x is simply a vector in E, which can be thought of as being based at x. The tangent bundle $T(U)$ is the product $U \times E$, and a point $(x, \mathbf{v})$ in $U \times E$ consists of base point x, vector $\mathbf{v}$.

More generally if x is a point in a manifold M, in which M is represented as a k-dimensional surface in some Euclidean space E, then a tangent vector of M at x is simply a vector in E based at x and tangent to M. The space of all tangent vectors to M at x is the tangent space $T_x(M)$ of M at x. Then $T(M)$, the tangent bundle of M, is the (disjoint) union of all these $T_x(M)$ as x varies over M, and is a manifold itself in a natural way.

In all the examples of mechanical systems, the initial step was to take the manifold M of positions or configurations. Then the space of states was the tangent bundle $T(M)$ of M. A state in all of these examples consists of a configuration together with a (generalized) velocity based at that configuration. This persists in much more general examples of mechanical systems.

DYNAMICAL SYSTEMS AS VECTOR FIELDS ON MANIFOLDS

On manifolds (*e.g.*, open sets of Euclidean space) differentiation makes sense, and a differentiable map (*i.e.*, transformation) from a manifold to itself with a differentiable inverse is called a diffeomorphism. In these terms the definition of a dynamical system can be restated as assigning to each time t (a real number), a diffeomorphism $\phi_t : M \to M$ with ϕ_0 the identity, $\phi_{t+s} = \phi_t \circ \phi_s$, and so that $\phi_t(x)$ is differentiable in (t, x). Thus for each x in the manifold, the trajectory $\phi_t(x) = x_t$ is a curve through x. The derivative of this curve $t \to \phi_t(x)$ at $t = 0$ can be taken, letting $X(x)$ be the derivative of $\phi_t(x)$ (see 353) to obtain the velocity or the tangent vector of the curve at x. Thus $X(x)$ is in the tangent space $T_x(M)$. The association $x \to X(x)$ is a vector field on M (or tangent vector field on M). More generally, a vector field on M is any differentiable map $X : M \to T(M)$ with $X(x)$ in $T_x(M)$. For example, if M is an open set in Euclidean space E, then $T(M) = M \times E$ and a vector field $Y : M \to M \times E$ must be of the form $Y(x) = (x, X(x))$ for some map $X : M \to E$. In this case by a slight abuse of language, X is called a vector field, so that a vector field can be thought of as assigning to each point x, in M, a vector in E based at x.

Diffeomorphisms

The notion of vector field on a manifold is of basic importance because among other things a dynamical system is characterized by its associated vector field. A converse to this statement is roughly true, and the converse could be regarded as the fundamental theorem of ordinary differential equations. More precisely, if $x \to X(x)$ is a vector field on a manifold M, the ordinary differential equation can be considered for $\phi_t(x)$, given by the derivative of $\phi_t(x)$ being equal to $X(x)$ (see 354).

This equation has a solution, at least for all t satisfying $t^-(x) < t < t^+(x)$ in which t^-, t^+ are real functions on M defining some (perhaps maximal) open set Q in $R \times M$ containing $O \times M$. This solution $\phi_t(x)$, defined for (t, x) in Q, then has the properties of a dynamical system except that it may not be defined for all t going to $\pm\infty$. This system of trajectories is still called a dynamical system (or flow), and in this sense vector fields on manifolds are equivalent to dynamical systems.

(350) $$L_n^m(x) = \frac{d^m}{dx^m} L_n(x)$$

(351) $$x\frac{d^2y}{dx^2} + (m+1-x)\frac{dy}{dx} + (n-m)y = 0$$

(352) $$(i_\lambda, i_\lambda, -i_\lambda, f(i_\lambda), v_\gamma - f(i_\lambda), v_\gamma) = (i_\rho, i_\lambda, i_\gamma, v_\rho, v_\lambda, v_\gamma)$$

(353) $$\left.\frac{d\phi_t(x)}{dt}\right]_{t=0} = X(x)$$

(354) $$\frac{d}{dt}\phi_t(x) = X(x)$$

EXAMPLES OF DYNAMICAL SYSTEMS

Particle in E^3 in a force field. Consider the example of a particle in three-dimensional Euclidean space. It is supposed now that there is a certain vector field, a force field defined on E^3, say $x \to F(x)$, $E^3 \to E^3$. Thus if the particle is at x in E^3, the force exerted on it is $F(x)$, and Newton's law reads $ma = F(x)$. Here m is a constant positive real number, the mass of the particle, and $a = x''(t)$ (or the second derivative of x with respect to t) is the acceleration with $t \to x(t)$ a curve in E^3 that represents the passage of the particle in time. For an actual or physical path, Newton's equation is satisfied, or $mx''(t) = F(x(t))$ for all t.

Newton's equation is seen to be equivalent to the pair of equations (see 355) relating the derivatives of position and velocity to the velocity and the force, respectively. In this last form, the map $(x, \mathbf{v}) \to (\mathbf{v}, F(x)/m)$ is a vector field on the state space for this mechanical system $T(E^3)$. This is the form of a vector field on the state space determining the passage in time of the space. The initial conditions are $(x, \mathbf{v})$, the position and velocity of the particle.

Conservative force fields

The force field $x \to F(x)$ is called conservative if it is of the form $F(x) = -\text{grad } V(x)$ in which V is some real function on E^3. Grad $V(x)$ is defined as the vector in which the ith component is $\partial v/\partial x_i$, for $i = 1, 2, 3$ (see 356), in terms of Cartesian coordinates. V is called the potential for the system. The classical example of a force field is that of gravitational force. If the source of this force is the point located at the centre of E^3, then (see 357) the force is inversely proportional to the square of the distance, involving the Euclidean norm on E^3 (see 358) in Cartesian coordinates. In this case, $F(x)$ is a conservative force field because a potential can be chosen (see 359) that is inversely proportional to the distance. This gives what is sometimes called the Kepler problem because the solution curves lead to the (Kepler) elliptical orbits of the Earth about the Sun. Strictly speaking, in the Kepler problem configuration space is $E^3 - 0$, because the vector field describing the dynamics is not defined at 0.

Van der Pol equation. The particular example of an electrical circuit given earlier is now considered. Physical laws of the inductor and capacitor give directly the differential equations for motion in terms of the representation of the state space found there. More precisely (see 360) the derivative of i_λ is proportional to v_λ and the derivative of v_γ is proportional to i_γ (see 360), in which L, the inductance, and C, the conductance, are constants of proportionality. For simplicity, L and C are taken as constants equal to unity. Here $v_\lambda = v_\gamma - f(i_\lambda)$ and $i_\gamma = -i_\lambda$ by Kirchhoff's laws and the resistor characteristic. The differential equations then become appropriately changed (see 361). The corresponding vector field on $\mathfrak{R}^2$ at the point (i_λ, v_γ) is $(v_\gamma - f(i_\lambda), -i_\lambda)$.

In the case in which f is linear, of the form $v_p = R_p i_p$, $R_p > 0$, then it can be deduced that all solution curves of this dynamical system tend toward the origin. The origin is a steady state and an attractor. This means that a solution at the origin stays at the origin for all time or that the vector field is zero at the origin (steady state). That the origin is an attractor means that nearby solutions tend to the origin as $t \to \infty$. Steady-state attractors play a central role in applications of dynamical systems.

Steady-state attractors

On the other hand, if a highly nonlinear characteristic is considered, say the derivative $f'(0) < 0$, then the unique zero $(0, f(0))$ of the vector field can be shown to be a source. This means that orbits tend to $(0, f(0))$ as $t \to -\infty$. If the characteristic $(i_\lambda, f(i_\lambda))$ stays well inside the first and third quadrants for large $|i_\lambda|$ (a natural assumption even for nonlinear resistors), then as $t \to \infty$ all orbits tend toward some bounded region in $\mathfrak{R}^2$. This follows from a general energy theorem. From these facts, the Poincare-Bendixon theorem implies the existence of a periodic orbit, and one will be an attractor in the general case. A periodic orbit, period $w > 0$ is an orbit where $\phi_{t+w}(x) = \phi_t(x)$ for all t. This amounts to an oscillatory behaviour of the physical variables. That it is an attractor means that all nearby solutions tend toward it and thus will also oscillate, at least to a high degree of approximation. Periodic attractors are of basic importance in applications of dynamical systems.

The Poincaré-Bendixon theorem

The characteristic defined (see 362) by $f(i_\lambda)$ being equal to $\mu(i_\lambda^3 - i_\lambda)$ satisfies the conditions of the previous paragraph. Here there is a unique periodic attractor; the differential equation is called van der Pol's equation.

Volterra-Lotka equations. The equations describing how the populations of rabbits and foxes vary in time are called the Volterra-Lotka equations and are (see 363) a pair of coupled first-order equations with positive constants α, β, λ, μ. For example, $dr/dt = \alpha r$ governs how the number of rabbits increases eating grass, without taking into account the foxes that eat them. The Volterra-Lotka equation can be interpreted as giving a vector field on the positive quadrant in $\mathfrak{R}^2$; its solution curves are in fact periodic so that the populations are expected to change in cycles under these conditions.

GRADIENT DYNAMICAL SYSTEMS

Systems on $\mathbf{E^3}$. A gradient dynamical system in E^3 is considered. For a classical description, it is supposed $f: E^3 \to \mathfrak{R}$ is a differentiable function and E^3 is provided with Cartesian coordinates $(x_1, x_2, x_3) = x$. Then $X(x) = \text{grad } f(x)$ is the vector field on E^3 that associates to the point x in E^3 the vector grad $f(x)$, defined in terms of the partial derivatives of f (see 364). Thus grad $f(x)$ can be thought of as a vector in E^3 based at x, and it will be perpendicular to the level surface $f^{-1}(c)$ at x, $f(x) = c$. As a particular example, take $f(x)$ equal to the sum of three terms $a_i x_i^2$ (see 365), in which each a_i is a negative constant. Then the differential equation $dx/dt = \text{grad } f(x)$ can be solved directly to see that each trajectory tends to the steady-state attractor $(0, 0, 0) = 0$ in E^3.

Open sets on $\mathbf{E^n}$. The last example can be generalized in several steps to gradient dynamical systems on abstract manifolds. A further step is to the case of a real differentiable function $f: U \to \mathfrak{R}$ with U an open set of n-dimensional Euclidean space E. Given Cartesian coordinates for E (see 366), grad $f(x)$ can be defined as before by assigning to each x in U, the vector in E the ith component of which is the partial derivative of f with respect to x_i (see 367). Thus grad f is obtained as a map; *i.e.* grad $F: U \to E$.

It is convenient to give a second equivalent description of grad f that extends to manifolds that leads to a modern exposition of Hamiltonian mechanics.

This requires a precise definition of what is meant by Euclidean space. Euclidean space is a finite-dimensional real vector space E together with an inner product, say B. B is first of all a bilinear form on E so that B is a map $B: E \times E \to \mathfrak{R}$ satisfying conditions of linearity (see 368) for x, $\mathfrak{L}(y)$, y, $\mathfrak{L}(y)$ in E and λ real. Moreover, B is symmetric so that $B(x, y) = B(y, x)$. Finally, B is positive definite so that $B(x, x) \geqslant 0$ and $B(x, x) = 0$ if, and only if, $x = 0$. From B a norm can be defined, $||x|| = B(x, x)$ and distance from x to y as $||x - y||$. A Cartesian coordinate system on E can be found (see 369) so that $B(x, y)$ is a sum of terms of the form $x_i y_i$.

Euclidean space

The derivative of $f: U \to \mathfrak{R}$ at x in U is now considered. This derivative of f at x is most naturally considered to be a linear map $Df(x): E \to \mathfrak{R}$ with the property that $B(\text{grad } f(x), w) = Df(x)(w)$ for all w in E. Depending on the approach, this could be the defining property for $Df(x)$ or for grad $f(x)$. In modern treatments of calculus, however, the definition of derivative is given as a linear transformation, so that the above property can be used to define grad f. In more detail, the linear space of maps from E to $\mathfrak{R}$ is denoted by E^* (the dual of E). Then the derivative of f is a map $Df: U \to E^*$. A bilinear form B on E defines a natural induced map $\phi_B: E \to E^*$ by $\phi_B(x)(y) = B(x, y)$ for x, y in E. For the inner product this map ϕ_B is seen to be an isomorphism, or in other words, B is nondegenerate. In these terms the definition is, grad $f(x) = \phi_B^{-1} Df(x)$. These arguments form the basis for the modern approach to gradient dynamical systems and Hamiltonian systems.

Two important properties of grad f follow. First grad $f(x) = 0$ if, and only if, the derivative $Df(x)$ of f at x is 0 (or x is a critical point for f). So x is a steady state for the dynamical system grad f if, and only if, x is a critical point for the function f. Second, along a trajectory $\phi_t(x)$ of grad f, f is nondecreasing and in fact increases if x is not a critical point (see 370).

Thus, following a trajectory of grad $f(x)$, the function is maximized if possible, and in this way gradient systems are used in maximization problems.

Surfaces. Next a surface M in E^3 is considered and a differentiable function $f: M^2 \to \mathfrak{R}$. If x is in M, the inner product B on E^3 restricts to give an inner product B_x

(355) $$\dot{x} = \mathbf{v}, \qquad \dot{\mathbf{v}} = \frac{1}{m} F \quad \text{on} \quad E^3 \times E^3 = T(E^3)$$

(356) $$\left(\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}, \frac{\partial V}{\partial x_3}\right)$$

(357) $$F(x) = -\frac{x}{\|x\|^3}$$

(358) $$(\|x\| = (\Sigma x_i^2)^{1/2}$$

(359) $$V(x) = -\frac{1}{\|x\|}$$

(360) $$L\frac{di_\lambda}{dt} = v_\lambda, \qquad C\frac{dv_\gamma}{dt} = i_\gamma$$

(361) $$\begin{cases} \dfrac{di_\lambda}{dt} = v_\gamma - f(i_\lambda) \\ \dfrac{dv_\gamma}{dt} = -i_\lambda \end{cases}$$

(362) $$f(i_\lambda) = \mu(i_\lambda^3 - i_\lambda), \qquad \mu > 0$$

(363) $$\frac{dr}{dt} = \alpha r - \lambda rf, \qquad \frac{df}{dt} = -\beta f + \mu rf$$

(364) $$\left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \frac{\partial f}{\partial x_3}(x)\right) = \text{grad } f(x)$$

(365) $$f(x) = a_1 x_1^2 + a_2 x_2^2 + a_3 x_3^2$$

(366) $$(x_1, \cdots, x_n) = x$$

(367) $$\left(\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_n}(x)\right)$$

(368) $$\begin{cases} B(x+\mathfrak{L}[x], y) = B(x, y) + B(\mathfrak{L}[x], y) \\ B(x, y+\mathfrak{L}[y]) = B(x, y) + B(x, \mathfrak{L}[y]) \\ B(\lambda x, y) = \lambda B(x, y) = B(x, \lambda y) \end{cases}$$

(369) $$B(x, y) = \sum_{i=1}^{n} x_i y_i$$

(370) $$\frac{d}{dt}\phi_t(x)\Big]_{t=0} = Df(x)(\operatorname{grad} f(x)) = B(\operatorname{grad} f(x), \operatorname{grad} f(x)) \geqslant 0$$

(371) $$J(x, \bar{x}) = x_1 \mathfrak{L}(x_1) - x_2 \mathfrak{L}(x_2)$$

(372) $$P(i_\lambda, v_\gamma) = -i_\lambda v_\gamma + \int_0^\lambda f(u)\, du$$

(373) $$\frac{di_\lambda}{dt} = -\frac{\partial P}{\partial i_\lambda}, \qquad \frac{dv_\gamma}{dt} = \frac{\partial P}{\partial v_\gamma}$$

(374) $$\frac{dx}{dt} = \operatorname{grad}_J P(x)$$

(375) $$B(x, \mathfrak{L}[x]) = \sum_{i=1}^{n} x_i \mathfrak{L}(x_i)$$

(376) $$\Omega(x, \mathfrak{L}[x]) = \sum_{i=1}^{n} x_i \mathfrak{L}(x_{i+n}) - \sum_{i=1}^{n} x_{i+n} \mathfrak{L}(x_i)$$

Riemannian structures

on the tangent space $T_x(M)$. The assignment $x \to B_x$ is called a Riemannian structure on M. With this structure, the constructions of the previous section carry over. Thus for each x, $Df(x)$ is an element of the dual space $T_x(M)^*$ (or $T_x{}^*(M)$). This map $x \to Df(x)$ [or $df(x)$ oftentimes] is an example of a 1-form on M. Then an isomorphism $\phi_{Bx}: T_x(M) \to T_x{}^*(M)$ is defined for each x via B_x just as ϕ_B before. The inverse image of $Df(x)$ under ϕ_{Bx} is then grad $f(x)$ for each x, and the previous properties of gradients follow in the same way.

A special case is the unit sphere S^2 in E^3 given as the set of x such that $B(x, x) = 1$. Then if $f: S^2 \to \mathfrak{R}$ is the projection on some line—for example, the height or z-coordinate in an (x, y, z)-coordinate system on E^3—then f has two critical points, a maximum and a minimum. It can be shown that except for these two points all the other points of S^2 under the gradient flow go from the minimum to the maximum.

Riemannian manifolds. The construction of gradient in the previous section extends verbatim to any k-dimensional surface M in E^n (or submanifold of E^n).

Finally it should be remarked that all of this can be done abstractly on Riemannian manifolds by the same construction. A Riemannian manifold is a manifold M together with an inner product B_x on $T_x(M)$ for each x differentiable in x. This inner product defines an isomorphism $T_x(M) \to T_x{}^*(M)$ for each x, or putting this information for all x together, $T(M) \to T^*(M)$. Here $T^*(M)$ is the disjoint union of $T_x{}^*(M)$ over all x in M and is the cotangent bundle of M. This isomorphism converts 1-forms to vector fields. Thus, given any function $f: M \to \mathfrak{R}$, then $f \to Df \to$ grad f.

Other extensions. Finally, some extensions of gradients are discussed. In the example of a pure exchange economy, not one function on the state space is maximized but several. These are the utility functions of the m consumers. In this case the notion of gradient leads to versions of what is called Pareto Optima.

In a different direction, instead of a (positive definite) inner product, an indefinite symmetric form on Euclidean space can be considered. For example, if J is the bilinear form on $\mathfrak{R}^2$ defined by being equal to the difference of $x_1\mathfrak{L}(x_1)$ and $x_2\mathfrak{L}(x_2)$ (see 371), then J is symmetric and nondegenerate; and if a real differentiable P is given on E^2 a vector field grad $_JP$ is obtained replacing B by J in the definition. It turns out that the equations of electrical circuit theory can be expressed in this form. For example, if for P on $\mathfrak{R}^2$, the function is taken to be $-i_\lambda v_\gamma$ added to the integral from 0 to λ of $f(u)$ (see 372), then the equations of the circuit discussed earlier have the form of the derivatives of i_λ and v_γ being proportional to appropriate partial derivatives of P (see 373). These equations may be expressed in terms of grad $_JP(x)$ (see 374) in which $x = (i_\lambda, v_\gamma)$.

HAMILTONIAN MECHANICS

Hamiltonian mechanics as developed on manifolds has the virtue of including quite general mechanical systems and shows basic principles most simply. To define gradients, the inner product on Euclidean space E was of central importance. Hamiltonians are developed similarly with the bilinear form B, or inner product, replaced by a symplectic bilinear form Ω on E. $B(x, \mathfrak{L}[x])$ is a sum of terms $x_i\mathfrak{L}(x_i)$ (see 375) for Cartesian coordinates $(x_1, \cdots, x_n) = x$ on E. E is taken to have even dimension $2n$ and a bilinear form Ω on E is defined (see 376). Not only is it seen that Ω is bilinear, but it also is clearly antisymmetric, $\Omega(x, \mathfrak{L}[x]) = -\Omega(x, \mathfrak{L}[x])$. It can also be seen that Ω is nondegenerate or that the map $\phi_\Omega = E \to E^*$ is an isomorphism with $\phi_\Omega(x)(\mathfrak{L}[x]) = \Omega(x, \mathfrak{L}[x])$. In fact, any nondegenerate, antisymmetric bilinear form Ω can be used in the following, and such an Ω is called a symplectic structure on E. Ω can be used to define Hamiltonian dynamical systems.

Open sets of $\mathbf{E^{2n}}$. If U is an open set of E in which E has a symplectic structure (and so has even dimension) and $H: U \to \mathfrak{R}$ is any differentiable function, called a Hamiltonian in this context, then the derivative $DH: U \to E^*$ is a 1-form on U, and $X_H = \phi_\Omega^{-1} DH$ is a vector field (or ordinary differential equation) on U. Thus X_H is a map from U to E, and $\Omega(X_H(x), Y)$ equals $DH(x)(Y)$ for all x in U and Y in E. Note that X_H is defined as grad H except that B has been replaced by Ω.

In terms of the Cartesian coordinates on E in which $\Omega(x, \mathfrak{L}[x])$ has a specific form (see 377), X_H has for its first n coordinates $\partial H/\partial x_{n+i}$ and for its second n coordinates $-\partial H/\partial x_i$ (see 377). If $y_i = x_{n+i}$, $i = 1, \cdots, n$ the ordinary differential equations represented by X_H are the classical Hamilton's equations in which the derivative of x_i equals $\partial H/\partial y_i$ and the derivative of y_i equals $-\partial H/\partial x_i$ (see 378).

Particle in $\mathbf{E^3}$. A special case of the foregoing is the earlier mechanical system of a particle of mass m moving in 3-space under a conservative force field $F(x) = -\operatorname{grad} V(x)$ with $V: E^3 \to \mathfrak{R}$, (potential) function. Then on the state space $E^3 \times E^3$ coordinates are chosen (x_1, x_2, x_3) for the configuration and y_i is proportional to $\mathbf{v}_i$ (see 379) for the second factor (or momentum coordinates). If Ω is taken as above and $H: E^3 \times E^3 \to \mathfrak{R}$ defined in terms of a sum of the $y_i{}^2$ and $V(x)$ (see 380), then Hamilton's equations coincide with Newton's equations (see 381). The term (see 382) involving the $y_i{}^2$ is called the kinetic energy, and V is the potential energy of this mechanical system; the sum $K + V$ is the total energy or simply energy.

Manifolds. The concept of symplectic structure can be generalized easily to manifolds. If W is an even-dimensional manifold and x is a point of W, then a symplectic structure Ω at x is simply a nondegenerate, antisymmetric bilinear from Ω_x on the tangent space $T_x(W)$. A manifold W with such Ω_x given for each x, Ω_x varying differentially with x, is called a symplectic manifold. As before, for each x, $\phi_{\Omega x}: T_x(W) \to T_x{}^*(W)$ is an isomorphism from the tangent space to its dual, the cotangent space at x. Given a Hamiltonian $H: W \to \mathfrak{R}$, then for each x, the derivative $DH(x)$ is in $T_x{}^*(W)$, and the above isomorphism converts the 1-form DH to a vector field X_H (see 383). The vector field X_H is then the Hamiltonian

$$(377)\quad \begin{cases} \Omega(x, \mathfrak{L}[x]) = \sum_{i=1}^{n} x_i\, \mathfrak{L}(x_{i+n}) - \sum_{i=1}^{n} x_{n+i}\, \mathfrak{L}(x_i) \\ X_H = \left(\frac{\partial H}{\partial x_{n+1}}, \cdots, \frac{\partial H}{\partial x_{2n}}, -\frac{\partial H}{\partial x_1}, \cdots, -\frac{\partial H}{\partial x_n}\right) \end{cases}$$

$$(378)\quad \frac{dx_i}{dt} = \frac{\partial H}{\partial y_i}, \qquad \frac{dy_i}{dt} = -\frac{\partial H}{\partial x_i}, \qquad i = 1, \cdots, n$$

$$(379)\quad (y_1, y_2, y_3),\ y_i = m\mathbf{v}_i$$

$$(380)\quad H(x, y) = \frac{1}{2m}\sum y_i^2 + V(x)$$

$$(381)\quad \frac{dx_i}{dt} = \mathbf{v}_i, \qquad \frac{m d\mathbf{v}_i}{dt} = -\frac{\partial V}{\partial x_i}$$

$$(382)\quad \frac{1}{2m}\sum y_i^2, \qquad \frac{1}{2}m\sum \mathbf{v}_i^2$$

$$(383)\quad X_H(x) = \phi_{\Omega_x}^{-1}\, DH(x)$$

$$(384)\quad (DH(x))(X_H(x)) = \Omega_x(X_H(x), X_H(x))$$

vector field (or ordinary differential equation) corresponding to the Hamiltonian H. As a Riemannian manifold is required to define gradient vector fields, a symplectic manifold is required in order to define Hamiltonian vector fields.

Conservation laws. Various physical principles can be derived in this setting. For example, conservation of energy amounts to the function H being constant on trajectories of X_H. This is the same thing as the assertion that for each x, $DH(x)$ evaluated at $X_H(x)$ is zero. But another condition holds (see 384) from the definition of X_H, and the last is zero since Ω_x is antisymmetric.

Liouville's theorem

Liouville's theorem asserts that Hamiltonian systems preserve volume. In this context, that volume on the abstract manifold W is the n-fold wedge product of Ω with itself. Without pursuing the technical definition of wedge product, it can be said that if Ω is defined as above on E in coordinate, then the Liouville volume is just ordinary volume on E. It can be first checked that Ω itself is invariant under a Hamiltonian flow X_H, and then it naturally follows that the wedge product, the Liouville volume is preserved. This can be contrasted with the gradient case. For example, neighbourhoods of a local maximum of a gradient flow are shrunk into smaller subsets by the dynamics. Volume is strictly decreased.

Simple mechanical systems. A simple mechanical system (M, K, V) has a manifold M that is the configuration space of the system. $V: M \to \Re$ is a function, the potential energy, and kinetic energy K is considered to be given by some Riemannian metric on M. Thus for each x in M there is an inner product K_x on $T_x(M)$ and $K(v) = K_x(v, v)$ for v in $T_x(M)$. This can be compared with the earlier mechanical system of a particle constrained to move on the 2-sphere in E^3. In this case, K_x on $T_x(S^2)$ is the restriction of inner product on E^3 times half the mass. Together these sum to give the energy $E: T(M) \to R$ as a real function on the tangent bundle of M as defined by $E(v) = K(v) + V(x)$ for v in $T_x(M)$.

To put this into the symplectic perspective, it is convenient to consider the cotangent bundle $T^*(M)$. There is a canonically defined symplectic structure Ω defined on $T^*(M)$, which extends the earlier example of Ω on $E^3 \times E^3$. Furthermore, the energy E can be transferred to $T^*(M)$ by a map called the Legendre transformation $L: T(M) \to T^*(M)$. More precisely, if $v \in T_x(M)$, then $L(v) = \phi_{Kx}(v)$ in which ϕ_{Kx} is the usual linear map induced by a bilinear form. Then the Hamiltonian $H: T^*(M) \to \Re$ is defined by $H = E \circ L^{-1}$. The corresponding Hamiltonian vector field X_H on $T^*(M) = W$ gives the dynamics for this mechanical system. The inverse of the Legendre transformation takes trajectories of X_H back to solutions of the Euler-Lagrange equations on $T(M)$. (S.Sm.)

BIBLIOGRAPHY. A good elementary introduction to the theory of ordinary differential equations is WALTER LEIGHTON, *Ordinary Differential Equations* (1963). Methods of solution are listed together with the solutions of many particular equations in ERICH KAMKE, *Differentialgleichungen, Lösungsmethoden und Lösungen*, 6th ed. (1959). The best accounts of the classical theory are contained in EARL A. CODDINGTON and NORMAN LEVINSON, *Theory of Ordinary Differential Equations* (1955); PHILIP HARTMAN, *Ordinary Differential Equations* (1964); and EINAR HILLE, *Lectures on Ordinary Differential Equations* (1968). The qualitative theory is discussed in V.V. NEMYTSKII and V.V. STEPANOV, *Qualitative Theory of Differential Equations* (1960). The theory of differential equations in Banach spaces is treated in HENRI CARTAN, *Calcul différentiel* (1967; Eng. trans., *Differential Calculus*, 1971).

Elementary accounts of partial differential equations, especially of the equations of mathematical physics, and of techniques of solving them may be found in A.J.W. SOMMERFELD, *Partialle Differentialgleichungen der Physik* (1943; Eng. trans., *Partial Differential Equations in Physics*, 1949); and IAN N. SNEDDON, *Elements of Partial Differential Equations* (1957). The classical theory is treated fully in IVAN G. PETROVSKY, *Lectures on Partial Differential Equations* (1954; orig. pub. in Russian, 1950); BERNARD EPSTEIN, *Partial Differential Equations: An Introduction* (1962); PAUL R. GARABEDIAN, *Partial Differential Equations* (1964). Finite-difference methods of obtaining approximate solutions are contained in GEORGE E. FORSYTHE and WOLFGANG R. WASOW, *Finite-Difference Methods for Partial Differential Equations* (1960). DOROTHY BERNSTEIN, *Existence Theorems in Partial Differential Equations* (1950), gives a complete account of existence and uniqueness theorems. For an introduction to the abstract theory, see AVNER FRIEDMAN, *Partial Differential Equations* (1969); and ROBERT W. CARROLL, *Abstract Methods in Partial Differential Equations* (1969).

Many works on various aspects of analysis contain discussions of special functions. In addition to these, there are also many works devoted entirely to special functions. MILTON ABRAMOWITZ and IRENE A. STEGUN (eds.), *Handbook of Mathematical Functions* (1964), is a large handbook covering many special functions with numerical tables, statements of mathematical properties, bibliographies, and other information. A very extensive treatise is ARTHUR ERDELYI *et al.*, *Higher Transcendental Functions*, 3 vol. (1953–55). EDMUND T. WHITTAKER and GEORGE N. WATSON, *A Course of Modern Analysis*, 4th ed. (1928, reprinted 1958), contains a good treatment of special functions, as does IAN N. SNEDDON, *Special Functions of Mathematical Physics and Chemistry* (1956). For spherical harmonics, see THOMAS M. MACROBERT, *Spherical Harmonics*, 2nd rev. ed. (1948). There are also many works entirely devoted to one or another particular special function.

The reader interested in dynamical systems on manifolds can obtain some additional background on manifolds from SHLOMO STERNBERG, *Lectures on Differential Geometry* (1961); SERGE LANG, *Introduction to Differentiable Manifolds* (1962), a very basic book; and MICHAEL SPIVAK, *Calculus on Manifolds* (1965), an especially readable elementary work. For more information on ordinary differential equations, in addition to the works mentioned above, see WITOLD HUREWICZ, *Lectures on Ordinary Differential Equations* (1958), a highly recommended and less imposing book than most on this subject; and SOLOMON LEFSHETZ, *Differential Equations: Geometric Theory* (1957), which discusses structural stability and the van der Pol equation. Some works on dynamical systems are STEPHEN SMALE, "What Is Global Analysis," *Am. Math. Mon.*, 76:4–9 (1969); and "Differentiable Dynamical Systems," *Bull. Am. Math. Soc.*, 73:747–817 (1967)—both of these include further references. For the mathematics of electrical circuits, see CHARLES A. DESOER and ERNEST S. KUH, *Basic Circuit Theory* (1969). GERARD DEBREU, *Theory of Value* (1959), gives a good background in mathematical economics. An excellent standard text with a traditional view is HERBERT GOLDSTEIN, *Classical Mechanics* (1950); some modern approaches are in LYNN H. LOOMIS and SHLOMO STERNBERG, *Advanced Calculus* (1968), see especially ch. 13; RALPH ABRAHAM, *Foundations of Mechanics* (1967); V.I. ARNOLD and ANDRE AVEZ, *Problèmes ergodiques de la mécanique classique* (1967; Eng. trans., *Ergodic Problems of Classical Mechanics*, 1968); and STEPHEN SMALE, "Topology and Mechanics," 2 pt., *Inventiones Mathematicae*, 10:305–331 and 11:45–64 (1970).

(I.N.S./S.Sm./Ed.)

Digestion, Disorders of

Generally speaking, digestion is the process whereby ingested food is converted into material suitable for absorption into, and assimilation by, the body. The digestive system includes the alimentary canal—from mouth to anus—in which this and related processes occur, and the various glands associated with it. This article is concerned with disturbed functions of the digestive system, rather than with the diseases of that system of which such disturbed functions are a reflection. Because many symptoms of these disorders may occur in the absence of demonstrable disease of the gastrointestinal tract or of its accessory glands—the liver, gallbladder, and pancreas—emphasis will be placed upon these symptoms, their causes, their effects, and, in many instances, their treatment.

Symptoms related to the mouth, pharynx, and esophagus. Sensation in the mouth may be affected by alterations in taste or smell, or by odours, either stemming from the mouth itself or derived from exhalations from the nasopharynx, oropharynx, or the stomach. Many disorders of this type are brought about by the effects of various drugs upon the nerve endings in this highly sensitive area of the body; others represent actual disturbances of the nervous elements themselves, either as a result of primary nervous system diseases or from various psychiatric disorders. Since the central representation of smell, taste, and spatial recognition occurs in different portions of the brain, diseases of that organ, such as tumours, inflammations, or impairment of the blood vessels, may cause persistent distortion of these perceptions.

Careful neurological and dental examinations are necessary to distinguish between central and local causes, although it is probable that most of the minor difficulties of this type do reflect local hygienic conditions of teeth and gums or are derived from chronic inflammation in the nasal sinuses. Persons who habitually use tobacco or other local irritants in large amounts may find their taste discrimination faulty; normalcy returns, however, after discontinuance of their use.

Disorders related to chewing

Many problems related to the cutting and grinding of food in the mouth, and its subsequent reduction to a mass capable of being easily swallowed, are of course directly related to the presence or absence of teeth and to alignment of the upper and lower jaws. Insufficient chewing can make for difficult, and sometimes painful, swallowing and may reduce the efficiency of digestion. In certain neurological disorders or local disease processes in the throat there is perhaps more impairment of swallowing than of chewing, but the individual so afflicted may complain primarily of difficulty in moving the food within the mouth into an area that provides the best position for cutting, chewing, or grinding. Reduction in amount or quality of salivary juices, or poor mobility of the joint between the jawbones and the skull—the temporomandibular joint—may account for difficulty in chewing.

In general, individuals are able to identify fairly sharply whether dysphagia, or difficulty in swallowing, is high (in the back of the throat or in the neck) or low (farther down under the breastbone where the chest and abdomen meet). The distinction is of importance, both in the methods used to investigate the two areas and in the kinds of, and outlook for recovery from, the various causes. For most persons localization of a sensation of swallowed food "sticking" in one of these two areas is rather precise. The sensation that there is "a lump in the throat," however, which in no way interferes with swallowing, is called globus hystericus, an important psychoneurotic entity with a usually favourable prognosis. Various neurological disorders may impair the function of the pharynx and the upper constrictor muscles of the esophagus, although this is usually transient in the case of a stroke. Some chronic neurologic disturbances cause failures of synchronization of the striated, or voluntary, muscles of the pharynx with the smooth, or involuntary, muscle of the esophagus, and this can be a serious problem. Such conditions can now be identified without difficulty by using combinations of cineradiography, or serial X-rays, and techniques of recording pressure within the interior, or lumen, of the tube.

Esophageal spasm and achalasia

Low dysphagia, as mentioned previously, refers to the sensation that food, after leaving the pharynx in a normal manner, sticks at a level below the breastbone or sternum. In younger persons the most common cause for this is either a spastic contraction of the esophageal muscle, resulting in difficult forward or downward propulsion for the swallowed food, or a primary failure of the esophageal muscle to provide sufficient pressure to drive the food along. The first of these conditions, known as diffuse esophageal spasm, seems to occur rather commonly, often brought on by rapid eating, emotional disorders, or in response to the reflux of gastric juice back into the lower esophagus. The second, known as achalasia, is brought about when a degenerative process in the nerves to the esophageal muscle causes the entire esophagus to dilate. In the swallowing process, in either case, pain is generated by ineffective attempts to push food along into the stomach through a lower esophageal barrier.

In older persons the sensation of food "sticking" in this area is more often caused by a disease process, frequently a tumour, involving the wall of the esophagus and providing a mechanical rather than a functional obstacle to the passage of food. In all disturbances of this sort, careful study of the esophagus by direct observation through the esophagoscope, by cineradiological and pressure techniques, and by the study of cells removed from the area, is of great help in diagnosis and treatment.

Between the esophagus and the stomach there exists a mechanism that not only opens to allow food to pass through but also closes to serve as a barrier against the reflux of the stomach contents back into the lower esophagus. The pressure exerted upon the gastric contents by stomach contractions is frequently augmented by the action of the abdominal muscles, as during defecation or sneezing, bending over and lifting, or suddenly assuming the horizontal position. Pressure in the chest can be abruptly lowered during forced inspiration. Thus the pressure gradient across the esophagogastric junction may be very great; when it is, it favours the flow of stomach contents back into the esophagus. Acid gastric juice produces abnormal motor behaviour of the esophagus and, often, a burning sensation that seems to travel from the tip of the breastbone up under it toward the neck. This moving sensation is called heartburn and is known medically as gastroesophageal reflux.

Gastroesophageal reflux or heartburn

Although gastroesophageal reflux has been the subject of much research in recent years, all of its features are not yet completely understood, nor are details of the normal mechanisms that prevent it. Some of these mechanisms are: a normal pressure barrier exerted by the lower 2–3 centimetres of esophagus, an intra-abdominal segment of esophagus that is compressed by increased intra-abdominal pressures, an angle of entry (of the esophagus into the stomach) that permits the stomach mucous membrane to fold back over its opening as a kind of valve, and a normal gastric motor function that prevents the contents of the stomach from accumulating and distending its walls. When reflux does occur, and it is probable that nearly every normal person suffers from it occasionally, little harm is done if the esophagus contracts and pushes the refluxing juice back into the stomach quickly.

Peptic esophagitis

When reflux is frequent, however, and the esophagus cannot rid itself easily of the acid content—because of defects in its motor function or abnormalities such as herniation of the stomach through the diaphragm (hiatus hernia)—an abnormal state that may become chronic is produced. This abnormal state is called peptic esophagitis. The mucous membrane involved by peptic esophagitis is abnormally fragile and bleeds easily; the hiatus hernia (if present) frequently enlarges with the increasing age of the patient, and with obesity. The fat pad of the abdominal wall acts as a wedge elevating intragastric pressure when the patient strains or bends over. When medical treatment fails to control the factors making up this complex, surgery may be required.

Symptoms related to the stomach. The stomach movements are in a rolling and wringing pattern, beginning about one-third of the way down the length of the organ, and propel the mixture of food and juices toward its outlet, the pylorus. Any disorder that affects the power or coordination of the stomach muscles is capable of producing symptoms ranging from those that are mildly unpleasant to others that are incompatible with life. The unpleasant sensations called anorexia and nausea seem to be mediated through the central nervous system, with reflex input from nerve endings in the stomach and duodenum. Sometimes the entire duration of a nausea-vomiting episode is so short that it appears to be vomiting alone, obscuring the presence of nausea. This is characteristically noted in persons with primary diseases of the brain, especially those with tumours or meningitis in which the cerebrospinal fluid is under increased pressure. In many diseases vomiting may not be preceded by nausea at all, and in others there may be a long time lag between nausea and vomiting. Seasickness is the paradigm of this relationship.

The stomach muscles are innervated by branches of the vagus nerve, travelling all the way down the esophagus from their point of emergence in the brainstem. Cutting these nerves, as is often done in surgical treatment of peptic ulcer, may produce temporary or more prolonged changes in the ability of the stomach to empty itself. Many drugs, particularly the anticholinergic medications, are often used in the treatment of peptic ulcer. Anticholinergic drugs exert an action comparable to that produced by cutting the vagus; *i.e.*, they block the impulses from nerves to muscles in the stomach wall, with resultant slowed gastric emptying, sometimes to the point of complete inability to empty. Certain diseases, in particular diabetes mellitus, are associated with degeneration of the nervous fibres in the stomach, and consequently with impairment of the stomach's emptying powers.

These conditions, as well as those diseases that actually produce obstruction to the outflow tract of the stomach, may all result in gastric retention. Thus ulcers of the antrum, or distal portion of the stomach, characteristically obstruct gastric emptying, as do tumours infiltrating the walls and duodenal ulcers just outside the pylorus or in the pyloric channel. In such circumstances the stomach fills up with fluid of its own production, as well as with partially digested food, and vomiting often occurs. The sequel of persistent vomiting is dehydration, and death ultimately will ensue if the condition is not corrected. Persons who vomit frequently lose not only water but also such electrolytes as potassium, deficit of which markedly interferes with the ability of muscles to contract and of the kidneys and heart to function normally. The ingestion of soluble alkali, or "antacid," drugs in this situation produces profound disturbances in the acid-base defenses of the body.

Secretory disturbances

Secretory disturbances of the stomach by themselves rarely produce symptoms. Any human population may be classified into component groups by the relative ability of these components to secrete hydrochloric acid or pepsin. It cannot be stated categorically that those who produce the most acid have symptoms as a result, any more than the reverse can be implied. What can be stated, however, is that duodenal ulcer occurs more often in those who produce the higher amounts of acid, but such occurrence is not determined only by the quantity of acid produced. A substance called "intrinsic factor" is secreted by the gastric cells and is necessary for the absorption of vitamin B_{12} from food, but only a tiny amount of this factor is needed to ensure such absorption. Some immunologic block to its action is usually found in those persons suffering from pernicious, or B_{12} deficient, anemia. The stomach also secretes a hormone called gastrin, a powerful stimulus to the secretion of acid by the parietal cells of the gastric wall. The interaction of this hormone with histamine and acetylcholine, both found in the stomach wall, is a complex system for regulating the stopping, starting, and amounts of acid-peptic secretion.

"Indigestion"

The notion of "indigestion" and of its relationship to foods, habits, or the daily problems of living is highly imprecise in most circumstances. In children specific reactions to certain nutrients may occur, and it is possible to demonstrate that such reactions may be due to allergy. Almost every adult knows that certain foods or beverages will give him some kind of digestive tract disturbance, usually not disabling, but nevertheless unpleasant. These offending agents, furthermore, usually will give him more trouble when he is tense, frightened, tired, or otherwise ill, than when he is well and relaxed. On occasion the unpleasant reaction may be explosive and potentially fatal, as with giant hives, bleeding into the gut, or acute swelling of the pharynx or larynx. Unpleasant gastrointestinal symptoms are the most common adverse reactions to therapeutic agents and, when they occur, seriously limit the usefulness of the drugs. The mechanism by which such reactions are produced is poorly understood, and it may be more dependent on characteristics of the user than upon the drug itself. The interaction between the brain and the gastrointestinal tract has been well demonstrated, and many disturbances of gut function are closely dependent upon the "set" of the central nervous system. Such rare but profound psychiatric disorders as anorexia nervosa (psychotic loss of appetite) can often kill patients through starvation and incessant vomiting.

For many individuals the presence of gas provokes great distress. In the stomach there is always some air as a result of swallowing. The constant passage of air from the mouth into the esophagus and the stomach (and the reverse) is a form of nervous tic in some highly neurotic persons, but occasional belching is engaged in by everyone. The drinking of carbonated beverages influences the ease with which gas can be belched, but the total volume of gas in the stomach, like the total volume of acid gastric juice, is not correlated with symptoms. Some persons carry large gastric air bubbles without discomfort; others have difficulty even when only a little is present.

The small intestine. This longest portion of the gastrointestinal tract is primarily responsible for the absorption of foodstuffs. It is exquisitely designed to advance the chyme received from the stomach, admix it with digestive enzymes from the pancreas and the constituents of the bile, and then keep this mixture churning in small squirtlike movements until the necessary nutrients have been extracted from it.

Motor disturbances

When the coordination of the inner circular and outer longitudinal muscular layers of the intestinal wall is impaired, there is usually an accumulation of excess contents in the lumen, with consequent distention of the wall. This distention may cause pain, and it usually results in hyperactive contractions of the normal segment next to the disturbed loop. Such contractions may be strenuous enough to produce severe pain, which is characteristically crampy; *i.e.*, occurring in cycles with relatively normal pain-free intervals interspersed among the painful ones. When a large portion of the gut is irritated by an infection —as in viral gastroenteritis—such crampy pain may be generalized throughout the abdomen, but usually small bowel pain tends to be "located" by the brain as in the midline, either at the level of the navel, or slightly above or below it. The brain does not "see" sensation from the small bowel as either left or right until some infection or adhesion of the peritoneal membrane makes the external covering of the organ fix itself to the abdominal wall, from which the somatic nerves can determine its position.

The most serious problems in the area of small intestinal motor disturbances arise from intestinal obstruction, which, in turn, can result from an actual encroachment on the bowel by an adhesive band, or an internal block produced by a tumour or gallstone. Just as profound an obstruction results when a portion of the intestine undergoes partial necrosis, or death, from failure of its blood supply. This necrotic section cannot pass peristaltic activity and, for all practical purposes, is serving as an obstruction. The death of the tissue, furthermore, results in the escape of highly toxic fluids from the intestinal contents, through the wall, producing peritonitis. Surgical

correction of all these types of obstruction is usually necessary early in the course of the illness.

The speed with which material is passed along the small bowel is increased by many factors, including the swallowing of air, the nature of foods—which may be physically very cold or highly spiced—and most importantly by emotions. The upright position during physical activity, on the other hand, causes the small bowel to swing about on its mesenteric root, so that merely assuming the horizontal position will often markedly slow intestinal motor activity.

Absorptive disturbances

For proper absorption of nutrients, coordination of motor and digestive activities is necessary. Of the mixture of fats, proteins, carbohydrates, minerals, and vitamins that are ingested, it appears that fat, generally, is most subject to malabsorption when conditions become unfavourable. Measurements of fat absorption, consequently, have been used for some time as an index of general intestinal malabsorption. The solubilizing of fats is prerequisite to their absorption and requires certain optimum conditions for the action of bile salts, phospholipids, and the enzyme lipase from the pancreas. Normal motor function must be present so that these materials can be aggressively and continuously mixed. The cells lining the intestine must be normal, and contact between them and the mixture to be absorbed must be long enough to permit the lipids to leave the intestine and be carried into the cells. Pathways leading from these cells into the lymph channels and blood vessels of the body must be unobstructed. Under normal circumstances a gross excess of fat can be ingested without appearing in the stool, but if any of the aforementioned necessary conditions cannot be met, it will show up in the stool, either as visible fat or in a form detectable only on chemical analysis, depending upon the amount of fat ingested.

A dramatic example of the interrelation of motor and absorptive defects is shown by persons who are deficient in lactase, the enzyme that splits lactose (milk sugar) into glucose and galactose, its two component hexoses. Shortly after drinking milk, lactase-deficient persons usually have severe intestinal cramping, followed later by watery diarrhea. What has happened is that the lactose in the milk is not broken down and stays in the gut, drawing water to it osmotically. This increased bulk of fluid and sugar distends the intestine, which then contracts actively. The rapid contractions become crampy, and the material is driven along the intestine into the colon, which cannot absorb all the water rapidly enough. The resulting watery, unformed stools are frequently acidic.

Diagnosis of disturbances of small intestinal function is made primarily from the patient's history, physical examination, X-ray films of the abdomen, and by the study of the stools under controlled dietary conditions. Some test substances can be administered and their recovery measured in the stools and in the blood. Instruments are available that can be passed into the intestine to secure sections of intestinal mucous membrane for biopsy. Special methods of studying this material have been most helpful in furthering the knowledge of disorders of absorption. Motor aspects of the intestine can be studied rather grossly by measuring "transit times" after giving X-ray contrast material, or more precisely by the use of pressure monitoring through inlying tubes. Recently, a new method for motor and pressure monitoring has been developed, involving the use of "radio pills"—tiny, encapsulated transmitters that, after being swallowed, send the information in the form of electronic signals. The capsules are ultimately eliminated in the stool.

The large intestine. The colon begins in the right lower quadrant of the abdomen, rises along the right side, crosses the entire abdomen, and descends along the left flank and hip bone into the pelvis, where it is called first the sigmoid (S-shaped) colon and then the rectum, or straight colon. It is a large organ with an extensive surface area and relatively sluggish muscular activity. Its functions are to absorb water from the material ejected into it by the small intestine and to prepare this material for excretion as feces. It has a nervous supply that is complex but, unlike that of the small intestine, sufficiently extensive to allow the brain to perceive whether motor function disturbances are to the left or right, at midline, or in the pelvis. The brain becomes increasingly more precise in appreciating what is going on within the large intestine as the latter approaches its terminus at the anal skin. This is necessary if voluntary control over defecation (continence of feces) is to be learned. It may take small children a long time to acquire this degree of learned control, and disease can cause an adult to lose it in a short period of time.

The kinds of foods eaten and general habit patterns in toilet training may be responsible for rather wide variations in the frequency of stools passed, or "bowel movements." The frequency may range, for 95 percent of all healthy persons in the Western world, from three a day to three a week. Five percent of the population, on the other hand, have stool habits that may differ markedly from this, and they suffer no ill effects as a consequence. The average stool, of mixed Western diets, weighs under 150 grams and is about 70 percent water, although it appears solid. A slight increase in the water content may easily cause it to fragment into unformed fecal matter.

Diarrhea

The frequent passage of unformed watery stools is referred to as diarrhea. Water is normally absorbed from the colonic content, principally in the ascending, or right, colon. Thus, any inflammatory, neoplastic, or vascular disturbance of that part of the colon will usually decrease the firmness of the stool, increase its 24-hour total weight, and often produce blood or other evidences of inflammation, apparent when the feces are examined under the microscope. Some parasites, in particular *Entamoeba histolytica*, invade the right colon, and their motile or cyst forms may be identified in the stools. Surgical removal of the right colon for various disorders usually results in loose stools for months, or even years, until some form of adaptive compensation by the colon assists the water-reabsorbing process.

The colon does not show regularly rhythmic progressive peristalsis as does the small intestine, but rather exhibits a series of stationary segmental contractions (haustra) as its principal motor pattern. The purpose of this segmentation is to keep the contents from passing along too rapidly. Within the individual closed compartments thus produced, the material is churned to facilitate the absorption of water from it. As the contents are passed slowly over to the left side of the colon, the segmenting waves become less frequent, and the material is sent by coordinated waves into the pelvic, or sigmoid, colon where it is held up just above the rectum by a strong musculature assemblage. This is the area of the pelvic floor that represents a convergence point of abdominal wall muscles, the muscles of the pelvic floor itself, and thickened muscles of the colon wall. It is in this area that the final molding and ejection of the stool into the upper rectum occurs. The upper rectum is the site for the initiation of the defecation reflex; distending it with a balloon, for example, will call forth many of the reflexes noted to occur in the passage of a normal stool. The reflexes necessary for the completion of the urge to defecate are localized in the upper rectum and extend into the lower spinal cord. Diseases of the nerves, or of the muscles, involved will diminish or abolish the reflex, so that the patient will not be aware of the fullness of his rectum. The very last stronghold for control of the defecatory urge or of continence is the external anal sphincter, a muscular circle just under the anal skin, which is under central (brain) control. A momentary contraction of this ring can serve to inhibit the propulsive force by which the stool enters the rectum, so that it slips back up into the pelvic colon. Damage to this sphincter muscle by disease or surgery obviously reduces its competence.

Gas in the colon normally has a volume of about 100 millilitres. Some is from swallowed air, but some also arises from bacterial action on various substrates present in the colonic lumen. In the dietary intake of a lactase-deficient patient, the unhydrolyzed lactose will enter the colon, where the lactose equivalent of that in a glass of

milk will be capable of liberating, after bacterial fermentation, the equivalent of two cups (500 millilitres) of gas (hydrogen), about 15 percent of which diffuses back into the blood, and the rest passes as flatus. Although the small intestine normally digests and absorbs most short chain sugars, the longer complex sugars (*e.g.*, those found in beans and some other legumes) escape such action and in the colon are converted by bacteria into other products, with hydrogen gas given off in the process. Gas-fluid mixtures do not provide peristaltic pressure distribution as smoothly as do fluids alone. Gas is also subject to physical laws, and in the upright position it will diffuse to the most superior portions of the colon, at the right and left "corners." Here it can be compressed by contraction of adjacent segments, giving rise to pain that is localized either near the liver and gallbladder, or under the diaphragm and heart. Many a person has been thought to have diseases of these organs when the entire problem was caused by increased gas in the colon. Reduction of air intake by slowing down the rate of food ingestion, decreasing the intake of carbonated beverages and whipped desserts that contain air bubbles, and in avoiding certain gas-producing foods, such as most beans and nuts, usually helps to reduce flatulence.

The "irritable colon syndrome" is a term applied to periodic episodes of abdominal pain radiating from the colon, associated with the passage of small stools—sometimes loose or diarrheal but more often hard and pellet-like. The triggering mechanism for this disturbance appears to be an unrelaxed sigmoid colon. The pain can be very severe, and the disorder is primarily related to nervous tension, fatigue, or occasionally to a recent bacterial or viral infection of the intestine.

Constipation

Constipation, the undue delay in passage of feces, may result from failures at any point in the sequence of events described earlier. The most common forms of this disorder, which rarely occur in primitive civilizations, are the irritable colon syndrome, in which the rectum is usually free of feces, and various forms of sensory or motor failure in the colon-rectal segment. Some individuals' senses may be dulled by brain disease, metabolic failure, or by drugs—they simply do not appreciate the normal signals arriving in their brains to inform them that the lower colon is full of stool that needs expelling. Other persons, quite aware of the defecation urge, cannot mobilize the effective pressures needed to empty their colons, due to poor abdominal musculature or to a poor pelvic floor, sometimes the result of surgery or childbirth. Some persons are apathetic, eat irregularly and with no gusto, drink inadequate fluids, and do not observe regular times to defecate. Differentiation among these many causes can be achieved only by careful study, proper rearrangement of diet, habit patterns, emotional states, and by physical measures. Constant use of laxatives in the past has been a major cause of chronic constipation, but this practice is becoming less frequent.

Liver, gallbladder, and pancreas. These complex accessory glands provide essential digestive materials to the gut. Diseases of the liver characteristically produce nausea, and when bile flow is interfered with, the patient usually becomes jaundiced. In some forms of liver disease, itching and fever may be present in varying degrees, along with the jaundice. When bile flow is slowed or stopped under these conditions, the stools become lighter in colour—from light yellow to a silvery gray—and the urine becomes dark with bile. The gallbladder is a muscular hollow sac that normally contracts after a meal containing fat, so as to discharge concentrated bile into the upper intestine. If it is blocked by a gallstone the contractions become more intense and may be very painful. Such pains are called biliary colic. If the gallstone leaves the outlet of the gallbladder and moves into the common bile duct, it may obstruct this larger conduit from the liver and produce obstructive jaundice. Such an obstruction usually is associated with nausea and vomiting; behind the obstruction infection often occurs, accompanied by fever and chills.

The pancreas usually makes its disorders evident by producing pain, often in the back, or by acute inflammation associated with nausea, vomiting, and sometimes collapse. The digestive functions of the pancreas may be perturbed by pancreatic disease, and this then results in an inadequate delivery of pancreatic enzymes into the intestine. Such pancreatic insufficiency results in the production of large bulky stools and by marked weight loss due to the maldigestion. Digestive enzymes from the pancreas are now available and can be used in substitution therapy for pancreatic insufficiency.

BIBLIOGRAPHY. H.W. DAVENPORT, *Physiology of the Digestive Tract*, 3rd ed. (1971), is a lucid, well-illustrated, and critical survey of the most important functions and disorders of the gastrointestinal tract, exclusive of the metabolic functions of the liver. The most complete treatment of the functions of the alimentary tract, including the pancreas, gallbladder, and liver, is the massive five-volume *Handbook of Physiology*, sect. 6, *Alimentary Canal*, ed. by C.F. CODE *et al.* (1966–68).

(A.I.M.)

Digestion, Human

Digestion, in the human body, is the process of dissolving and chemically breaking down ingested food into simple compounds that are easily absorbed into the body. Digestion occurs chiefly through the actions of enzymes secreted into the alimentary canal and is aided by certain mechanical actions such as chewing and the motor action of the stomach, small intestine, and the large intestine. Thus, the ultimate aim of digestion is to reduce food to relatively simple chemical compounds that go into solution easily so that they can be absorbed by the mucous membrane that lines the alimentary canal.

The alimentary canal

The three primary activities of the alimentary canal of man are: secretion, motility, and absorption. These activities will be described, in the sections that follow, in some detail as they relate to the process of digestion. It is only after absorption that the various nutriments become available to the organism for assimilation into its own cells for such various metabolic activities as energy production, growth, repair, and conversion into other chemical compounds essential for life processes.

The alimentary canal begins at the mouth and ends at the anus. In between are the pharynx, esophagus, stomach, small intestine (comprised of the duodenum, jejunum, and ileum), large intestine (colon), and rectum (see DIGESTIVE SYSTEM, HUMAN). The contents of the alimentary canal are really external to the body, and it is only after transport across the epithelial lining of the alimentary canal that the chemical end products of digestion become available for use within the body.

In man, as in other animals, special enzymes are required for the digestion of the three principal food sources—carbohydrates, proteins, and fats. The digestion of each of these substances will be considered in the various parts of the alimentary canal of man. Minerals and vitamins, also essential for life, are absorbed directly in the form of soluble compounds.

DIGESTION IN THE MOUTH

Little digestion of food takes place in the mouth. Important mechanical and other events, however, usually do occur, and these prepare food for transport through the upper alimentary canal and thus aid the digestive processes in the stomach and small intestine.

Mastication

Mechanical activities. Mastication, or chewing, is the first mechanical process to which food is subjected in the mouth. Movements of the lower jaw in chewing are brought about by the muscles of mastication; these include the masseter, the temporal, the internal and external pterygoid, and the buccinator muscles. In chewing, the lower jaw or mandible may be moved in any one of several directions. A rotational movement of the jaw produces the grinding movements of the molar teeth. This kind of chewing is accomplished by the alternate contractions of the right and left pterygoid muscles.

The sensitivity of the peridental membrane, rather than the power of the muscles of mastication, limits the force of the bite. In man the force between upper and lower

molar teeth has been measured to be in excess of 250 pounds. The more sensitive teeth such as the incisors may exert a pressure of 30 to 50 pounds. The act of mastication may be voluntary, but for the most part it is a reflex act. Chewing movements can be elicited by electrical stimulation of appropriate areas of the cerebral cortex of the brain.

Mastication is not essential for adequate digestion and nutritional well-being. Chewing does aid digestion, however, by reducing food to small particles and mixing it with the saliva secreted by the salivary glands. The saliva lubricates and moistens dry food while the chewing distributes the saliva throughout the food mass. The lubrication of the food mass is primarily accomplished by the mucus, also secreted by the salivary glands.

Tongue movements

The movements of the tongue against the hard palate and the cheeks help to form a bolus of the food by moistening and mixing the material with saliva. After this intra-oral preparation, the bolus is manoeuvred to a position at the back of the mouth on the upper surface of the tongue. This has been termed the "preparatory position," as the bolus is always manoeuvred to a constant position on the dorsal surface of the tongue just before swallowing.

The first stage of deglutition, or swallowing, consists of passage of material through the oral cavity into the pharynx. This stage of swallowing is initiated voluntarily. The front portion of the tongue is then retracted and depressed, mastication ceases, respiration is inhibited reflexly, and the back portion of the tongue is elevated and retracted against the hard palate. This action, produced by the strong muscles of the tongue, forces the bolus into the pharynx. It has been demonstrated that during the first stage of swallowing negative pressure changes occur in the anterior part of the mouth, near the incisor teeth. In the posterior part positive pressure changes occur, brought about by the forceful contraction of the tongue against the hard and soft palates. Pressures as great as 100 centimetres of water have been recorded in the posterior part of the oral cavity during swallowing.

Composition of saliva

Chemical functions. The composition of saliva is variable, but principally it contains water, the same inorganic ions commonly found in the blood plasma, and a number of organic constituents. The variability of the composition of saliva is accounted for by the fact that the different salivary glands contribute different constituents. In addition, the concentration of many of the common inorganic constituents depends upon the rate of salivary flow.

The amount of saliva secreted by an individual in 24 hours usually amounts to 1,000 to 1,500 millilitres, or, roughly, a quart. This amounts, on the average, to about one millilitre per minute. Although saliva is slightly acid in reaction, the bicarbonates and phosphates contained within it serve as buffers and maintain the pH, or hydrogen ion concentration, of saliva relatively constant under ordinary conditions.

Bicarbonate concentration in saliva rises with increased salivary flow. The concentration of bicarbonate also is directly influenced by the partial pressure of the carbon dioxide in the arterial blood. Thus, with an increased arterial partial pressure of carbon dioxide, concentration of bicarbonate in saliva is increased. The chloride concentration of saliva also shows considerable variation. The total chloride and bicarbonate ratio, however, tends to remain rather constant. Chloride concentration of saliva from different glands may show considerable variation and, in general, there is a direct relationship between the chloride concentration and flow rate. At all flow rates, however, the chloride concentration is below that in blood plasma. At low flow rates the chloride concentration may be as little as five milliequivalents per litre, whereas at high flow rates it may rise to 70 milliequivalents per litre. The salivary glands are capable of secreting bromide and iodide in a manner similar to that of chloride. Sodium concentration in saliva is highly dependent upon flow rate. At low flow rate the concentration may be less than five milliequivalents per litre, whereas at high flow rates it may exceed 100 milliequivalents per litre. The potassium concentration in saliva is relatively high and exceeds that in the blood plasma. Ordinary mixed saliva usually contains from 8 to 20 milliequivalents per litre of potassium, which is 1½ to 4 times the concentration in plasma. The concentration of calcium in human saliva, like sodium, also increases with rate of flow. At high flow rates, a concentration of three to four milliequivalents of saliva per litre may be found.

Organic constituents of saliva consist of salivary proteins, free amino acids, lysozyme, specific blood group substances, and amylase. Glucose is normally absent from saliva even in individuals with diabetes. When saliva is subjected to electrophoretic analysis, the main protein band consists of salivary amylase. Amylase is the enzyme that digests starch, breaking it down into its sugar components, maltose and glucose. Digestion in the mouth consists primarily of the action of amylase on starch. There are no enzymes in saliva for the digestion of proteins and fats. The details of starch digestion by amylase, from both the salivary glands and the pancreas, will be described later.

Functions of saliva

The main functions of saliva are to initiate the digestion of starch and to provide a protective secretion in the mouth to keep the mucous membranes moist. Saliva has many other functions, however, including maintaining oral hygiene, which is accomplished by its solvent and cleansing action. The constant flow of saliva exerts a cleansing effect on the mouth and teeth and keeps the oral cavity and teeth comparatively free from food residues, sloughed epithelial cells, and foreign particles. By removing material that may serve as culture media, saliva inhibits the growth of bacteria. The lysozyme in saliva also serves a protective function, for it has the ability of lysing, or dissolving certain bacteria. Taste is mediated by chemical mechanisms, and in order to be tasted, substances must be in solution for the taste buds to be stimulated. Saliva provides the solvent for the solution of food materials.

The secretion of saliva also provides a mechanism whereby certain organic and inorganic substances can be excreted. A number of drugs and chemical substances are excreted in the saliva when introduced into the body. Examples of such substances include mercury, lead and potassium iodide, morphine, and certain antibiotics; *e.g.*, penicillin, streptomycin, chlortetracycline. Ethyl alcohol is excreted by the salivary glands, and the recommendation has been made that the alcohol content in saliva be used for medicolegal purposes.

Saliva is not essential to life. Its absence, however, results in a number of inconveniences, including dryness of the oral mucous membrane, poor oral hygiene because of bacterial overgrowth, and a greatly diminished sense of taste.

DIGESTION IN THE PHARYNX AND ESOPHAGUS

The primary digestive function of the pharynx (throat) and esophagus is to transport swallowed materials from the mouth to the stomach. The pharynx and esophagus have no significant secretory or absorptive functions. Transport of swallowed material through the pharynx constitutes the second stage of deglutition, and transport through the esophagus constitutes the third stage of deglutition.

The second and third stages of deglutition are entirely reflex actions and once initiated cannot voluntarily be interrupted. The initiation of the second and third stages of swallowing is brought about by contact of swallowed materials with the pharyngeal and peripharyngeal structures. The swallowing act, once initiated, is dominant over other functions occurring in this area. Thus, respiration and speech are interrupted by the second stage of swallowing. The complexity of the act and the high degree of integration necessary are indicated by the fact that the entire second stage of deglutition is completed in less than one second.

Pharyngeal transport

With the initiation of the swallowing act the bolus of food on the base of the tongue is rapidly propelled into the pharynx by the downward and backward movement of the tongue. At the same time the larynx moves upward and forward under the base of the tongue. Contraction of

the superior pharyngeal constricture muscles occurs, initiating a rapid pharyngeal peristaltic, or squeezing, contraction that moves down the pharynx propelling the bolus in front of it. The walls and structures of the lower pharynx are elevated to engulf the oncoming bolus. The epiglottis protecting the entrance of the larynx tends to divert the bolus to either side of the pharynx. The cricopharyngeus muscle, or upper esophageal sphincter, which has kept the esophagus closed until now, relaxes as the bolus approaches and allows it to enter the upper esophagus. The pharyngeal peristaltic contraction continues into the esophagus to become the primary esophageal peristaltic contraction. The swallowed bolus moves through the pharynx at a speed of 20 to 25 feet per second, and on entering the esophagus it often is projected deep into this organ.

Since the pharynx serves both the digestive tract and the respiratory tract, there are four possible outlets for swallowed material from the oral pharynx. These are: back into the mouth, up into the nasal pharynx, forward into the larynx, and downward into the esophagus. The swallowing reflex is so well coordinated that food normally takes only one of these possible paths, namely, that into the esophagus. Return into the mouth is prevented by the position of the tongue against the hard palate. Entrance into the nasal pharynx is prevented by elevation of the soft palate against the posterior pharyngeal wall. Entry into the larynx is prevented by this structure being drawn under the base of the tongue and also by the epiglottis diverting material away from the laryngeal opening.

Esophageal transport

The esophagus lies within the thoracic cavity and the pressure within its lumen is slightly below atmospheric pressure. At its upper and lower end the esophagus is sealed off by two sphincter muscles. These remain closed, opening only in response to the deglutition reflex.

Above, the cricopharyngeus muscle serves as the upper esophageal sphincter, separating the pharynx from the upper esophagus; below, the esophagus and stomach are separated by the lower esophageal sphincter.

Transport through the esophagus is accomplished by the primary esophageal peristaltic contractions. This is a continuation of the peristaltic contraction that originated in the pharynx. The primary esophageal peristaltic contraction is produced by a ring of advancing contracting muscles. Transport of material through the esophagus requires approximately ten seconds. This peristaltic wave serves to create a pressure gradient and sweep the bolus ahead of it. Pressures of 50 to 100 centimetres of water are created during the swallowing of liquids and solids. When the bolus arrives at the lower esophageal sphincter, which lies within the hiatus of the diaphragm, the sphincter relaxes and the swallowed bolus enters the stomach.

When a solid bolus is swallowed, it moves through the esophagus just in front of the advancing peristaltic contraction. If the bolus is too large or if the peristaltic contraction is too weak, the bolus may become arrested in the middle or lower esophagus. If this occurs, secondary peristaltic contractions originate just proximal to the bolus as a result of the local distension of the esophageal walls by the bolus, and propel it on into the stomach.

When a liquid bolus is swallowed, its transport through the esophagus depends somewhat on the position of the body and the effects of gravity. When swallowed in a horizontal or head-down position liquids are handled in the same manner as solids, with the liquid bolus moving immediately ahead of the advancing peristaltic contraction. (The high pressures and strong contractions of the esophageal peristaltic wave make it possible for animals with very long necks, as the giraffe, to transport liquid material through the esophagus to a height of many feet.) When the body is in the upright position liquids enter the esophagus and fall to the lower end and there await the arrival of the peristaltic contraction and the opening of the lower esophageal sphincter, a matter of some eight to ten seconds, before being emptied into the stomach.

DIGESTION IN THE STOMACH

The stomach serves as a temporary reservoir for ingested food and liquids. The size, shape, and capacity of the human stomach are extremely variable, not only from one individual to another, but also within the same person. The stomach can accommodate itself to the ingestion of small or very large meals. The stomach is capable of dilating to accommodate a meal or liquids whose volume is in excess of a litre without increasing intraluminal pressure within the stomach.

The purpose of the stomach in digestion is to mix the food with the gastric juices and partly to solubilize the ingested meal, preparing it for further digestion in the small bowel. Factors that regulate and control digestion in the stomach include its motor activity, the secretion of hydrochloric acid and enzymes, the rate of emptying the gastric contents into the small bowel, and the various types of chemical reactions taking place.

Gastric motility. Three types of motor activity of the human stomach have been observed and described. The first is a small contraction wave of the stomach wall that originates in the upper part of the stomach and slowly moves down over the organ toward the pylorus, the opening into the small intestine at the distal end of the stomach. This type of contraction produces a slight indentation of the stomach wall and it is thought to serve the purpose of mixing the gastric contents. The second type of motor activity is also a contracting wave but is peristaltic in nature. These contractions also originate in the upper part of the stomach and are slowly propagated over the organ to the pylorus. This type of gastric contraction produces a deep indentation in the wall of the stomach. When the peristaltic wave approaches the antrum, or the narrowing end, of the stomach, the indentation completely occludes the lumen, thus compartmentalizing it. The contracting wave then moves over the antrum, propelling the material ahead of it through the pylorus into the duodenum. This type of contraction serves as a pumping mechanism for emptying the contents of the gastric antrum. The third type of gastric motor activity that has been observed in man is best described as a tonus contraction. This type of motor activity decreases the size of the lumen of the stomach, as all parts of the gastric wall seem to contract simultaneously. It is this type of activity that accounts for the stomach's ability to accommodate itself to varying volumes of gastric content. The contraction of the gastric wall tends to be independent of the other two types of motor activity, particularly in different parts of the stomach. The mixing contractions and peristaltic contractions are normally superimposed upon gastric tonus.

The three types of stomach movements

Gastric motility can be recorded in man by having the subject swallow a small balloon connected to the end of a soft tube or catheter. The balloon is filled with water or air and the tube is connected externally to a recording device. With this method it is possible to determine the rate, duration, and pressures of mixing and peristaltic contractions. Gastric tonus is recorded by using balloons that detect volume changes. More recently investigators have been making these measurements by means of a "radio pill," a micro-miniaturized electronic transmitting device that is swallowed by the subject and transmits signals from the alimentary canal. These are received, decoded, and interpreted by the physician or technician. Both the mixing and peristaltic contractions of the human stomach occur at a constant rate of three per minute when recorded from the gastric antrum. This rate of three per minute is now recognized as the basic rhythm of gastric activity in man. Neither disease nor drugs alter this basic rhythm, although some drugs are capable of abolishing both types of contractions or stimulating the strength of contractions. The small mixing contractions of the stomach produce a pressure of approximately five centimetres of water. The peristaltic contractions may produce pressures as great as 50 to 60 centimetres of water.

Gastric secretion. The gastric mucosa that lines the stomach of man secretes approximately 1,200 to 1,500 millilitres (more than a quart) of gastric juice per day. This juice is highly acidic, because of its hydrochloric acid content, and is rich in enzymes. Gastric juice solubilizes food particles, initiates digestion, particularly of

protein, and converts the gastric contents to a semiliquid state called chyme, thus preparing it for introduction into the small intestine and further digestive processes.

Composition of gastric juice

Although man secretes gastric juice constantly, the composition of the juice is quite variable, depending upon the stimulus that evoked the gastric mucosa to secrete. It is useful to consider gastric juice as a mixture of secretions from the parietal cells, located in the gastric glands, and from the nonparietal structures in the gastric mucosa. The gastric juice is made up mainly of water, hydrochloric acid, and other inorganic constituents, including sodium, potassium, calcium, phosphate, bicarbonate, and sulfate. The organic constituents of gastric juice include mucus, the gastric enzymes, and the intrinsic factor of Castle. Each is described below.

Parietal-cell juice. Parietal-cell juice is essentially a solution of hydrochloric acid whose concentration is about 0.1 Normal (*i.e.*, 0.1 gram-molecular weight of hydrochloric acid per litre of solution). Hydrochloric acid is formed within the parietal cells, collected in small canaliculi within the cell, and emptied into the lumen of the gastric gland. The secretion is then carried to the surface of the mucosa and emptied out into the lumen of the stomach. Basal gastric secretion is measured in man, after an overnight fast, by placing a tube within the stomach to aspirate all secretions for an hour after awakening. A normal adult will usually secrete from two to four milliequivalents of hydrochloric acid per hour under these conditions. Patients with duodenal ulcer may secrete two to three times this amount.

Nonparietal-cell juice. In addition to the parietal cells, there are many other cells in the gastric mucosa that contribute to the composition of gastric juice. The zymogenic or chief cells of the gastric glands are the source of the enzyme pepsin. These cells contain zymogenic granules packed with the pro-enzyme pepsinogen. Pepsinogen is released into the lumen of the gastric gland and is transported to the lumen of the stomach. There it is converted into active pepsin and initiates the digestion of proteins. The gastric juice also contains small amounts of other enzymes. These include a gastric lipase, lysozyme, urease, and carbonic anhydrase. These latter enzymes are less important than pepsin in digestive function. Their exact cellular origin is unknown. Recent studies on human pepsin have demonstrated that at least four pepsin enzymes are contained in gastric juice. Pepsinogen is activated to pepsin by the presence of hydrochloric acid.

Gastric mucus. An important constituent of gastric juice is mucus. Mucus is secreted by the cells that line the neck of the gastric glands and by the surface epithelium cells. Two types of mucus have been recognized in gastric juice. One is called visible mucus, as it has a whitish colour and is thick, viscous, and even jelly-like. The other type of mucus is transparent or soluble mucus. This type of mucus forms a protective layer, approximately one millimetre thick, over the gastric mucosa. It is thought that it is this type of mucus that protects the gastric mucosa from the highly acid gastric juice that contains active pepsins and prevents the juice from digesting the stomach itself. Soluble mucus is alkaline in reaction and thus inactivates any pepsin molecules that diffuse into it.

Cause of pernicious anemia

Intrinsic factor of Castle. The gastric mucosa also secretes a mucoprotein known as the intrinsic factor of Castle, which combines with vitamin B_{12} in the diet. The binding of vitamin B_{12} by the intrinsic factor is necessary for the absorption of B_{12} from the small intestine. Vitamin B_{12} is essential for the normal maturation of red cells in the bone marrow (see VITAMIN; NUTRITION AND DIET, HUMAN). When the gastric mucosa is unable to secrete intrinsic factor in the gastric juice, there is no binding of B_{12}, and pernicious anemia results.

Phases of gastric secretion. The process of gastric secretion can be conveniently divided into phases, depending upon the primary mechanisms that evoke the gastric mucosa to secrete gastric juice. A brief description to each phase follows.

Cephalic phase. The cephalic phase of gastric secretion occurs in response to stimuli received by the senses; that is, taste, smell, sight, and sound. This phase of gastric secretion is entirely reflex in origin and is mediated by the vagus (10th cranial) nerve. Thus, at the sight or smell of an appetizing meal, the gastric mucosa is stimulated to secrete gastric juice in anticipation of the meal that will soon arrive in the stomach. Ivan Petrovich Pavlov, the famous Russian physiologist, originally demonstrated this mechanism of gastric secretion in dogs years ago. He prepared a small gastric pouch, with intact vagal innervation, that could be drained to the outside through a cannula for the collection of gastric juice. Each day when the dogs were fed, a bell would ring at the beginning of the meal. Gastric juice secreted by the small gastric pouch was collected for a period of time at the start of the meal. After a few days or weeks merely ringing the bell would cause the small gastric pouch to respond by pouring out copious amounts of gastric juice. When the vagus nerve is cut in experimental animals or man, the cephalic phase of gastric secretion is abolished.

Gastric juice secreted in response to vagal stimulation either directly, by electrical impulses, or by stimuli received through the special senses is rich in enzymes and highly acid.

Gastric phase. The gastric phase of gastric secretion begins when food enters the stomach. This can occur in a normal way after eating a meal or experimentally by introducing food into the stomach of an animal through a cannula connected to the outside. Pavlov observed that when food was placed through a gastric cannula into the stomach without the dog's knowledge, gastric secretion would begin after about 15 minutes. The gastric phase of gastric secretion is brought about by certain mechanical, chemical, and hormonal events. Distension of the stomach causes the mucosa to secrete. As the food becomes solubilized in gastric juice, the products of protein digestion elicit further gastric secretion. Distension of the gastric antrum by food, or the contact of the antral mucosa with the products of protein digestion, causes the antrum to release gastrin, a hormone that stimulates the gastric mucosa to secrete hydrochloric acid and pepsin. The gastric phase of gastric secretion lasts throughout the period of time that food is in contact with the gastric mucosa, and this particular phase is terminated when gastric emptying is complete.

Intestinal phase. When gastric chyme enters the small intestine it initiates the intestinal phase of gastric secretion. When food is placed experimentally directly into the small intestine by means of a tube, gastric secretion is elicited. The intestinal phase of gastric secretion may last for as long as 8 to 10 hours. Products of protein digestion are most potent in eliciting this phase of gastric secretion, but carbohydrates, fats, ethyl alcohol, and even water, when placed directly into the intestine, will elicit gastric secretion. The intestinal phase of gastric secretion, although lasting much longer than the cephalic or gastric phases, is the weakest of the three. Of the three phases of gastric secretion described above, it is the gastric phase during which the greatest amount of hydrochloric acid is secreted.

Interdigestive phase. Even after a prolonged fast the gastric mucosa continues to secrete small amounts of gastric juice containing hydrochloric acid and pepsin. This has been termed the interdigestive phase, or basal secretion, of the gastric mucosa.

Regulation of gastric secretion. As mentioned previously, gastric secretion in man is stimulated and regulated primarily by neural (vagus) and humoral mechanisms.

Neural and humoral stimulation

Stimulation of the gastric mucosa by way of the vagus nerves has already been mentioned; intact vagal innervation is essential for the cephalic phase of gastric secretion. Vagal innervation is also important in the gastric phase of gastric secretion because the vagus potentiates both the release of gastrin from the gastric antrum and also the response of the parietal cells to secrete hydrochloric acid.

The primary humoral mechanism regulating gastric secretion is the release of the gastric hormone gastrin from the gastric antrum. Small amounts of gastrin are apparently released continuously from the gastric antrum when the vagus is intact. As digestion proceeds in the stomach,

the products of protein digestion stimulate the antrum to release additional quantities of gastrin. Gastrin then stimulates parietal cells to secrete large amounts of hydrochloric acid. As the contents of the stomach become more and more acid, the acid chyme comes in contact with the antral mucosa. This then inhibits the release of gastrin from the antrum. Thus, there is a "feedback" mechanism related to the degree of acidity of the gastric contents. When the contents of the gastric antrum are less acid the mucosa releases gastrin. When the contents of the gastric antrum are more acid the release of gastrin is inhibited.

Another humoral substance that regulates gastric secretion is enterogastrone, a gastrointestinal hormone released from the duodenum in response to stimulation by various types of food coming in contact with the duodenal mucosa. Fat is the most potent stimulus for the release of the hormone. Enterogastrone inhibits both gastric secretion and gastric motility. Thus, as food enters the duodenum from the stomach, enterogastrone acts as a brake on both gastric emptying and gastric secretion. Two additional hormones, glucagon from the pancreas and secretin from the duodenal mucosa, exert some inhibitory effects on gastric secretion.

Gastric digestion. Digestion is not carried to completion in the stomach. Experimental evidence in man shows that approximately 30 to 50 percent of ingested starches and approximately 10 to 15 percent of ingested protein may be digested in the stomach. Little, if any, of ingested fat is digested there. Thus, the stomach is not primarily a digestive organ and is not essential to life. Its main function is to convert the ingested meal into chyme by mixing it with gastric juice in preparation for delivery into the small intestine, where most of the digestive processes occur.

Digestion of carbohydrates. When a meal has been thoroughly chewed and adequately mixed with saliva, the amylase in the saliva will continue its digestive action on starches in the stomach as long as the acidity of the gastric contents is not too high. Each digestive enzyme has an optimum hydrogen ion concentration, or pH, at which it can function most effectively. The optimum pH of amylase is approximately 6.8, slightly on the acid side of neutral (7.0). Thus, as the gastric contents are mixed with gastric juice and hydrochloric acid, the pH of gastric contents decreases (*i.e.*, becomes more acid) and the amylase activity diminishes accordingly.

Digestion of proteins. The proteolytic enzyme secreted in the gastric juice is pepsin. Pepsin has an optimum pH of approximately 2.0 and splits proteins mainly into the less complex proteose and peptone fragments of the protein molecules. Few amino acids or polypeptides are released from protein by the action of pepsin. Thus, protein digestion in the stomach is only partial—about 10 to 15 percent, as mentioned previously. Gastric secretion in response to a meal will have the highest concentration of pepsin within the first hour. As the volume of gastric juice increases, the concentration of pepsin falls. During the later stages of gastric digestion, when the volume of gastric juice diminishes, the concentration of pepsin increases.

Pepsin is not the only proteolytic enzyme in gastric juice. Others include cathepsin and gastricin. These proteolytic enzymes have an optimum pH of 4.0 and 3.0, respectively. Their activity is of relatively minor importance.

Digestion of fats. Little if any digestion of fats normally occurs in the human stomach. There is some evidence, however, for a gastric lipase, a tributerase with an optimum pH between 4.0 and 5.0 and inactive at pH below 2.5. The origin of this lipase is unknown and it acts only on short-chain fats; *i.e.*, those with less than five carbon atoms. Much of the fat in the human diet contains 18 carbons in the fatty acid chains; tributerase has little if any activity on these fats and is of negligible significance.

Gastric absorption. Although the stomach absorbs few products of digestion, it is capable of absorbing a great many other substances.

Foodstuffs. Components of the diet that can be absorbed through the gastric mucosa include glucose and other simple sugars, amino acids, and some fat-soluble substances.

Water. Water moves freely from the gastric contents across the gastric mucosa into the blood. The net absorption of water from the stomach, however, is small, because water moves just as easily from the blood across the gastric mucosa to the lumen of the stomach. To demonstrate absorption of water from the stomach, isotopic, or "heavy," water is introduced into the stomach by a tube, and the presence of the isotope in the blood is then determined. In man about 60 percent of heavy water placed in the stomach is absorbed into the blood in 30 minutes. In normal young adults the mean rate of water absorption from the stomach is about 2 percent of the gastric contents per hour.

Alcohol and drugs. A number of alcohols, including ethyl alcohol, are readily absorbed from the stomach. The membranes of the cells that line the stomach are composed of lipids or fat substances, thus items in the diet that are fat soluble penetrate the membranes of the cells lining the stomach. Since ethyl alcohol and a number of other drugs are somewhat soluble in fat, they are absorbed directly from the stomach. The pH of the gastric contents also determines whether or not some substances will be absorbed. At a very low pH, for example, aspirin is absorbed almost as rapidly as water from the stomach, but as the pH of the stomach rises it is absorbed much more slowly.

Gastric emptying. Gastric emptying time is determined by the rate of the antral pumping contractions (three per minute) and the volume of material emptied into the duodenum by each antral contraction. The stroke volume of each antral contraction varies, depending upon the nature of the contents being pumped. Usually, however, this amounts to volumes of three to five millilitres for each antral contraction.

Emptying time

An ordinary meal is usually completely emptied from the stomach in two to three hours. Both the physical and chemical composition of the meal influence the emptying rate. Fluids are emptied more rapidly than solids, carbohydrates more rapidly than proteins, and proteins more rapidly than fats. Liquid meals that are isotonic to blood, that is, having the same osmotic pressure as blood, are emptied more rapidly than hypotonic or hypertonic liquids. Different substances are emptied from the stomach at differing rates because of the various neural and humoral factors that regulate gastric emptying.

When food is received into the stomach from the esophagus, it tends to form concentric layers. The contents at the periphery of the ingested meal are in contact with the gastric mucosa where they are mixed with the gastric secretions and subjected to the contracting waves of the wall of the stomach. As the peristaltic waves pass over the intragastric contents they tend to push the liquefied peripheral portion of the gastric contents into the antrum. When the antrum becomes compartmentalized from the body of the stomach, its liquid contents are pumped into the duodenum. The process is repetitive and constant until the stomach is emptied.

DIGESTION IN THE SMALL INTESTINE

The small intestine is the principal digestive organ of the human alimentary canal. Three principal activities within the small intestine are especially adapted for its role in digestion: (1) its motor activity allows both mixing and transport of intraluminal contents; (2) secretions into the small intestine provide the necessary enzymes and other constituents essential for normal digestion; (3) it has highly selective absorptive capabilities. These absorptive functions of the small intestine are made possible by the specialized structures of the intestinal mucosa that line the small bowel.

Morphology of the small intestine

The morphology of the small intestine is arranged so that the inner mucosal surface, in contact with the intraluminal contents, greatly exceeds the external surface area. Some of the increased internal surface area is due to special folds in the intestinal mucosa known as plicae

circulares. This internal area in addition is greatly increased by the presence of intestinal villi—small finger-like structures about one millimetre in height and one-tenth of a millimetre in diameter, projecting from the surface. The villi are covered by tall columnar epithelial cells. The luminal surface of the epithelial cells have microvilli that exponentially increase the absorptive surface presented to the intraluminal contents.

Anatomically the small intestine is divided into three parts: the duodenum, the jejunum, and the ileum. The most proximal portion of the duodenum communicates with the stomach through the pyloric valve. The duodenum is approximately 9 to 11 inches long and receives bile, secreted by the liver, from the gallbladder and also the secretions from the pancreas. The duodenal mucosa itself is the origin of a number of gastrointestinal hormones that regulate not only gastric and pancreatic secretion but also emptying of the stomach and gallbladder. The electrical "pacemaker" that regulates small intestinal motility also is located in the duodenum.

The jejunum is the upper third of the remainder of the small intestine. Its mucosal villi are tall and well developed. By the time ingested food has passed through this segment of the small intestine, most of the digestive processes are completed, as is the absorption of the products of digestion.

The ileum is the remaining two-thirds of the small intestine, similar in structure and function to the jejunum but with less well-developed intestinal villi. If for any reason normal digestive and absorptive processes are not completed in the jejunum, they are carried to completion in the proximal ileum. The ileum also provides a special site for the absorption of vitamin B_{12} and bile salts.

Small intestinal motility. The primary purpose of the movements of the small intestine is to provide mixing and transport of intraluminal contents. Two types of motor activity have been recognized and designated as (1) segmenting contractions, and (2) peristaltic contractions. Characteristic of small intestinal motility is the inherent ability of the smooth muscle making up the wall of the intestine to contract spontaneously and rhythmically. This phenomenon is independent of any extrinsic nerve supply to the small intestine and creates pressure gradients from one adjacent segment of the intestine to another. The pressure gradients, in turn, are primarily responsible for transport within the small intestine.

Segmenting contractions

The predominant motor action of the small intestine is the segmenting contraction. A segmenting contraction may be defined as a localized circumferential contraction, principally of the circular muscle of the intestinal wall. The contraction involves only a short segment of the bowel wall, usually less than one to two centimetres, and constricts the lumen, tending to divide its contents. The contractions occur at a maximum rate of 11 per minute in the duodenum and eight per minute in the lower ileum. Thus, from the duodenum to the ileum there is a gradual decrease in the maximal rate of segmenting contractions, from above downward. This has been described as the "gradient" of small intestinal motility. Most often segmenting contractions occur in an irregular manner. At times, however, they may occur in an extremely regular or rhythmic pattern, and at a maximum rate for that particular site of the small intestine. This type of recurring segmenting contraction has been termed rhythmic segmentation.

Rhythmic segmentation may occur in a localized segment of small intestine with the segmenting contractions being stationary; or such action may occur in a progressive manner, with each subsequent segmenting contraction occurring slightly distal to the preceding one. The latter type of rhythmic segmentation is known as progressive segmentation.

The rate of segmenting contractions is controlled by a pacemaker located in the duodenum in the neighbourhood of the ampulla of Vater, a swelling or pouch in the duodenal wall (see below). This pacemaker sends electrical impulses down the small intestine at a rate of 11 cycles per minute in the duodenum, gradually decreasing to eight cycles per minute in the ileum. These electrical changes are propagated down the small intestine in the longitudinal muscle layer of the wall of the intestine. This cyclical slow-wave electrical activity is called basic electrical rhythm. Superimposed upon the slow-wave electrical activity may be fast, spikelike electrical changes. This type of electrical activity originates in the circular muscle layer of the intestinal wall and occurs when the circular layer contracts to form a segmenting contraction.

Peristaltic contractions

A peristaltic contraction may be defined as an advancing ring, or wave of contraction, passing along a segment of the small intestine. A peristaltic contraction normally occurs only over a short segment (a half-dozen or so centimetres) and moves at a rate of about one to two centimetres per minute. This type of small intestinal motor activity is primarily for transport of intraluminal contents from above downward, usually through one intestinal segment at a time.

When some inflammatory condition of the small bowel exists or when irritating substances are present in the intraluminal contents, a peristaltic contraction may travel over a considerable distance of the small intestine. This type of peristaltic contraction has been termed the peristaltic rush. Common infectious diarrhea is frequently associated with peristaltic rushes. Most cathartics produce their diarrheal effect by irritating the intestinal mucosa or increasing the contents, particularly with fluid. Either will induce a peristaltic rush.

Digestive secretions into the small intestine. The sources of digestive secretions into the small intestine are numerous. The gastric chyme that is emptied into the duodenum contains gastric secretions that continue their digestive processes while in the small intestine. One of the major sources of digestive secretion, however, is the pancreas, a large gland that produces both digestive enzymes and hormones (see DIGESTIVE SYSTEM, HUMAN), which empties its "juice" into the duodenum through both the major pancreatic duct at the ampulla of Vater and the accessory pancreatic duct a few centimetres proximal to the ampulla. Pancreatic juice contains enzymes that digest proteins, fats, and carbohydrates. The common bile duct delivers secretions of the liver, via the gall bladder, to the duodenum through the ampulla of Vater. Secretion from the mucosal surface of the small intestine is minimal and usually is no more than a few millilitres per hour for any one segment of intestine.

Secretions into the small intestine are under both neural (vagus) and hormonal control.

Secretions from the small bowel. Secretions from the small intestine are known as succus entericus. Except in the duodenum, the quantity of fluid secreted by the small intestine is never very great, even under conditions of stimulation.

Composition of fluids

The composition of the fluids secreted by the intestinal mucous membrane varies somewhat in the different parts of the intestine. In the duodenum, for example, where the mucous glands of Brunner are located, the secretion contains more mucus than is found elsewhere. In general the secretion of the small intestine is a thin, colourless, or slightly straw-coloured fluid, somewhat opalescent, and containing flecks of mucus. The intestinal juice consists of water, inorganic salts, and organic material. The inorganic salts are those commonly present in body fluids, with the bicarbonate concentration higher than it is in blood. The organic matter of the juice consists of mucus, enzymes, and cellular debris.

A great many enzymes have been reported as occurring in intestinal secretion. These include a pepsin-like protease (from the duodenum only), an amylase, a lipase, at least two peptidases, sucrase (invertase), maltase, lactase, enterokinase, alkaline phosphatase, nucleophosphatases and nucleocytases, and others (see also ENZYME).

Stimulation of the vagus nerve in experimental animals causes a moderate secretion of juice from the duodenum and smaller amounts from the rest of the small intestine after a latent period of one to one and one-half hours. Stimulation of sympathetic nerves supplying the small intestine causes no secretion, but cutting the nerves results in an increase in secretion.

The most effective stimulus for secretion of succus entericus is local mechanical or chemical stimulation of the intestinal mucous membrane. Such stimuli are always present in the digesting intestine in the form of chyme and the food particles that it contains.

A hormone called enterocrinin, extracted from intestinal mucosa, when injected intravenously or subcutaneously, causes a denervated jejunal loop to secrete intestinal juice.

Secretions from the pancreas. The volume of the pancreatic juice may vary from 500 to 1,000 millilitres per day. Pancreatic juice is colourless, odourless, of low viscosity, and alkaline due to the presence of sodium bicarbonate. The pH of pancreatic juice is usually 8.0 to 8.3. Pancreatic digestive secretions consist mostly of water, sodium bicarbonate, and organic constituents—mostly enzymes—and inorganic components, including small amounts of calcium, magnesium, zinc, potassium, chloride, phosphate, and sulfate.

The principal enzymes secreted by the pancreas include trypsin and chymotrypsin, which digest protein; lipase, which acts on fats; and amylase, which digests starch. In addition to trypsin and chymotrypsin, other proteolytic enzymes are also present. These include carboxypeptidase, ribonuclease, elastase, and a collaginase. In addition to pancreatic lipase, pancreatic juice also contains two phospholipases known as phospholipase A and B.

Regulation of pancreatic secretion

Regulation is provided by both neural and hormonal stimulation. Vagal stimulation causes the pancreas to secrete a juice rich in enzymes. The secretion of water and sodium bicarbonate is mainly under the control of the gastrointestinal hormone secretin, elaborated in the duodenal mucosa, released into the bloodstream, and carried to the pancreas where it stimulates the centroacinar cells to secrete water and sodium bicarbonate. The concentration of the sodium bicarbonate secreted by these cells is approximately 150–160 milliequivalents per litre. In its passage through the pancreatic ducts to the duodenum, some bicarbonate is exchanged for chloride. It is thought that the bicarbonate in pancreatic juice is dependent upon flow rate and that the greater the flow rate, the higher the concentration of bicarbonate.

Pancreatic enzymes are elaborated in the acinar (or exocrine) cells of the pancreatic glands and stored in the form of zymogen granules until released into the lumen of the acinus. The release of the zymogen granules is also under the control of both neural and humoral mechanisms. Cholinergic or vagal stimulation enhances the effect of the gastrointestinal hormone pancreozymin. This hormone, like secretin, is also elaborated within the duodenal mucosa, released into the bloodstream, and carried to its target organ, the acinar cells, and there stimulates them to release pancreatic enzymes.

Both secretin and pancreozymin have been isolated and purified. In man they can be used to test the exocrine function of the pancreas. Both are administered intravenously, and pancreatic juice is collected from the duodenum by a long intestinal tube. After stimulation of the pancreas with secretin, the volume of juice secreted increases greatly, and the concentration of bicarbonate goes up with increased flow rate. After an intravenous injection of pancreozymin, the secreted juice is rich in proteins, including the principal enzymes trypsin, lipase, and amylase.

Secretions from the liver. The secretions from the liver are contained in bile. Bile is formed by the hepatocytes of the liver. Initially it is collected within the biliary canaliculi located between adjacent liver cells. These small bile ducts join together to form larger ones, and ultimately the bile is collected into the two main hepatic ducts and then into the common bile duct. Under normal conditions, bile is stored in the gallbladder, where it becomes concentrated as the gallbladder mucosa absorbs water and some salts from the hepatic bile. The steady flow of bile from the liver serves as an avenue for the secretion of substances important in digestion and for the excretion of certain others.

Composition of bile

Liver bile consists of about 97.5 percent water and 2.5 percent solids. The main constituents of the solids that make up bile are bile salts. Other important constituents include certain inorganic salts, as are found in blood plasma, cholesterol, bile pigments, and small amounts of fatty acids, lecithin, and fats. The pH of the liver bile is slightly alkaline, being 8.0 to 8.6; gallbladder bile is usually neutral. Gallbladder bile contains a much higher concentration of organic solids than does liver bile, and the concentration of individual bile constituents may vary widely.

When gastric chyme is introduced into the duodenum it causes the duodenal mucosa to release a gastrointestinal hormone called cholecystokinin (now thought by some investigators to be the same as pancreozymin) that causes the gallbladder to contract. Thus, as gastric emptying occurs, the contraction of the gallbladder empties the concentrated bile into the small bowel to be mixed with the gastric chyme.

Bile salts

The bile salts are the most important constituents of bile as far as digestion is concerned. In man, four bile acids have been identified: cholic, chenodeoxycholic, deoxycholic, and lithocholic acids. The bile salts are composed of the corresponding bile acid combined with sodium. They are important to digestion because they activate pancreatic lipase, produce emulsification of fats, and promote the absorption of fats by combining with the products of fat digestion to form fat micelles, colloid-like particles. Although fat is absorbed primarily in the jejunum, the bile salts are not reabsorbed until they reach the lower ileum. Of all the bile salts delivered into the duodenum, approximately 90 percent are reabsorbed from the terminal ileum. There is thus a recirculation of bile salts back to the liver where they are subsequently resecreted into the duodenum. This is known as the enterohepatic circulation of bile salts.

Of the many products excreted by the liver in the bile, bilirubin is one of the most important. Bilirubin, a bile pigment, is an end product of hemoglobin breakdown and of porphyrin metabolism. In the liver it is conjugated with glucuronic acid and excreted in the conjugated form. In the small intestine, bilirubin is reduced to urobilogen and absorbed back into the blood, carried to the kidney, and excreted in the urine. Bacterial action in the small intestine and colon may unconjugate bilirubin, and the unconjugated bilirubin is absorbed into the blood and returned to the liver where it is re-excreted into the bile in the conjugated form. Thus, as in the case of bile salts, there is an enterohepatic circulation of bilirubin. Biliverdin, a green pigment, as the name implies, is also present in the bile but is of minor importance.

Small intestine chemical digestion. In the human diet the three principal components requiring digestion are carbohydrates, proteins, and fats. Most of the digestive processes that solubilize these substances and reduce them to relatively simple organic compounds occur in the proximal small intestine. The carbohydrates, fats, and proteins are substrates for the specific enzymes previously described.

Digestion of carbohydrates. Carbohydrates ordinarily make up about 50 to 60 percent of the total human diet. The amount eaten per day may range from 300 to 600 grams, representing 1,200 to 2,400 calories per day. Most of these carbohydrates are in the form of starches or polysaccharides—long-chain compounds containing many sugar molecules. The diet also contains some oligosaccharides, which are composed of two to ten simple sugars; disaccharides, compounds of two linked simple sugars, such as sucrose (common table sugar), maltose (malt sugar), and lactose (milk sugar); and monosaccharides or simple sugars such as glucose, or dextrose, and fructose.

Since carbohydrates are absorbed only in the form of monosaccharides or disaccharides, the polysaccharides and oligosaccharides must be reduced to these simpler components. This is accomplished by the enzyme amylase, which is secreted in the saliva by the parotid glands and in the pancreatic juice by the pancreas. The action of amylase on starch is to reduce it to maltose and glucose. On complete digestion of starch by amylase, starch is reduced to approximately 88 percent maltose and 12

Amylase action

percent glucose. Amylase also acts on other polysaccharides and oligosaccharides and reduces them to mono- and disaccharides.

Digestion of proteins. The amount of protein in the diet usually amounts to about 50 to 75 grams per day. Adults require 0.5 to 0.7 gram of protein per kilogram of body weight per day to remain in nitrogen balance. Growing children may require as much as four to five times this amount. The products of protein digestion are amino acids. Proteins are absorbed only as amino acids.

Pepsin, trypsin, chymotrypsin

In the gastric juice of man four pepsins have been identified. They initiate protein digestion in the stomach and reduce the ingested protein to simpler substances as proteoses and peptones, but only 10 to 15 percent of the ingested protein is reduced to amino acids by gastric digestion. In the small intestine, protein digestion is carried to completion by the action of pancreatic trypsin and chymotrypsin and by the proteases and peptidases that are released by the breakdown of epithelial cells peeled off the small bowel mucosa. Protein digestion is usually completed by the time the meal has passed through the first 100 centimetres of the proximal small intestine.

Digestion of fats. Little or no digestion of fats occurs in the stomach, and most ingested fats are delivered to the duodenum chemically unchanged. Ingested fat usually is from both animal and vegetable sources. The amount of fat in the diet varies with custom and the availability of fats. The average American or northern European diet contains from 50 to 100 grams of fat per day. In other parts of the world, particularly in colder climates, two to three times that amount may be consumed daily.

Because fats are insoluble in water, special digestive mechanisms are required to put them into solution and then to reduce them to simple absorbable units. Dietary fats consist mainly of triglycerides. A triglyceride is a fat molecule composed of a unit of glycerol linked to three units of fatty acids. The fatty acids in animal fat are mostly of the 16 or 18 carbon-atom-chain type. Pancreatic lipase hydrolyzes a triglyceride by splitting off the fatty acids that are connected to the ends of the glycerol molecule. This produces two molecules of fatty acids and a monoglyceride. Some triglycerides may be hydrolyzed completely into glycerol and fatty acids.

Bile salts secreted in the bile into the duodenum emulsify these products of fat digestion by combining with them to form fat micelles. Like carbohydrates and proteins, most fat is absorbed by the time it passes through the first 100 centimetres of the jejunum of the small intestine. A normal intestinal mucosa will absorb all dietary fat. If bile salts or pancreatic lipase are deficient, fat is not completely digested and will be excreted in the stool. Steatorrhea, an excess of fat in the stool, is a condition found in a number of diseases; *e.g.*, tropical sprue or adult celiac disease.

Small intestine absorption. Absorption of all food by the small intestine occurs principally in the upper part of that organ. The duodenum, although the shortest portion of the small intestine, plays an extremely important role in intestinal absorption. As noted previously, the duodenum is not only the receptor of the acid gastric chyme but is also the recipient of the pancreatic and liver secretions. It is in the duodenum that the intestinal contents are rendered isotonic with the blood plasma. The bicarbonate secreted by the pancreas neutralizes the acid secreted by the stomach. This brings the intestinal contents to the optimal pH, allowing the various enzymes to act on their substrates at peak efficiency. The duodenum is also the source of a number of important gastrointestinal hormones that regulate gastric emptying, gastric secretion, pancreatic secretion, and contraction of the gallbladder. These hormones, along with neural impulses from the autonomic nervous system, provide for autoregulatory mechanisms for normal digestive processes.

Absorption of electrolytes and water. Most salts and minerals, as well as water, are readily absorbed from all portions of the small intestine. Sodium, in the average daily diet to the extent of 12–15 grams, is absorbed by an active process, which means that an expenditure of energy is required for absorption. The necessary metabolic energy is provided by the epithelial cells of the small intestinal mucosa. Sodium is moved from the lumen of the bowel across the mucosa against both a concentration gradient, or increase, and an electrochemical gradient. The ion is absorbed more readily from the jejunum than from other parts of the small intestine.

Electrolytes

Potassium is absorbed at about 5 percent of the rate of sodium. It is thought that potassium moves across the intestinal mucosa in a passive manner or by diffusion as a consequence of water absorption. Chloride is readily absorbed in the small bowel and probably takes place as a consequence of sodium absorption.

Water absorption

Water appears to be dependent upon the absorption of electrolytes and occurs throughout the small intestine. The chief site of water absorption, however, is the jejunum. It has been shown by the use of isotopically labelled water that the upper small intestine absorbs approximately 95 percent of a 50-gram sample within ten minutes. Water, as mentioned previously, moves across the intestinal mucosa both ways. Thus, there is passage from the lumen of the intestine to the blood and from the blood to the lumen of the intestine. If the contents of the lumen are hypotonic, water moves more rapidly from the lumen to the blood. If the contents of the intestinal lumen are hypertonic, water moves more rapidly from the blood into the lumen. This two-way movement of water tends to maintain the intestinal contents in an isotonic state.

Absorption of carbohydrates. Carbohydrates are absorbed in the form of monosaccharides or disaccharides. Monosaccharides are absorbed more readily from the upper than from the lower small intestine. Glucose absorption is dependent upon the presence of sodium and is an active process; thus glucose can move across the intestinal mucosa against a concentration gradient. It is quite possible that there may be a specific pathway for the absorption of glucose and other monosaccharides that are actively transported across the mucosa, because certain structural requirements within the sugar molecule are necessary for the sugar to be accepted in the transport mechanism. Glucose and galactose have these structural requirements. Fructose, another simple sugar, is apparently absorbed by simple diffusion across the intestinal mucosa.

Disaccharides are absorbed subsequent to the action of specific enzymes, called disaccharidases, located within the microvilli of the epithelial cells. These enzymes split the disaccharides into simple sugars. Thus, sucrase hydrolyzes a molecule of sucrose into one molecule of glucose and one molecule of fructose. Maltase splits maltose into two molecules of glucose, and lactase converts lactose into glucose and galactose. The monosaccharides are then transported into the cells of the epithelium of the mucosa.

In certain individuals there is a deficiency or a congenital absence of one or more of the disaccharidases. The most common condition is that of a deficiency or absence of lactase. If these individuals drink milk, they cannot digest the milk sugar, lactose, and the nonabsorbable disaccharide remains in the intestine and frequently causes diarrhea. This is a common factor in so-called milk intolerance (see METABOLISM, DISEASES OF).

Absorption of protein. All digestible proteins are reduced by the proteolytic enzymes to amino acids and are absorbed as such. Amino acids may be classified into groups depending upon their optical rotatory characteristics (*i.e.*, whether they rotate polarized light to the left —levo; or to the right—dextro) and in terms of reactivity or pH. Amino acid absorption is an active process. Levorotatory amino acids are absorbed extremely rapidly—much more rapidly than are dextrorotatory amino acids —in fact, almost as quickly as they are released from protein or peptide. Neutral amino acids have certain structural requirements for active transport. If these specific structural arrangements are disturbed, active transport will not occur. Basic amino acids are transported at about 5 to 10 percent of the rate of neutral levorotatory amino acids.

Amino acids

Absorption of fats. The products of fat digestion are glycerol, free fatty acids, and monoglycerides. The glycerol is absorbed directly through the cell wall of the epithelial cells lining the mucosa. As explained previously, the free fatty acids and monoglycerides are combined with bile salts to form fat micelles. The fatty acids and monoglycerides are then absorbed from the micelles by the epithelial cells. Inside the cell the glycerol and fatty acids, as well as the monoglycerides, are reconstituted into new triglycerides or fat, the structure of which is characteristic of the species—in this case, of man. Molecules of triglycerides are then combined to form chylomicrons, which are microspheres of triglyceride coated with a layer of phospholipid. In this form they are transported by the lymph out of the intestinal mucosa and into the lymphatic system. Glycerol and free fatty acids that are not reconstituted into triglycerides within the mucosal cells are transported from the cells via the portal blood to and from the liver. The bile salts left behind in the upper small intestine move, along with other digestive residue, to the lower ileum, where they are absorbed and then recirculated through the liver.

Absorption of other substances. Calcium is absorbed from the small intestine by an active process. Apparently the amount of calcium absorbed from the intestine is controlled to meet the metabolic needs of calcium by the body. Parathormone, the hormone of the parathyroid glands, and vitamin D increase calcium absorption. The absorption of calcium appears to be a two-step process. First there is the uptake of calcium by the epithelial cells of the mucosa. The second step is the accumulation of calcium within the epithelial cells. Thus the calcium is not transported immediately across the mucosa into the blood. The entry of calcium into the cell is related, within certain limits, to the concentration of calcium within the lumen of the bowel. It appears that transport of calcium from the cells to the blood has a limiting rate, and if the calcium is not needed by the body it will be returned to the intestinal lumen when the epithelial cell is desquamated.

Absorption of iron

Iron in the amounts of 10 to 20 milligrams per day is contained in the average diet, of which only 0.5 to 1.0 milligram is absorbed. The absorption of iron is an active process and, like calcium, apparently occurs in a two-step fashion, first by entry into the cell and, second, storage in the cell with a slow release to the blood. The transfer of iron from the cells to the blood is regulated by the bodily need for iron. The total body store of iron is approximately four grams. Excessive amounts of iron stored in the epithelial cells of the mucosa are returned to the gut by desquamation of the epithelial cells. In the case of blood loss or during menstruation, the amount of iron absorbed a day may reach 1.0 to 1.5 milligrams.

Vitamin B_{12} is absorbed only in the terminal ileum. It first must be bound with the intrinsic factor of Castle, which, as noted above, is elaborated by the parietal cells of the gastric mucosa. The exact mechanism whereby vitamin B_{12} enters the intestinal mucosa is not known. The average diet contains adequate amounts of vitamin B_{12} and the normal requirement is only about one microgram per day.

DIGESTION IN THE LARGE INTESTINE

The large intestine or colon receives from the small intestine approximately one pint (500 millilitres) of intestinal chyme per day. This is ultimately reduced, in the colon, to approximately 150 millilitres, representing the volume of feces excreted each day. The primary function of the colon is to absorb water, sodium, and chloride from the digestive residue as it passes through.

Large intestinal motility. Twenty-four to 48 hours after oral administration of barium chloride, the opaque material commonly used in X-ray examination of the alimentary canal, the large intestine, or colon, will be filled with the substance and visible under fluoroscopy. A few minutes' observation shows no colonic motility. With longer periods of study using cineradiographic equipment, colonic segmenting contractions can be seen occurring at a rate of about two per minute in any one segment. The segmenting contractions are similar to those of the small intestine and produce sacculations, or recesses, called haustra. These contractions primarily serve the purpose of mixing the colon contents to aid in the absorption of water, sodium, and chloride.

Mass movement

Another movement of the large intestine transports the intraluminal contents. This type of motor activity is termed mass movement and usually occurs only two or three times per day; it is frequently associated with eating and thought to be initiated by what is called the gastrocolic reflex. When mass movements occur, materials in the cecum (ascending colon) may be transported to the transverse colon, or material in the transverse colon transported to the descending colon. When mass movements occur in the descending colon, the intestinal contents fill the rectum and initiate the defecation reflex. Since the primary function of the colon is an absorptive one in man, the types of motility it exhibits are particularly adapted to enhance this function.

Large intestinal secretion. The large intestinal mucosa does not have the villi observed in the mucosa of the small intestine but is more glandular, with many crypts. The surface of the mucosa is covered by tall columnar epithelial cells. The crypts are lined with mucous glands. The principal substances secreted by the colon are mucus, potassium, and bicarbonate. The mucus aids in lubricating the intestinal contents and facilitates their transport by the mass movement contractions.

Large intestinal chemical digestion. In a normal individual, digestion of carbohydrates, proteins, and fats is completed in the upper part of the small intestine. Thus, by the time that the intestinal contents have reached the ileocecal valve, all normal digestive and absorptive processes of foodstuffs have occurred. The intestinal chyme delivered to the colon is isotonic and consists mainly of nonabsorbable substances such as cellulose and other fibrous materials. Man has no enzymes that adequately digest cellulose. Cellulose and pectin taken in the diet may be broken down by colonic bacteria. The large intestine of man contains large numbers of bacteria as well as yeast and fungi. These micro-organisms, referred to as intestinal flora, serve the function of reducing undigested food and undigestible items in the diet to substances suitable for elimination in the feces. Only in abnormal conditions where there is a deficiency or lack of intestinal enzymes or bile, as occurs in some diseases of pancreas, liver, or small intestine, does a significant amount of undigested food reach the colon. When this does occur, the undigested material is broken down by the intestinal flora.

Large intestinal absorption. Absorption of water and storage of fecal material until elimination by defecation are the main functions of the colon. The net absorption of water by the colon represents about 350 millilitres per day. The capacity of the colon to absorb water, however, has been determined to be two to three litres per day, and in the process of absorbing water the contents of the large intestine remain isotonic.

In addition to water, significant amounts of salt and chloride are absorbed through the colonic mucosa. Isotonicity is maintained as the colon mucosa secretes potassium and bicarbonate while sodium and chloride are absorbed. Colonic absorption results in the daily formation of about 150 millilitres of fecal material. Fecal material is composed of about one-third solids and two-thirds water. Bacteria make up 10 to 20 percent of the total weight of solids excreted in the feces each day.

ELIMINATION

Sources of gas in the digestive tract

Normally the digestive tract of man contains from 100 to 200 millilitres of gas. Approximately one-third of this amount is found in the stomach and two-thirds in the large intestine. Not much gas is normally present in the small bowel. Gas in the digestive tract comes primarily from air swallowed with meals. The amount of air ingested with meals varies greatly but may be as great as 500 millilitres in some individuals. Air ingested with meals is usually eructated but may pass on down through the small intestine to the colon. Some of the gases in the

large intestine are the result of bacterial action on the intraluminal contents. The composition of gas in the stomach depends on when it was swallowed. Immediately after eating, the composition of stomach gas is very similar to that of ordinary air; that is, 20 percent oxygen and 80 percent nitrogen. After remaining in the stomach for a period of time, however, the composition approaches that of expired air from the lungs; that is, 15 to 16 percent oxygen, 5 to 9 percent carbon dioxide, and the remainder nitrogen. In the colon approximately half of the gas has its origin from swallowed air; the remainder is that produced by the colonic bacteria. The composition of colonic gas is made up mostly of nitrogen and carbon dioxide, with nitrogen representing 50 to 60 percent and carbon dioxide 30 to 40 percent, and the remaining 5 to 10 percent is composed of hydrogen sulfide, methane, and hydrogen. Colonic gas is normally expelled by the act of flatulation. Movement of gas in the digestive tract produces gurgling sounds known as borborygmus.

Defecation. The transformation of ileal contents to fecal material by the colon has been described previously. When mass movements of the colon force fecal material into the rectum, the defecation reflex is initiated. The rectum is normally empty, but when it is filled with gas, liquids, or solids to the extent that the intraluminal pressure is raised to 15 to 20 centimetres of water, the impulse to defecate occurs.

The defecation reflex

The act of defecation is, in the adult, preceded by a voluntary effort consisting of assuming an appropriate posture, voluntary relaxation of the external anal sphincter, and usually compression of the abdominal contents by means of straining efforts. These movements, in turn, probably give rise to stimuli that augment the visceral reflexes, although these originate primarily in the distended rectum. Centres that control defecation reflexes are found in the hypothalamus of the brain, in two regions of the spinal cord, and in the ganglionic plexus of the intestine. As the result of these reflexes, the internal anal sphincter relaxes. The mass contraction of the colon carries its contents into the pelvic colon, which in turn transfers them into the rectum, eventually to be evacuated by way of the anus. Thus, the entire distal colon from the splenic flexure to the anus may be emptied at one time. A prominent mechanical feature of the final act of defecation is contraction of the longitudinal muscles of the distal and pelvic colon. The shortening of the distal colon tends to elevate the pelvic colon and obliterate the angle that it normally makes with the rectum. The straightening and shortening of the passage facilitates evacuation.

The act of defecation is a reflex that is under some degree of voluntary control. The voluntary regulation consists of the ability to inhibit the reflex under normal circumstances and to initiate it voluntarily provided the necessary visceral stimulus (recent distension of the rectum) is still present. If the impulse to defecate is ignored, it will disappear but reoccur as the pressure increases. Eventually involuntary reflex action takes over and defecation is no longer under voluntary control.

Formation and composition of feces. As mentioned previously, by the time food material has progressed through the small intestine, practically all the digestible material has been absorbed and carried away by the blood or lymph. The residue consists of water, small amounts of undigested and unabsorbed food, the remains of desquamated mucosal cells, and the undigestible and unabsorbable fractions of the digestive secretion. In addition, there may be present various species of bacteria and the products of their metabolism. This material is emptied from the terminal ileum into the colon, where it is subjected to further alteration through absorption of water and the action of bacteria.

Food residues provide an excellent culture medium for bacteria, and the interior of the colon is a nearly ideal environment for their growth; in consequence they multiply enormously, so that from 10 to 20 percent of the solid weight of the feces may be made up of bacterial cells. The colonic contents are further modified by absorption of water and inorganic salts, so that of the 75 to 180 grams of feces daily excreted only about 60 to 70 percent is normally water; this proportion varies a great deal, however, according to the consistency of the feces.

Feces normally contain from 5 to 25 percent of fatty material, including such substances as neutral fats, free fatty acids, soaps, and sterols. Fats are present in appreciable amounts even on a fat-free diet, proving that some fat is excreted through the intestine. About one-third of the fecal lipid consists of sterols, chiefly cholesterol. The daily output of feces contains from 0.5 to 1.5 grams of nitrogen as nitrogen compounds. About half the nitrogen represents nitrogenous constituents of the fecal bacteria; the rest represents unabsorbed intestinal secretions and digestive fluids, mucus, desquamated mucosal epithelial cells, and a small amount of food residue; some digestive enzymes also may be present. Usually absorption and digestion of protein are so nearly complete that only small amounts of food protein nitrogen escape into the feces; likewise, only small quantities of digestive enzymes are normally present, since they are usually destroyed by other enzymes or by the action of bacteria. Digestive enzymes may be increased in the stool in diarrhea and may then cause severe skin irritations.

Since the pH of the feces is usually between 7.0 and 7.5, the inorganic constituents are mainly substances that are poorly soluble in alkaline pH ranges; these are chiefly calcium phosphate and oxalate and iron phosphate. Small amounts of magnesium, potassium, and sodium salts are also present in the stool. Since cellulose and lignin are not digested, they may be found in the feces whenever they are present in the diet, particularly in fibrous foods. The total volume of the feces may be considerably influenced by the amount of cellulose or lignin ingested.

The dark brown colour of the normal stool is due chiefly to stercobilin and urobilin, reduction products of the action of bacteria on the red bile pigment, bilirubin. The main cause of odour is the presence of skatole and indole, produced by bacterial action, and, to a lesser degree, the presence of hydrogen sulfide and methyl mercaptan. The odour is more pronounced on a high protein diet because of the formation of larger quantities of these odoriferous substances.

Even when no food is taken, about seven to eight grams of feces are excreted daily. This consists of desquamated mucosal cells, digestive secretions, bacteria, and bacterial residue. For the fate of ingested foods after digestion and absorption, see METABOLISM.

BIBLIOGRAPHY. Additional information on this subject may be found in the following texts: B.P. BABKIN, *Secretory Mechanism of the Digestive Glands*, 2nd ed. rev. (1950); A.S.V. BURGEN and N.G. EMMELIN, *Physiology of the Salivary Glands* (1961); C.F. CODE (sect. ed.), *Alimentary Canal*, sect. 6, vol. 1–5, *Handbook of Physiology* (1966–68); H.W. DAVENPORT, *Physiology of the Digestive Tract*, 3rd ed. (1971); N.C. HIGHTOWER, JR., "Section on Digestion," in *The Physiological Basis of Medical Practice*, 8th ed., pp. 1061–1260 (1966); B.A. HOUSSAY *et al.*, *Human Physiology*, 2nd ed. (1955); T.H. WILSON, *Intestinal Absorption* (1962); and A.V. WOLF, *Thirst* (1958).

(N.C.H.)

Digestion and Digestive Systems

In order to sustain themselves all organisms must obtain nutrients from the surrounding environment. Some nutrients serve as raw materials for the synthesis of cellular material; others (*e.g.*, many vitamins) act as regulators of chemical reactions in living cells; and still others, upon oxidation in living cells, yield energy. Not all nutrients, however, are in a form suitable for immediate use by an organism; some must undergo physical and chemical changes before they can serve as energy or cell substance. This article deals with digestion and digestive systems generally, referring to specific organisms to clarify the account; it emphasizes the vertebrate condition in the latter half. For aspects strictly concerning human digestion see DIGESTIVE SYSTEM, HUMAN; DIGESTIVE SYSTEM DISEASES.

GENERAL FEATURES

Distinctions and definitions. Through the act of eating, or ingestion, nutrients are taken from the environ-

ment. Many nutrient molecules are so large and complex that they must be split into smaller molecules before they can be used by the organism. This process of breaking down food into molecular particles of usable size and content is called digestion. Unusable components are expelled from the organism by a process called egestion. Some plants, many micro-organisms, and all animals perform these three functions—ingestion, digestion, and egestion (often grouped under the term alimentation)—but, as expected, the details differ considerably from group to group.

Autotrophy and heterotrophy

The problems associated with nutrient intake and processing differ greatly depending on whether the organism is autotrophic or heterotrophic. Autotrophic organisms are those that can manufacture the large energy-rich organic compounds necessary for life from simple inorganic raw materials; consequently, they require only simple nutrients from the environment. By contrast, heterotrophic organisms cannot manufacture complex organic compounds from simple inorganic ones, and so they must obtain preformed organic molecules directly from the environment.

Green plants constitute by far the majority of the Earth's autotrophic organisms. During the process of photosynthesis, they use light energy to synthesize organic materials from carbon dioxide and water. Both compounds can be absorbed easily across the membranes of cells—in a typical land plant, carbon dioxide is absorbed from the air by leaf cells, and water is absorbed from the soil by root cells—and used directly in photosynthesis; *i.e.*, neither of them requires digestion. The only other nutrients needed by most green plants are minerals such as nitrogen, phosphorus, and potassium, which also can be absorbed directly and require no digestion. There are, however, a few green plants (*e.g.*, sundew, Venus'-flytrap, pitcher plant) that supplement their inorganic diet with organic compounds (particularly protein) obtained by trapping and digesting insects and other small animals.

Heterotrophy characterizes all animals, most microorganisms, and plants and plantlike organisms (*e.g.*, fungi) that lack the pigment chlorophyll, which is necessary for photosynthesis. These organisms must ingest organic nutrients—carbohydrates, proteins, and lipids (fats)—and, by digestion, rearrange them into a form suitable for their own particular needs.

Ingestion. As already explained, the nutrients procured by most green plants are small inorganic molecules that can move with relative ease across cell membranes. Heterotrophic organisms such as bacteria and fungi, which require organic nutrients yet lack adaptations for ingesting bulk food, also rely on direct absorption of small nutrient molecules across cell membranes; molecules of carbohydrates, proteins, or lipids, however, are too large and complex to move easily across cell membranes. Bacteria and fungi circumvent this problem by secreting digestive enzymes onto the food material; these enzymes catalyze the splitting of the large molecules into smaller units that are then absorbed into the cells. In other words, the bacteria and fungi perform extracellular digestion—digestion outside cells—before ingesting the food. This method of feeding is often referred to as osmotrophic nutrition.

Ingestion by simple cells

Like bacteria, protozoans are unicellular organisms, but their method of feeding is quite different. They ingest relatively large particles of food and carry out intracellular digestion—digestion inside cells—a method of feeding called phagotrophic nutrition. To a lesser degree many protozoans also are osmotrophic. Some organisms (*e.g.*, *Amoeba*) put out pseudopodia ("false feet"), which flow around the food particle until it is completely enclosed in a membrane-bounded chamber called a food vacuole; this process (Figure 1A) is called phagocytosis. Other protozoans (*e.g.*, *Paramecium*) pinch off food vacuoles from the end of a prominent oral groove into which food particles are drawn by the beating of numerous small, hairlike projections (cilia). In still other cases of phagotrophic nutrition, tiny particles of food adhere to the membranous surface of the cell, which then folds

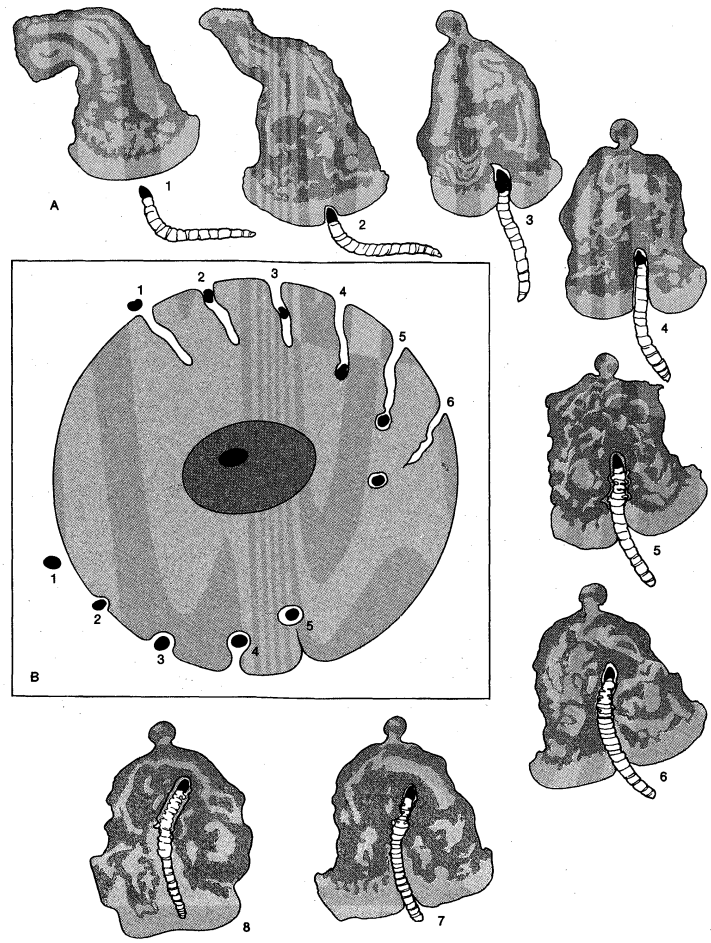


Figure 1. *The ingestion of food by cells.*
(A) Phagocytosis, or the engulfment of a food particle; (B) two forms of pinocytosis, or the pinching off of food vacuoles.
Adapted from (A) James G. Hirsch in A.E. Nourse, *The Body;* (B) From W.T. Keeton, *Biological Science,* © 1967 by W.W. Norton & Company, Inc.

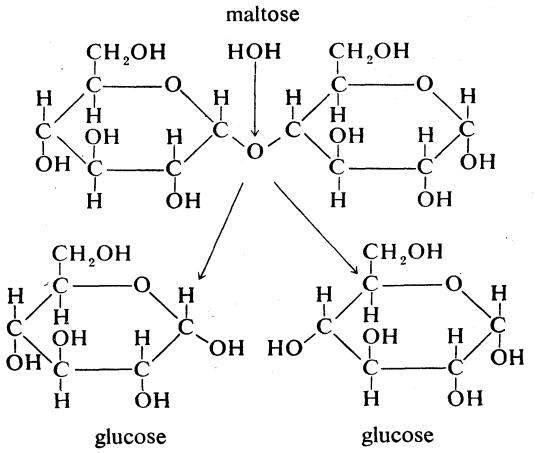
inward and is pinched off as a vacuole; this process (Figure 1B) is called pinocytosis. The food particles contained in vacuoles formed through phagocytosis or pinocytosis have not entered the cell in the fullest sense until they are digested into molecules able to cross the membrane of the vacuole and become incorporated into the cellular substance.

Most multicellular animals possess some sort of digestive cavity—a chamber opening to the exterior via a mouth—in which digestion takes place. Large particles of food are broken down to units of more manageable size within the cavity before being taken into cells and reassembled (or assimilated) as cellular substance.

Digestion. The enzymatic splitting of large and complex molecules into smaller ones is effective only if the enzyme molecules come into direct contact with the molecules of the material they are to digest. In animals that ingest very large pieces of food, only the molecules at the surface are exposed to the digestive enzymes. Digestion can proceed more efficiently, therefore, if the bulk food is first mechanically broken down, exposing more molecules for digestion. Among the variety of devices that have evolved to perform such mechanical processing of food are the chewing teeth of mammals and the muscular gizzards of birds.

The chemical basis of digestion

The chemical reactions involved in digestion can be clarified by an account of the digestion of maltose sugar. Maltose is, technically, a double sugar, since it is composed of two molecules of the simple sugar glucose bonded together. The digestive enzyme maltase catalyzes a reaction in which a molecule of water is inserted at the point at which the two glucose units are linked, thereby disconnecting them, as illustrated in the following structural diagrams.



In chemical terms, the maltose has been hydrolyzed. All digestive enzymes act in a similar way and thus are hydrolyzing enzymes.

Many other nutrient molecules are much more complex, being polymers, or long chains of simple component units. Starch, for example, is a carbohydrate, like maltose, but its molecules are composed of thousands of glucose units bonded together. Even so, the digestion of starch is essentially the same as the digestion of maltose: each linkage between adjacent glucose units is hydrolyzed, with the result that the starch molecule is split into thousands of glucose molecules. Protein molecules also are polymers, but their constituent units are amino acids instead of simple sugars. Proteolytic (*i.e.*, protein-digesting) enzymes split the protein chains by hydrolyzing the bonds between adjacent amino acids. Because as many as 20 different kinds of amino acids may act as building blocks for proteins, the complete digestion of a protein into its amino acids requires the concerted action of several different proteolytic enzymes, each capable of hydrolyzing the bonds between particular pairs of amino acids. Fat molecules, too, are composed of smaller building-block units (the alcohol glycerol plus three fatty acid groups); they are hydrolyzed by the enzyme lipase.

Various other classes of compounds are digested by hydrolytic enzymes specific for them. Not all of these occur in every organism; for example, few animals possess cellulase (cellulose-digesting enzyme), despite the fact that cellulose constitutes a high percentage of the total bulk of the food ingested by plant-eating animals (herbivores). Some herbivores, nonetheless, benefit from the cellulose in their diet because their digestive tracts contain micro-organisms (known as symbionts) capable of digesting cellulose; the herbivores absorb some of the products of their symbionts' digestive activity.

So far, emphasis has been placed on the role of digestion in converting large complex molecules into smaller simpler ones that can move across membranes, thus permitting absorption of food into cells. The same processes occur when substances must be moved from cell to cell within a multicellular organism. Thus, green plants, which do not have to digest incoming nutrients, digest stored material, such as starch, before it can be transported from storage organs (tubers, bulbs, corms) to points of utilization, such as growing buds.

The elimination of indigestible matter

Egestion. Animals that ingest bulk food unavoidably take in some matter that they are incapable of using; for example, since man lacks cellulase, the cellulose he ingests in vegetables and fruits is indigestible. It cannot be absorbed from the digestive tract and must be expelled from the body.

In the case of unicellular organisms that form food vacuoles, the vacuoles eventually fuse with the cell membrane and then rupture, releasing indigestible wastes to the outside (Figure 2). Multicellular animals periodically release such waste from their digestive tracts either by regurgitating it through the mouth or by eliminating it as feces through the anus.

INVERTEBRATE DIGESTIVE SYSTEMS

Vacuolar systems. Unicellular organisms that ingest food particles via vacuoles rely on intracellular digestion to prepare the nutrients for use. The enzymes that catalyze this digestion, being very potent chemicals capable of breaking down the cell substance itself, are held until needed in special packets, or vesicles, called lysosomes; the membrane of a lysosome is both impermeable to the enzymes and capable of resisting their hydrolytic action. Soon after a food vacuole is formed, a lysosome fuses with it (Figure 2). Food material and digestive enzymes are mixed in the resulting composite vesicle, which is sometimes called a digestive vacuole. This vacuole moves in an orderly fashion through the cell, during which passage the products of digestion are absorbed, leaving the indigestible material, which is eventually expelled.

Vacuolar digestion is not restricted to unicellular organisms, however. Many multicellular invertebrates partly digest their food extracellularly before phagocytizing the remainder, which is then digested intracellularly by the same process described in the preceding paragraph.

Channel-network system. The sponges, among the simplest multicellular organisms, have what amounts to diversionary water channels that serve to bring water and food to their component cells. The channels are lined with special cells bearing whiplike structures called flagella that create water currents. A steady flow of water inward through smaller secondary channels and then out the main, or excurrent canal, carries with it bits of food. The lining cells capture the food particles and enclose them in food vacuoles, wherein the matter is digested as in protozoans, by intracellular means.

Saccular systems. With the evolution of multicellular organisms came a corresponding evolution of cellular specialization, resulting in a division of labour among cells; in this way, certain cells became specialized to perform the function of digestion for the entire organism. Cnidarians, especially hydra, provide a comparatively simple example. These radially symmetrical animals have a saclike body composed of two principal layers of cells (Figure 3B). The cells of the outer layer function as a

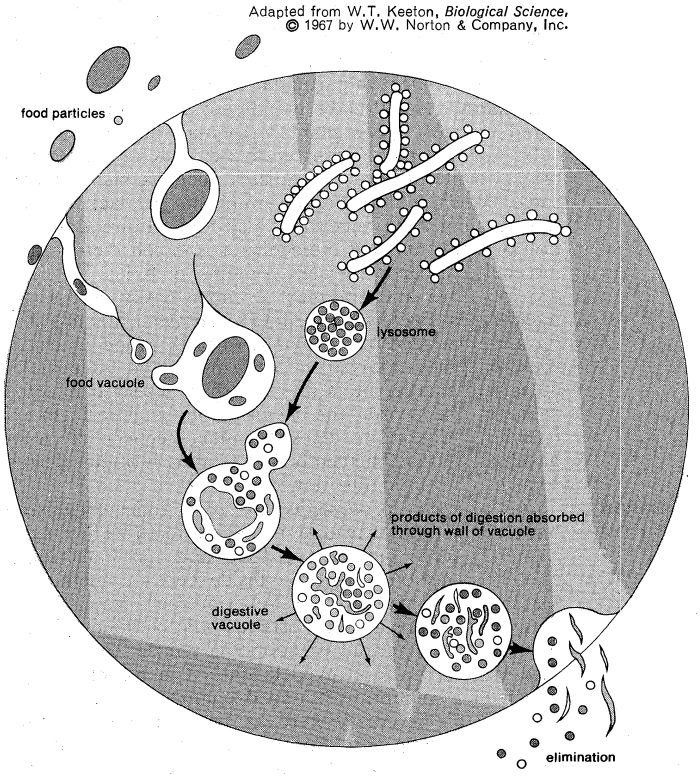


Figure 2: The role of lysosomes in intracellular digestion. Digestion takes place when a food vacuole and a lysosome unite, forming a digestive vacuole. The products of digestion are absorbed across the vacuolar membrane, and the indigestible wastes are ultimately expelled to the outside.

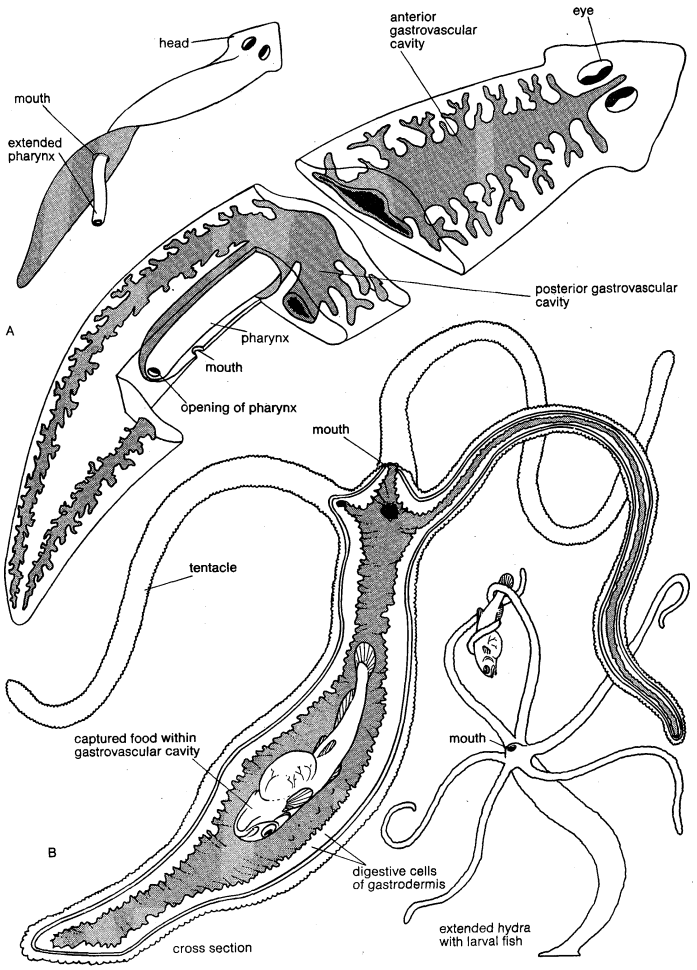


Figure 3: *Saccular systems.*
(A) Much-branched gastrovascular cavity and extruded pharynx of a planaria; (B) gastrovascular cavity of a hydra.
Adapted from W.T. Keeton, *Biological Science*, © (1967) by W.W. Norton & Company, Inc.

The gastrovascular cavity

protective and sensory covering (epithelium); those of the inner layer, or gastrodermis, which lines the central cavity of the body, act as a nutritive epithelium. The central cavity, which functions as a digestive cavity, has only one opening to the outside; the opening acts both as a mouth for ingestion of food and as an anus for egestion of wastes. Such a digestive cavity is called a gastrovascular cavity, because in many animals it has vessel-like branches that convey the contents to all parts of the body.

Once prey, captured by a hydra's tentacles, has been passed through the mouth into the gastrovascular cavity, digestive enzymes are secreted into the cavity by the gastrodermal cells, and extracellular digestion begins. In cnidarians, this extracellular digestion is limited largely to partial hydrolysis of proteins. As soon as the food has been partially disintegrated, the gastrodermal cells engulf the fragments by phagocytosis, and digestion is completed intracellularly within food vacuoles.

Many flatworms (phylum Platyhelminthes) also have gastrovascular cavities, even though their bodies are in many other ways much more complex than those of cnidarians. In planarians, for example, the mouth opens into a tubular chamber called the pharynx, which in turn leads into an extensively branched gastrovascular cavity that ramifies throughout the body (Figure 3A). As in cnidarians, some extracellular digestion occurs in the planarian gastrovascular cavity, with the small food particles then being engulfed by gastrodermal cells and digested intracellularly. The additional process of extracellular digestion frees cnidarians and flatworms from exclusive reliance on intracellular digestion.

Tubular systems. Most animals above the level of cnidarians and flatworms have a complete digestive tract; *i.e.*, a tube with two openings—a mouth and an anus. There are obvious advantages of such a system over a gastrovascular cavity, among them the fact that food moves in one direction through the tubular system, which can be divided into a series of distinct sections, each specialized for a different function. A section may be specialized for mechanical breakdown of bulk food, for temporary storage, for enzymatic digestion, for absorption of the products of digestion, for reabsorption of water, and for storage of wastes. The overall result is greater efficiency, as well as the potential for special evolutionary modifications for different modes of existence.

The earthworm tubular system

The digestive system of an earthworm is an example of a tubular system (Figure 4C). Food, in the form of decaying organic matter mixed with soil, is drawn into the mouth by the sucking action of a muscular pharynx. From the pharynx and then through a connecting passage called the esophagus, the food enters a relatively thin-walled storage chamber, or crop. Next, the food enters the gizzard, a compartment with thick muscular walls, and is ground up by a churning action, the grinding often being facilitated by bits of stone taken in with the food. The pulverized food, suspended in water, then passes into the long intestine, in which digestion and absorption take place. Most of the digestion is extracellular; cells of the intestinal lining secrete hydrolytic enzymes into the cavity of the intestine, and the end products of digestion, the simple compounds from which large molecules are formed, are absorbed. Finally, toward the rear of the intestine, some of the water is reabsorbed, and the indigestible residue is ultimately eliminated through the anus.

Not all large animals eat and grind up large pieces of food. Many are filter feeders; *i.e.*, they strain small particles of organic matter from water. Clams and many other mollusks filter water through tiny pores in their gills and trap microscopic food particles in streams of mucus that flow along the gills and enter the mouth; the mucus is kept moving by beating cilia. In such mollusks, digestion is largely intracellular, as might be expected in animals that eat microscopic food. Current theory holds that the earliest vertebrates were filter feeders. Some of the largest whales are examples of modern-day filter-feeding vertebrates; they strain small planktonic organisms from vast quantities of water.

Possession of a storage organ, such as the crop of the earthworm, enables an animal to take in large amounts of readily available food in a short time and to draw upon this stored matter over an extended period. Such a discontinuous feeding habit makes it possible for an animal to devote time to activities other than feeding. The majority of higher animals have evolved adaptations for discontinuous feeding, thereby gaining time for a behaviorally more varied existence.

Discontinuous feeding is frequently also of adaptive advantage in the feeding process itself. An animal's proper food, for example, may occur only in widely scattered locations; if it had to eat constantly to maintain itself, the animal would be unable to spend time searching for a new food supply or capturing more prey when the original supply had been depleted. The animal would thus have to live in an area in which there was an essentially unlimited and continuous source of food.

Food storage in mosquitoes

Animal food-storage organs are quite variable. In some animals they take the form of blind sacs (diverticula) branching off the digestive tract. Female mosquitoes, for example, have a large diverticulum that opens off the anterior portion of the digestive tract and runs posteriorly, occupying much of the abdominal cavity. The female mosquito locates a suitable animal, pierces its skin, and sucks blood until the diverticulum is filled. One large meal may suffice for the entire process of locating a site and laying her eggs—a matter of four or five days.

THE VERTEBRATE DIGESTIVE SYSTEM

The account below, based on the digestive system of higher mammals, proceeds compartmentally from the mouth to the anus. Other vertebrates are mentioned when they illustrate some important departure from the model.

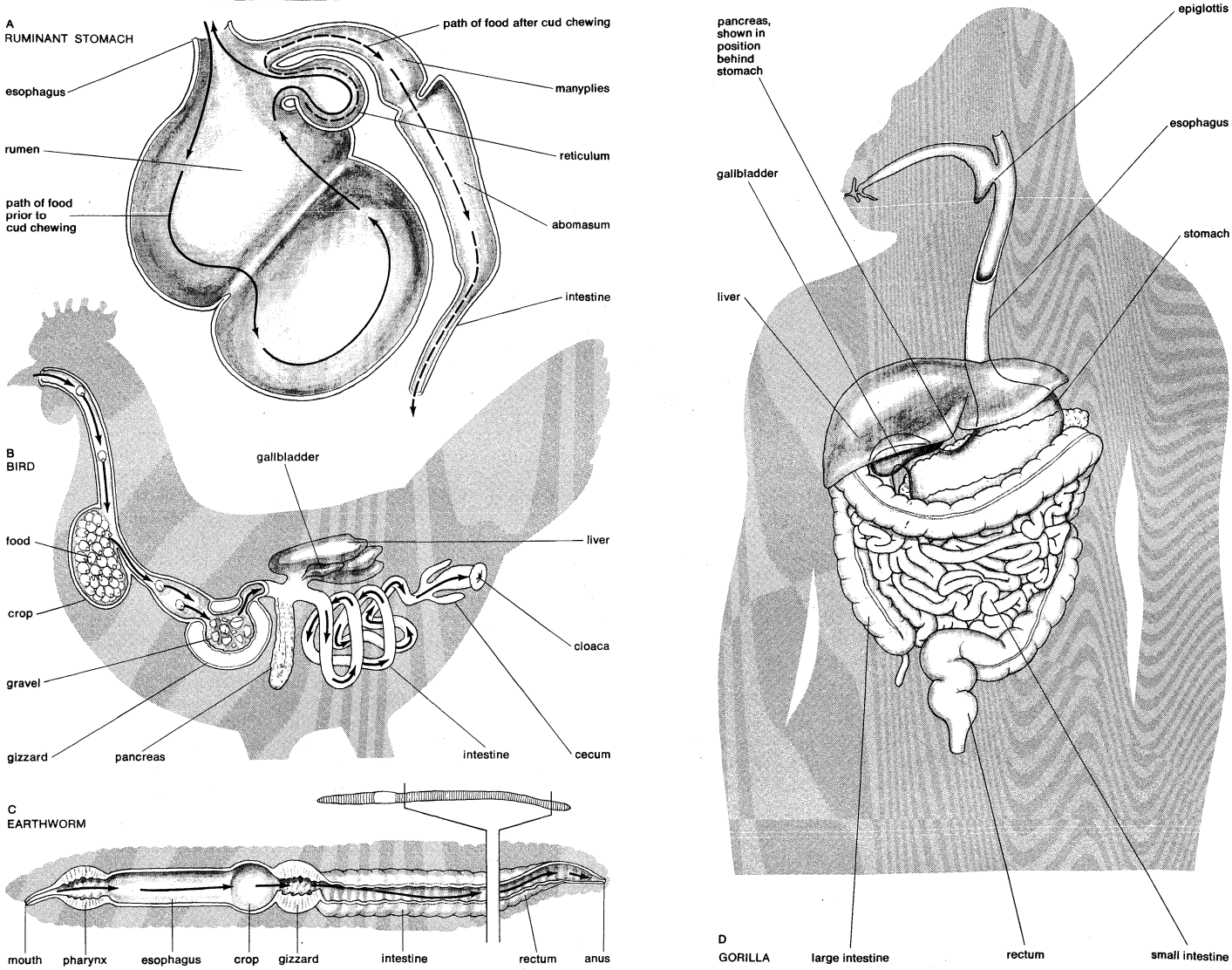


Figure 4: Vertebrate and invertebrate digestive systems.

The oral cavity and pharynx. The first chamber of the typical mammalian digestive tract is the oral cavity, which contains teeth that function in the mechanical breakdown of food by biting, shearing, and chewing.

Mammalian teeth are of several different types, each adapted to a different function. In front are the chisel-shaped incisors used for biting. Beside them are the more pointed canine teeth, specialized for tearing food. Behind the canines are premolars and molars with flattened, ridged surfaces, which function in grinding, pounding, and crushing food.

Teeth specialization

The teeth of different vertebrate species are specialized in a variety of ways. The teeth of snakes, for example, are very thin and sharp and usually curve backward; they function in capturing prey but not in chewing, for snakes swallow their food whole. The teeth of carnivorous mammals, such as cats and dogs, are more pointed than those of primates, including man; the canines are long, and the premolars lack flat grinding surfaces, being more adapted to cutting and shearing (often the more posterior molars are lost). On the other hand, such herbivores as cows and horses have very large, flat premolars and molars with complex ridges and cusps; the canines are often totally absent. Sharp pointed teeth, poorly adapted for chewing, generally characterize meat eaters such as snakes, dogs, and cats; and broad flat teeth, well adapted for chewing, characterize vegetarians. Why the difference? Very few animals can digest cellulose, yet the plant cells used as food by herbivores are enclosed in cellulose cell walls that must be broken down before the cell contents can be exposed to the action of digestive enzymes. By contrast, the animal cells in meat are not encased in nondigestible matter and can be acted upon directly by digestive enzymes. Consequently, chewing is not so essential for carnivores as for herbivores. Dogs gulp their food; cows and horses spend much time chewing. Carnivores have other problems, however—they must capture and kill prey, for which their sharp teeth, capable of piercing, cutting, and tearing, are well adapted. Man, an omnivore (an eater of plant and animal tissue), has teeth that belong, functionally and structurally, somewhere between the extremes of specialization attained by the teeth of carnivores and herbivores.

One class of vertebrates—the birds—have no true teeth. Mechanical breakdown of food is accomplished in these animals in a muscular gizzard (Figure 4B), located behind the stomach; in the gizzard, hard food is ground with grit formed of ingested rocks and pebbles.

The oral cavity has other functions in addition to those associated with the teeth. It is here that food is tasted and smelled, mixed with saliva secreted by several sets of salivary glands. The saliva dissolves some of the food and acts as a lubricant, facilitating passage through the subsequent portions of the digestive tract. The saliva of some mammals, including man, also contains a starch-digesting enzyme called amylase (ptyalin), which initiates the process of enzymatic hydrolysis; it splits starch into molecules of the double sugar maltose. Many carnivores, such as dogs and cats, have no amylase in their saliva, since their natural diet contains very little starch.

The muscular tongue of mammals manipulates the food as it is chewed, forming it into a mass called a bolus in

Function of the tongue

preparation for swallowing. The tongue pushes the bolus backward through the pharynx and into the esophagus. Food and air cross in the pharynx, which functions also as part of the respiratory passageway (see RESPIRATION; RESPIRATORY SYSTEMS). Consequently, swallowing involves a complex set of reflexes that closes off the opening into the nasal passages and trachea (windpipe), thereby forcing the food to move into the esophagus. When these reflexes do not occur in proper sequence, food enters the wrong passage, causing gagging or choking.

The esophagus. The esophagus, a tube in the upper portion of the abdominal cavity, extends from the pharynx through the neck and chest to the stomach. Food moves quickly through the esophagus, pushed along by waves of muscular contraction in a process called peristalsis. The contraction of circular muscles in the wall of the esophagus just behind the food bolus squeezes the food along the digestive tract.

At the junction between the esophagus and the stomach is a special ring of muscle called a sphincter, which, when contracted, closes the entrance to the stomach. It is normally closed, thus preventing the contents of the stomach from moving back into the esophagus, but it opens when a peristaltic contraction from the esophagus reaches it.

In some vertebrates, the esophagus is not merely a tubular connection between the pharynx and the stomach. In many birds, for example, an expanded region of the esophagus anterior to the stomach forms a thin-walled crop (Figure 4B), functionally analogous to the earthworm's crop, which, rather than the stomach, is the bird's principal organ for the temporary storage of food. Some birds also use the crop to carry food to their young; they fill it with food and then fly to the nest, whereupon they disgorge the food for their young.

The four "stomachs" of ruminants

Ruminant mammals, such as the cow, are often said to have four "stomachs" (Figure 4A). Actually, the first three of these chambers—rumen, reticulum, and omasum—are thought to be derived from sections of the esophagus. Vast numbers of bacteria and protozoans live in the rumen and reticulum. When food enters these chambers, the microbes begin to digest and ferment it, breaking down not only protein, starch, and fats, but cellulose as well. The larger, coarser material is periodically regurgitated as the cud; after further chewing, the cud is reswallowed. Slowly the products of microbial action, and some of the microbes themselves, move into the cow's true stomach and intestine, in which further digestion and absorption take place. Since the cow, like other mammals, has no cellulose-digesting enzymes of its own, it relies upon the digestive activity of these symbiotic microbes in its digestive tract. Much of the cellulose in the cow's herbivorous diet, which otherwise would have no nutritive value, is thereby made available to the cow.

The stomach. The primate stomach lies in the upper portion of the abdomen, just below the lower ribs (Figure 4D). It is a muscular sac, with a wall composed of three layers: an inner mucous membrane of connective tissue and many glands, a thick middle layer of smooth muscle, and an outer layer of connective tissue. The muscle layer contains fibres running around the stomach, others running longitudinally, and still others oriented diagonally. Thus, the stomach is capable of a great variety of movements. When it contains food, the stomach is swept by powerful waves of contraction, which churn the food, mixing and breaking it. In this manner, the stomach supplements the mechanical action of the teeth.

The glands of the stomach lining are of several types. Some secrete mucus, which covers the stomach lining (hence the name mucosa, or mucous membrane, for the inner layer of the stomach wall); others secrete gastric juice, a mixture of hydrochloric acid and digestive enzymes.

The principal enzyme of the gastric juice is pepsin, which digests protein. Unlike most protein-digesting enzymes, pepsin functions only in a strongly acid medium. This acidic requirement is characteristic of vertebrates; most invertebrates do not have proteolytic enzymes that are active in strongly acid solutions. Pepsin does not completely hydrolyze protein to its amino acid components. It splits the so-called peptide bonds that link certain amino acids, particularly tyrosine and phenylalanine. Proteolytic enzymes such as pepsin are relatively specific for certain amino acids; that is, they will react only with certain ones. This is because enzymes have a specific region, called the active site, at which the catalytic action of the enzyme occurs. The fit between an enzyme, such as pepsin, and an amino acid, such as tyrosine, must be a good one, therefore, or the reaction cannot take place. The structural configuration around the various peptide bonds in a protein varies, depending on which two amino acids are joined by the bond. Consequently, some bonds may not fit the active site of a particular enzyme.

Protection of the mucosa against its own enzymes

Discussion of protein digestion immediately raises the question of why the digestive tract itself—composed mostly of protein—is not digested by the proteolytic enzymes. There are two reasons. First, the gastric glands do not secrete active pepsin; instead, they secrete pepsinogen, which has no proteolytic activity and which, so long as it is stored in the glands of the stomach wall, poses no threat to that wall. Pepsinogen is changed to active pepsin only after exposure to acid in the lumen (cavity) of the stomach. Second, the wall of the digestive tract is covered with mucus, which apparently shields it from the enzymes it secretes. Sometimes, however, this defense breaks down, and the enzymes do begin to eat away at the lining; the resulting sore is known as an ulcer.

In addition to pepsin, the gastric juice of some mammals—*e.g.*, calves (but not humans)—contains another enzyme, rennin, which clumps milk proteins, thus taking them out of solution and making them more susceptible to the action of proteolytic enzymes. Few animals other than mammals have such an enzyme. Although rennin aids digestion, it is not itself a digestive enzyme, since it does not catalyze a hydrolysis reaction.

The small intestine and the pancreas. Upon leaving the stomach, the partially digested food, in a soupy mixture called chyme, passes through the pyloric sphincter into the small intestine, that portion of the digestive tract in which most of the digestion and absorption takes place. The first section of the small intestine, attached to the stomach, is called the duodenum; it leads into a long, coiled section lying lower in the abdominal cavity. The relative length of the small intestine differs in different species. It is usually very long and much coiled in herbivores, shorter in carnivores, and of medium length in omnivores. These differences, like those of the teeth, are correlated with the difficulty of digesting plant material. Even if the cellulose of plant cells has been well broken down, it is mixed with the digestible portions of the cells and interferes with their disintegration by digestive enzymes. This interference makes digestion and absorption of plant material much less efficient than animal material, with the result that, in animals with a herbivorous diet, a long intestine is an adaptive advantage because a maximum amount of nutrients can be extracted.

In primates, the small intestine, where absorption of the products of digestion occurs, has special structural adaptations that increase its absorptive surface area. Clearly, its great length plays a role. But examination of its internal surface (Figure 5A) reveals three other modifications that increase the surface area to an even greater degree: (1) the lining of the intestine is thrown into numerous folds and ridges; (2) small, fingerlike outgrowths, called villi, cover the entire surface of the mucosa; and (3) the individual epithelial cells covering the folds and villi have a border consisting of countless, closely packed, cylindrical projections called microvilli. Microvilli are revealed only by the electron microscope. Thus, the total internal surface of the small intestine, including folds, villi, and microvilli, is extremely large.

The extensive surface area of the intestine

Other vertebrates show other adaptations for increasing the absorptive surface area of the small intestine. For instance, it is not unusual for special blind sacs, called ceca, to branch from the anterior end of the small intestine in certain fishes and from the posterior end in many birds (Figure 4B). Another adaptation is the spiral valve of many primitive fishes, including sharks (Figure 5B). The spiral valve is an epithelial fold extending the length of

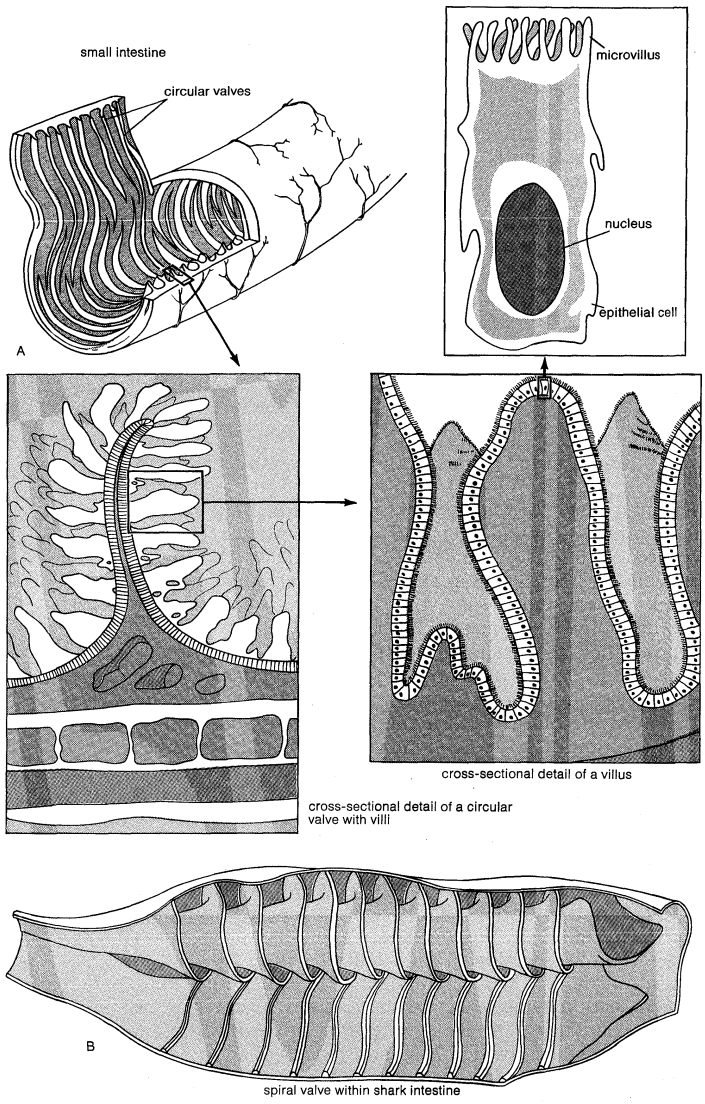


Figure 5: Structural modifications for increasing the surface areas of the small intestine.
Adapted from *Human Biology* by G.A. Baitsell; Copyright 1950; used with permission of McGraw-Hill Book Co.

the intestine. Like a screw tightly enclosed in a tube, it forms a spiral in the intestine, to whose wall its base is attached; hence, food cannot move in a straight path but follows the spiral of the valve, thereby contacting more epithelial surface than it could by moving straight through a smooth tubular intestine of the same length.

When partially digested food passes from the stomach into the duodenum, it stimulates the release of a large number of different digestive enzymes into the lumen of the intestine. These enzymes are secreted from two principal sources, the pancreas and the intestinal glands. The pancreas is a soft glandular organ lying posterior to the stomach; it originates in the embryo as an outgrowth of the digestive tract and retains a connection to the duodenum via the pancreatic duct. When food enters the duodenum, the pancreas secretes a mixture of enzymes that flow through the pancreatic duct into the duodenum. Included in this mixture are enzymes that digest all three principal classes of foods—carbohydrates, fats, and proteins—as well as some that digest nucleic acids.

Pancreatic enzymes

One of the pancreatic enzymes is pancreatic amylase (sometimes also called diastase, or amylopsin), which acts like salivary amylase, splitting starch into maltose. It is far more important than salivary amylase, however, for it digests most of the starch. Lipase is the fat-digesting enzyme of the pancreas. It splits molecules of fat into glycerol and fatty acids; most fat digestion is catalyzed by this pancreatic enzyme. Two important proteolytic enzymes from the pancreas are trypsin and chymotrypsin. Like pepsin, they are incapable of splitting all the linkages between the amino acids in a protein molecule. Each cleaves only the linkages adjacent to certain specific amino acids: trypsin splits the peptide bonds adjacent to lysine and arginine; chymotrypsin, those adjacent to tyrosine, phenylalanine, tryptophan, methionine, and leucine. Like pepsin, both trypsin and chymotrypsin are secreted in inactive forms, called trypsinogen and chymotrypsinogen, respectively. Trypsinogen is converted into active trypsin in the intestine by enterokinase, an enzyme secreted by intestinal glands. Chymotrypsinogen is converted into active chymotrypsin by trypsin.

The combined action of pepsin in the stomach and of trypsin and chymotrypsin in the intestine results in a splitting of proteins into fragments of varying lengths but does not produce many free amino acids. A great variety of other enzymes, each highly specific in its action, completes the digestion of proteins. One, for example, cuts off the amino acid at one end of a chain; another hydrolyzes the linkage of the amino acid at the other end. Still others break apart pairs of amino acids in specific ways. Most of these enzymes are secreted by glands in the intestine, but some are produced in the pancreas.

Just as certain enzymes from the intestinal glands complete the digestion of protein, other intestinal enzymes complete the digestion of carbohydrate begun by salivary and pancreatic amylase. These enzymes split double sugars into simple sugars. For example, maltase splits maltose, sucrase splits sucrose, and lactase splits lactose.

The liver. The liver is a relatively large organ occupying much of the space in the upper part of the abdomen (Figure 4D). It performs many important functions, among them the production of bile, which aids in fat digestion. Bile, produced throughout the liver, is collected by a series of branching ducts; it empties into a small storage organ, the gallbladder, located on the surface of the liver. When food enters the duodenum, the muscular wall of the gallbladder is stimulated to contract, and bile is forced down the gall duct into the duodenum. Bile is not a digestive enzyme; it is a complex solution of bile salts, bile pigments, and cholesterol. The bile salts act as emulsifying agents, causing large fat droplets to separate into many tiny droplets suspended in water (an action much like that of a detergent). The many small fat droplets expose much more surface area to the action of lipase than would a few large droplets. Bile salts apparently also aid in fat absorption. The bile pigments and cholesterol play no obvious role in digestion. The pigments, produced by destruction of red blood cells in the liver, give the characteristic brown colour to feces.

The function of bile salts

The large intestine. In primates the small intestine joins the large intestine (or colon) usually in the lower right portion of the abdominal cavity (Figure 4D). A blind sac, the colic cecum, projects from the large intestine near the point of juncture. A small fingerlike process, the appendix, extends from the tip of the cecum. The cecum of some mammals, particularly herbivores (*e.g.*, horses), is large and contains many micro-organisms capable of digesting cellulose.

From the cecum, the large intestine ascends on the right side to the midregion of the abdominal cavity, then crosses to the left side, and descends again. The three sections thus formed are frequently termed the ascending, transverse, and descending colons. The details of shape and length vary greatly from species to species.

One of the chief functions of the colon is reabsorption of much of the water resulting from the digestive process. Occasionally, the intestine becomes irritated, and abnormally strong peristalsis moves material through it too fast for sufficient water to be reabsorbed, resulting in watery stools, or diarrhea. Conversely, if material is moved too slowly, excessive water is reabsorbed, and constipation results. A good diet contains sufficient roughage (indigestible material, primarily cellulose) to provide the bulk needed to stimulate sufficient peristalsis in the large intestine to prevent constipation. A second function of the colon is the excretion of certain salts, such as those of calcium and iron, when their concentra-

tion in the blood is too high; the salts are excreted into the colon and are eliminated from the body in the feces.

The large intestine contains large numbers of bacteria, which live on the undigested food in the colon. Their significance is not yet fully understood; about half the dry weight of the feces consists of masses of bacteria.

The rectum and cloaca

The last portion of the large intestine functions as a storage chamber for the feces until defecation. In most mammals (including man) this chamber, called the rectum, is exclusively a part of the digestive system and terminates in a muscular sphincter that controls the anal opening. In many other vertebrates (*e.g.*, amphibians, reptiles, birds), however, the digestive, excretory, and reproductive systems share one terminal chamber; the walls of the chamber, called a cloaca (Figure 4B), frequently have very powerful water reabsorbing properties.

EMBRYOLOGY AND EVOLUTION OF THE VERTEBRATE DIGESTIVE SYSTEM

Embryonic development. In amphioxus, an invertebrate member of the Chordata (the phylum to which all vertebrates belong), early divisions of the fertilized egg cell give rise to an embryo that is hollow and nearly spherical. Then an invagination (infolding) of cells at the vegetal (yolk) pole of the embryo converts the initially single-layered embryo into a two-layered one, a process called gastrulation. The new inner layer of cells, called endoderm (sometimes entoderm), surrounds a cavity, the archenteron, which has an opening to the exterior at the point at which invagination occurred; this opening is called the blastopore. The archenteron eventually becomes the cavity of the digestive tract, and the blastopore becomes the anus; the mouth arises as a new opening. In some invertebrates the reverse is true: the blastopore becomes the mouth, and the anus is the new opening.

The early stages of embryonic development in most vertebrates are not as simple as in amphioxus, largely because the egg cells contain much yolk or, in mammals, undergo specialized changes preparatory to implantation in the uterus. Thus, gastrulation is seldom a simple involution at the vegetal pole, and the blastopore, if indeed a "pore" really appears at all, usually becomes overgrown with cells. Nevertheless, in all vertebrate embryos, an endodermal-lined cavity arises by some process that may be regarded as analogous to gastrulation in amphioxus, and this cavity develops into the digestive tract. Ordinarily, endoderm lines the yolk sac; forms a tube, called the foregut, that pushes forward into the head; and forms a second tube, the hindgut, that pushes into the posterior part of the embryonic body. Eventually, the surface tissue (ectoderm) of the embryo forms a small anterior invagination, the stomodeum, that meets the end of the foregut, and a similar posterior invagination, the proctodeum, that meets the end of the hindgut. Rupture of the tissues separating the stomodeum from the foregut and the proctodeum from the hindgut forms a digestive tract with two openings to the exterior.

Embryonic derivations

It is apparent from the above description that short sections at both the anterior and the posterior ends of the digestive tract are of ectodermal origin. These correspond roughly to the oral cavity and to the anal canal, respectively. All the rest of the digestive tract, from the pharynx through the large intestine, is of endodermal origin. However, only the lining of the digestive tract is endodermal; the walls contain layers of muscle and connective tissue, which are of middle layer (mesodermal) origin. The endodermal lining gives rise by outpocketing to numerous organs, including the thyroid gland, gills or lungs, thymus, liver, pancreas, and urinary bladder.

Evolutionary development. In amphioxus, the digestive tract consists of only three components: the oral cavity, the pharynx, and a tubular postpharyngeal gut without subdivisions. The same condition holds in the most primitive living vertebrates, the cyclostomes (lampreys and hagfishes). In higher vertebrates, however, the postpharyngeal gut is almost always subdivided into a series of regions both anatomically and functionally distinct. By far the most common is the esophagus–stomach–small intestine–large intestine–rectum sequence.

The oral cavity and pharynx vary considerably among the vertebrate classes. The variation correlates with the evolutionary changes in the respiratory system that accompanied the rise of terrestrial forms from aquatic ancestors. In most modern-day bony fishes, the nares (corresponding to a mammal's nostrils) function only as entrances to the olfactory organs, there being no connection between them and the mouth, as occurs in mammals. The structure called the palate, which in mammals separates the nasal and oral cavities, does not exist in fishes. Respiratory water is taken directly into the mouth and then forced back into the pharynx, where it flows across gills located in a series of slits leading from the pharynx to the exterior. Thus, the pharynx, with its gills, is an extremely important chamber in these animals.

The terrestrial vertebrates, which must extract oxygen from air instead of from water, evolved a second major function for the nares that they inherited from their piscine ancestors. While retaining a smell function, these openings became the principal entrance of air for breathing. In amphibians—the earliest land vertebrates—air enters the external nares (nasal openings) and then passes through the internal nares, which are evolutionarily newer openings, into the front of the oral cavity, whence it moves into the pharynx and then into the trachea. There being no palate, no separate nasal cavity exists in these animals; both the oral cavity and the pharynx are common passages for the digestive and respiratory systems.

In most reptiles and birds, a pair of longitudinal folds in the roof of the oral cavity forms a passage that leads air from the internal nares to the pharynx. Complete separation of nasal and oral cavities by a palate, however, is found only in crocodilians and in mammals. In mammals, the bony, hard palate is supplemented posteriorly by a thick, membranous, soft palate.

Changes in the esophagus

Having lost its gills and its former importance as the area of gas exchange, the pharynx of terrestrial vertebrates is usually a short and relatively unimportant link in the digestive tract, simply leading to the esophagus. In most bony fishes, the esophagus is a short connecting link between the large pharynx and the stomach, but during evolution, with the reduction of the pharynx in land vertebrates and the relegation of the stomach to the abdominal portion of the trunk, posterior to the lungs, the esophagus became longer and more prominent. It is usually a simple tube, however, except in the case of the avian crop and the ruminant extra "stomachs."

Most vertebrates above the level of the cyclostomes have a stomach (though there is none in such fishes as the chimaeras, lungfishes, and a few bony fishes), but its size and shape are quite variable. Perhaps the greatest departure from the general plan is seen in birds, in which the anterior storage part of the stomach is small (the crop taking over most of this function), and the posterior part is specialized as the thick-walled muscular gizzard.

The intestine varies greatly among the various vertebrates in length, in specializations for increasing absorptive surface area, and in the degree of distinction between small intestine and large intestine. The chief functions, however, remain enzymatic digestion and absorption. In all cases, the liver and the pancreas, both of which arise as outpocketings from the anterior part of the small intestine just posterior to its junction with the stomach, play major roles in the digestive process.

The final chamber of the digestive tract is a common cloaca in elasmobranch fishes and in lungfishes, but in most ray-finned fishes there is a rectum instead; *i.e.*, the urinary and reproductive tubes, which do not join the digestive tube, have their own separate opening to the exterior. In this regard, then, the modern-day ray-finned fishes are more specialized than amphibians, reptiles, and birds, which retain a cloaca, presumably inherited from a primitive fish ancestor. A cloaca is also retained in the egg-laying mammals (monotremes), and, in a much reduced form, in the pouched mammals (marsupials). Even in placental mammals, a short-lived cloaca appears in the embryo, but the urogenital ducts eventually develop their own openings, and, as a consequence of this, mammalian adults have a rectum rather than a cloaca.

BIODYNAMICS OF THE VERTEBRATE DIGESTIVE SYSTEM

Control of salivary secretion. The salivary glands are controlled by the two divisions of the autonomic nervous system (sympathetic and parasympathetic), and it is generally held that this innervation is exclusively responsible for regulation of the glands' secretory activity. No hormone appears to be involved.

There is normally some secretion of saliva all the time, whether or not food is in the mouth. When something touches the gums, the tongue, or some region of the mouth lining, or when chewing occurs, however, the rate of secretion rises. The stimulating substance need not be food—dry sand in the mouth, or even simply moving the jaws and tongue when the mouth is empty, are effective in increasing the salivary flow. This coupling of direct stimulation to the oral mucosa with increased salivation is known as the unconditioned salivary reflex. When an individual learns that a particular sight, sound, smell, or other stimulus is regularly associated with food, that stimulus alone may suffice to stimulate increased salivary flow. In other words, the mouth waters without any direct stimulation of the oral mucosa; this response is known as the conditioned salivary reflex.

Drugs that affect salivary flow

A variety of drugs are capable of increasing or decreasing salivary flow. Among those increasing the flow are sympathomimetic agents, drugs that produce effects similar to those elicited by the sympathetic nervous system (*e.g.*, adrenaline, noradrenaline, amphetamine), and parasympathomimetic agents, drugs that mimic the effects of the parasympathetic nervous system (*e.g.*, acetylcholine, pilocarpine). Among the drugs that decrease salivary flow are antagonists of adrenaline and noradrenaline (*e.g.*, ergotamine, dibenamine) and antagonists of acetylcholine (*e.g.*, atropine, scopolamine).

Control of gastric secretion. The secretion of gastric juice is regulated by both neural and hormonal mechanisms. Among the former are parasympathetic pathways (from the vagus nerve), which provide direct stimulation to the gastric mucosa, and sympathetic pathways, which influence gastric secretion indirectly via their control over the blood vessels that supply the mucosa. The hormonal control primarily results from the stimulatory action of gastrin, a hormone produced by certain cells in the mucosa of the pyloric region of the stomach, and from the inhibitory action of enterogastrone, a hormone produced by cells in the wall of the duodenum upon their stimulation by fats.

Control of gastric secretion in response to ingested food is customarily divided into three phases—cephalic, gastric, and intestinal. During the cephalic phase, stimulation of the mucosal glands is caused entirely by nervous impulses from the vagus nerve. The flow of gastric juice may be initiated by a variety of stimuli, even by merely thinking about food. If the vagus nerve is cut, this phase of gastric secretion is eliminated.

The gastric phase begins after food has reached the stomach. It is apparently controlled by gastrin and, to a lesser extent, by the vagus nerve. Among the most effective stimuli for gastrin release are various compounds from partly digested meat; however, mechanical stimulation by indigestible food particles or, in laboratory experiments, simple distension of the stomach by a balloon results in some release of gastrin. Though produced by cells that are located near its target cells, gastrin does not move directly from the one to the other; instead, it is picked up by the blood and carried to the gastric glands, where it stimulates increased secretion of gastric juice.

In the intestinal phase of gastric secretion, certain food substances in the duodenum stimulate release of regulatory chemicals that influence the gastric mucosa. This function was mentioned earlier, regarding the action of fats in stimulating release of enterogastrone, which inhibits secretion of gastric juice. Other food materials apparently cause release of a substance that stimulates secretion of gastric juice, but its nature is not yet known.

The influence of emotions

In many mammals, emotional states may markedly affect gastric secretion; fear, anger, or frustration, for example, may inhibit gastric flow. Chronic anxiety, however, tends to increase the flow, with the result that the stomach contents become more acid, which may lead to formation of gastric or duodenal (peptic) ulcers. Though both gastric and duodenal ulcers can be produced experimentally by exposure of the mucosa to a constant drip of hydrochloric acid, or by treatment with the compound histamine, which stimulates the acid-secreting cells, the causative connection in nonexperimental situations between high gastric acidity and ulceration is much clearer in duodenal ulcers than in gastric ulcers.

Control of pancreatic secretion. The secretion of pancreatic juice is regulated by both neural and hormonal mechanisms. Neural mechanisms, dependent upon sympathetic and parasympathetic pathways, can elicit increased flow both before and after food reaches the duodenum. By contrast, hormonal mechanisms act only after certain substances move into the duodenum and stimulate specific cells of the duodenal mucosa.

The hormones produced by the duodenal mucosa are secretin and pancreozymin. Secretin is released in response to dilute hydrochloric acid, whereas release of pancreozymin is initiated by various products of partial protein digestion. Secretin stimulates the pancreas to secrete the water and bicarbonate components of pancreatic juice, and pancreozymin stimulates secretion of the enzyme components.

Control of intestinal enzyme secretion. Local mechanical or chemical stimulation of the intestinal mucosa is the most effective stimulus for secretion of intestinal juices. Parasympathetic nerves are probably the chief elements involved in bringing about the response. There is some evidence, however, that a hormone, tentatively called enterocrinin, may also play a role.

Control of intestinal movements. The intestine undergoes varied movements, of which three are described below: rhythmic segmenting contractions, peristalsis, and mass movements.

Stationary intestinal contractions

Segmenting contractions, especially prominent in the small intestine, are rhythmic alternations of contraction and relaxation at regularly spaced intervals along the intestine. They function in agitating the gut contents, thereby facilitating the subdivision of food particles, the mixing of food with intestinal secretions, and the bringing of digestive products into direct contact with the absorptive mucosa. These contractions appear to be myogenic; *i.e.*, they originate in muscular tissue without the necessity of nervous stimulation (although nervous stimulation may increase or decrease the amplitude of such contractions).

Peristaltic movements, which have already been described, are waves of contraction that move along the intestine. They push the gut contents along their prescribed path. Peristaltic movements do not occur as smooth waves progressing along a relaxed intestine from one end to the other at regular intervals. The movement is actually superimposed upon the rhythmic segmenting contractions, the individual waves often travelling only a few inches or a few feet and the waves appearing at irregular intervals. Both mechanical and chemical stimuli can promote peristalsis; their effectiveness is related, in part, to the irritability of the intestine and its muscle tone. Irritant cathartic drugs may trigger several successive, unusually strong peristaltic waves that sweep without interruption along the entire length of the intestine.

Mass movements are the principal propulsive movements of the large intestine. They involve simultaneous contractions of large segments of the colon and hence are not true peristaltic waves. The contractions are exceedingly powerful and can therefore push along hard, dry feces. When a mass movement pushes feces into the rectum, the increase in pressure within the rectum stimulates the urge to defecate. Since eating is a powerful stimulant to mass movements in the colon, the desire to defecate often occurs shortly after food is taken.

BIBLIOGRAPHY. E.F. ANNISON and D. LEWIS, *Metabolism in the Rumen* (1959), a review of digestion in ruminants; L.B. AREY, *Developmental Anatomy*, 7th ed. (1965), a college textbook that covers the embryological development of the vertebrate digestive tract in ch. 11; B.I. BALINSKY, *An Introduction to Embryology*, 3rd ed. (1970), a general embryology textbook, with coverage of the digestive system in ch. 15; R.D.

BARNES, *Invertebrate Zoology*, 2nd ed. (1968), a textbook covering the major invertebrate animal phyla, with details on the digestive system of each particular group; D.P. CUTHBERTSON and A.T. PHILLIPSON in *Biochemistry and Physiology of Nutrition*, ed. by G.H. BOURNE and G.W. KIDDER, vol. 2, pp. 128–161 (1953), a review of the microbiology of digestion; H. HERAN, "Ein Beitrag zur Verdauungsphysiologie von *Lumbricus Terrestris* L.," *Z. Vergl. Physiol.*, 39:44–62 (1956), an account of digestion in earthworms; J.B. JENNINGS, "Studies on Feeding, Digestion, and Food Storage in Free-Living Flatworms (Platyhelminthes: Turbellaria)," *Biol. Bull.*, 112:63–80 (1957); W.T. KEETON, *Biological Science*, 2nd ed. (1972), an introductory college text, with treatment of digestion in ch. 5; J.E. MORTON, *Guts: The Form and Function of the Digestive System* (1967); J.A.C. NICOL, *The Biology of Marine Animals*, 2nd ed. (1967), includes a general account of feeding and digestion in marine organisms; C.L. PROSSER and F.A. BROWN, *Comparative Animal Physiology*, 2nd ed. (1961), a textbook that compares the digestive systems of a variety of animal phyla in ch. 5; A.S. ROMER, *The Vertebrate Body*, 4th ed. (1970), a clear discussion of the anatomy and evolution of the major vertebrate systems, including the digestive system in ch. 11–12; H.J. VONK, "Comparative Physiology (Nutrition, Feeding, and Digestion)," *A. Rev. Physiol.*, 17:483–498 (1955); C.M. YONGE, "Feeding Mechanisms in the Invertebrates," *Biol. Rev.*, 3:21–76 (1928); "Evolution and Adaptation in the Digestive System of the Metazoa," *ibid.*, 12:87–115 (1937); and *Tabulae Biologicae*, vol. 21, pt. 3–4 (1954), a review on feeding and digestion in various invertebrates.

(W.T.Ke.)

Digestive System, Human

The digestive system consists of (1) the digestive tract—the series of structures and organs through which food passes during its processing into forms absorbable into the bloodstream and also the structures through which solid wastes pass in the process of elimination—and (2) other organs that contribute juices necessary for the digestive process.

The digestive tract begins at the lips and ends at the anus. It consists of the mouth, or oral cavity, with its teeth, for grinding the food, and its tongue, which serves to knead the food, mix it with saliva, and start it on its way to the stomach; the throat, or pharynx; the esophagus, or gullet; the stomach; the small intestine, consisting of the duodenum, the jejunum, and the ileum; and the large intestine, consisting of the cecum, a closed-end sac connecting with the ileum, the ascending colon, the transverse colon, the descending colon, and the sigmoid colon, which terminates in the rectum. Glands contributing digestive juices include the salivary glands, the gastric glands in the stomach lining, the pancreas, and the liver and its adjuncts—the gallbladder and bile ducts.

Digestive systems in animals other than man are described in the article DIGESTION AND DIGESTIVE SYSTEMS. That article and the article DIGESTION, HUMAN contain descriptions of the processes of digestion, including the hormonal control of the secretion of digestive juices. The articles ENDOCRINE SYSTEM, HUMAN; HORMONE; and ENDOCRINE SYSTEMS also touch upon hormonal control of digestion. The present article is focussed on the form and structure of the parts of the human digestive system.

MOUTH AND ORAL STRUCTURES

The mouth. The mouth, or oral cavity, forms the entrance to the alimentary, or digestive, canal, also called the digestive tract. The structures within the mouth and the walls of the oral cavity are highly specialized in order to serve the functions of chewing food, mixing it with saliva, tasting it, and initiating the process of swallowing. (Another function for which the structures of the mouth have been adapted is speaking.) The exterior opening of the mouth is bordered by the lips. Behind, the oral cavity opens into the pharynx, or throat. The walls of the mouth are formed by the lips in front and the cheeks on each side. The roof is composed of the hard and the soft palate. The floor of the mouth is formed by the tongue, the tissue beneath it, and the lower jaw (mandible).

The walls, roof, and floor of the mouth

The lips. The lips are the fleshy folds surrounding the opening of the mouth. At their corners, on either side, they are continuous with the cheeks. The muscle of the lips, the orbicularis oris, accounts for the extreme mobility of these structures. The outer surface of the lips is covered with skin; the inner surface, with mucous membrane. The red margin of the lips, called the vermilion border, has a covering that is intermediate between skin and mucous membrane. The mucous membrane, or mucosa, of the lips and cheek has a characteristic reddish hue that is caused by the rich network of blood vessels that is visible through the thick but transparent covering layers of squamous, or scalelike, cells. The tissue beneath the covering of the lips contains numerous glands. Each lip at its midline is connected to the gum with a fold of mucous membrane, called a frenulum. The frenula of the lips can be felt with the tip of the tongue and can easily be seen when the lips are rolled back.

The lips are extremely sensitive; the upper and lower lips obtain their nerve supply, respectively, from branches of the maxillary and mandibular divisions of the trigeminal, or fifth cranial, nerve. There is a bountiful supply of arterial blood, from branches of the labial (lip) and buccal (cheek) arteries. Venous drainage is provided by tributaries of the anterior facial vein. The lips are also amply supplied with lymphatic vessels.

The cheeks. The cheeks, the sides of the mouth, are continuous with the lips and have a similar structure; the muscle of the cheeks is the buccinator (the "trumpeter") muscle. A distinct fat pad is found in the subcutaneous tissue (the tissue beneath the skin) of the cheek; this pad is especially large in infants and is known as the sucking pad. On the inner surface of each cheek, opposite the second upper molar tooth (the second grinder tooth from the end), is the slight elevation that marks the opening of the parotid duct, leading from the parotid salivary gland, which is located in front of the ear. Just behind are four to five mucus-secreting glands, the ducts of which open opposite the last molar tooth. The mucous membrane of the cheek, like that of the rest of the mouth, is composed of stratified squamous epithelium (a covering composed of layers of scalelike cells) and continues onto the gums.

The roof of the mouth. The roof of the mouth is concave both from side to side and from front to back and is formed by the hard and soft palate. The hard palate is formed by the horizontal portions of the two palatine bones and the palatine portions of the maxillae, or upper jaws. The hard palate is covered by a thick, somewhat pale mucous membrane that is continuous with that of the gums and is bound to the upper jaw and palate bones by firm fibrous tissue. In the midline is a slight ridge, the palatine raphe, which ends in front in a small elevation, the palatine papilla; this marks the position of the anterior palatine canal, which carries an artery. From the anterior part of the raphe, five or six transverse ridges, or rugae, of the mucous membrane run outward.

The hard and soft palates

The soft palate is continuous with the hard palate in front and posteriorly has a free margin. The uvula, which varies greatly in size and shape, is a projection of the soft palate and hangs free from its rear margin. The soft palate is composed of a strong, thin, fibrous sheet, the palatine aponeurosis, and the glossopalatine and pharyngopalatine muscles. Both the oral and pharyngeal (mouth and throat) surfaces are covered by mucous membrane that contains many mucous glands.

The floor of the mouth. The floor of the mouth can be seen only when the tongue is raised. In the midline is a prominent fold (frenulum linguae) like that which binds each lip to the gums, and on each side of this is a slight elevation called a sublingual papilla, onto the summit of which the ducts of the submaxillary salivary glands open. Running outward and backward from this is a ridge (the plica sublingualis) that marks the upper edge of the sublingual (under the tongue) salivary gland and onto which most of the ducts of that gland open.

The gums and teeth. *The gums.* The gums consist of mucous membranes connected by thick fibrous tissue to the membrane surrounding the bones of the jaw. Around the base of the crown (exposed portion) of each tooth the gum membrane rises to form a little collar. The gum tissues are rich in blood vessels, receiving branches from the alveolar arteries; these vessels, called alveolar from their relationship to the alveoli dentales, or tooth sockets,

also supply the teeth and the spongy bone of the upper and lower jaws, in which the teeth are lodged. The veins and lymphatics correspond essentially to the arteries.

Teeth. The teeth are hard white structures found in the mouth of man and many animals and are used for mastication of food, seizing and holding objects, combat, cutting, chiselling, and other purposes. Each tooth consists of a crown and one or more roots. The crown is the functional part that is visible above the gum. The root is the unseen portion that supports and fastens the tooth in the jawbone. The shapes of the crowns and the roots vary in the different parts of the mouth and from one animal to another. Man normally has two sets of teeth during his lifetime. The first set is acquired gradually between the ages of six months and two years. As the jaws grow and expand, these teeth are replaced one by one by the teeth of the secondary set. The first set is known as the milk, deciduous, or primary dentition. The teeth on one side of the jaw are essentially a mirror image of those located on the opposite side. The upper teeth differ from the lower and are complementary to them. There are five deciduous teeth and eight permanent teeth in each quarter of the mouth, making a total of 32 permanent teeth to succeed the 20 deciduous ones.

Distinctions between deciduous and permanent dentitions

The deciduous dentition differs from the permanent in its smaller size, whiter colour, greater constriction of the necks of the teeth, and in the fact that the roots of the molars are widely spread to provide a wider base in order to accommodate the developing permanent teeth (the molars of the deciduous dentition are replaced by the permanent premolars). In the permanent dentition there are two incisors in each quarter of the mouth. These are used primarily for biting and cutting and have a delicate tactile sense that enables them to be used for identifying objects in the mouth by nibbling. The biting portion is wide and thin, in a sort of chisel shape or cutting edge. The upper central incisor is the largest. The lateral incisor, next in line, is smaller and sometimes irregular in form. The lower incisors are smaller and similar to each other. They bite against the upper incisors either edge to edge or slightly inward on the tongue side. An edge-to-edge biting position is acquired more and more as the teeth wear down. The cuspids, four in number, are next to the incisors and at the corner of the mouth, the third tooth from the midline. The cuspid, or canine, is frequently pointed and rather peglike in shape. It is absent in many mammals that have developed protective horns; in some other animals, it is a great defensive weapon by virtue of its size. In man it has the function of cutting and tearing. Its length and position are such that it does not normally protrude beyond the level of the other teeth, and this permits a side to side chewing motion that is not possible in the other primates, which have long canines. The two bicuspids, or premolars (total number, eight), are next in line. They replace the baby molars but have a somewhat different function. Each has two cusps, or elevations; the lower second bicuspid frequently has a small third cusp. The upper first bicuspid generally has two roots and sometimes one; the second generally has one root, sometimes two; the lower bicuspids generally have one root each. The three molars in each quadrant (total number, 12), the teeth farthest back in the mouth, are used exclusively for grinding. The upper molars have four (sometimes three) cusps, whereas the lowers have five (sometimes four). The number and position of the cusps and the character of the grooves between them differ among individuals. Generally, the upper molars have three roots and the lowers two. In persons of European origin, the first molar usually is large, the second smaller, and the third the smallest. In some other groups the third molar is larger than the second, though smaller than the first (see TEETH AND GUMS, HUMAN).

Functions of the tongue

Tongue. The tongue, a muscular organ located on the floor of the mouth, in man is an extremely mobile structure and an important accessory organ in such motor functions as speech, chewing, and swallowing. In conjunction with the cheeks it is able to guide and maintain food between the upper and lower teeth until mastication is completed. In persons whose tongues have been excised, speech is defective. The tongue's motility aids in creating a negative pressure within the oral cavity, thus enabling mammals to suckle.

The mucous membrane that covers the tongue varies greatly. Especially important as a peripheral sense organ, it contains groups of specialized epithelial cells, known as taste buds, that carry stimuli from the oral cavity to the central nervous system. Furthermore, the tongue's glands produce some of the saliva necessary for swallowing.

The general appearance of the mucous membrane has interested physicians for centuries, and the custom of having the patient stick out his tongue during an examination is of some use, since certain changes in its appearance reflect disturbances in other organs and systems.

The mammalian tongue consists of a mass of interwoven, striated (striped) muscles covered with mucous membrane and interspersed with some glands and a variable amount of fat. By its extrinsic muscles, the tongue is attached to the lower jaw, the hyoid bone (a U-shaped bone between the lower jaw and the larynx), the skull, the soft palate, and the pharynx. It is bound to the floor of the mouth and to the epiglottis (a plate of cartilage that serves as a lid for the larynx) by folds of its mucous membrane.

The tongue presents on its upper surface a median groove, which traced posteriorly ends in a small pit, the foramen cecum linguae, which is the point of origin of the thyroid gland and forms the apex of a V-shaped groove, the terminal sulcus. The terminal sulcus marks the division of the tongue into two parts. The larger part, called the body of the tongue, is in front and belongs to the floor of the mouth, while the smaller posterior part, or root, forms the forward wall of the oral pharynx. (The pharynx and its divisions are taken up in a later section.)

The lingual papillae

The upper surface of the body of the tongue, called the dorsum, is rough in appearance because of numerous small projections, the lingual papillae, of which several kinds are recognized. Filiform (thread-shaped) and conical papillae are arranged in V-shaped rows, being numerous over the whole of the dorsum. Each contains a core of fibrous tissue with minute blood vessels. Fungiform papillae, similar in structure but less numerous, are easily distinguished by their larger size and reddish colour. They are found immediately in front of, and parallel to, the terminal sulcus. On most of the fungiform papillae are taste buds, usually seven to 11 in number. Each fungiform papilla consists of a flat central mound surrounded by a depression resembling a moat. On the sides of the moat are taste buds, and into the bottom of the fossa open ducts of Ebner's glands (which secrete a watery fluid that aids in forming saliva). Foliate papillae, rudimentary in man, are represented by three to eight vertical folds at the side of the hinder part of the dorsum.

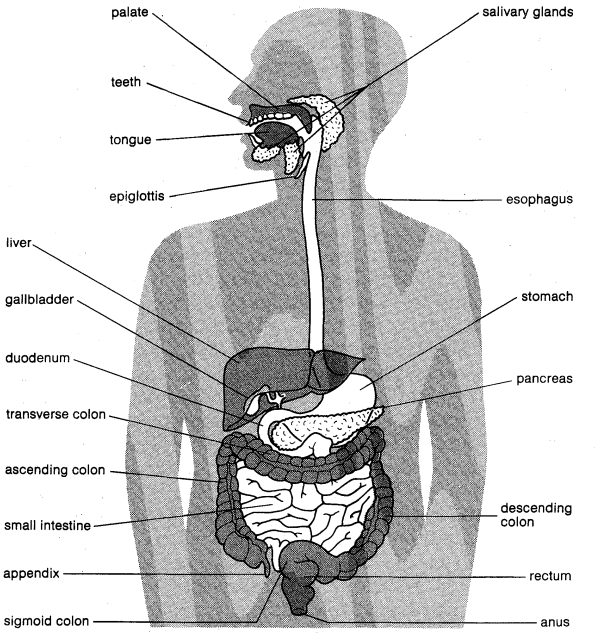
The undersurface of the body of the tongue is covered in its free portion by a thin, smooth, mucous membrane. A prominent median fold, the frenulum linguae, connects the tongue with the mandible, lower jaw, and the floor of the mouth; on each side of this structure is an irregular, fringed fold, the plica fimbriata. Between the frenulum and the plica fimbriata, on each side, the lingual vein side shines through the mucosa.

The root of the tongue, which is thought of as part of the oral pharynx, differs from the body in several respects, including structure and appearance. In appearance, the mucous membrane of the root is warty because of the underlying nodules of lymphoid tissue, the lingual follicles, collectively designated as the lingual tonsil. The mucosa is not firmly adherent to the underlying structures, and a loose median fold of mucous membrane passes from the root of the tongue to the epiglottis and separates the valleculae, or little valleys, in which foreign bodies often become lodged. The lateral boundaries of the valleculae are two folds of mucous membrane.

The substance of the tongue is composed chiefly of interlacing striated muscle fibres arranged symmetrically on each side of a median fibrous septum, or partition. They are innervated by the hypoglossal (under the tongue) nerve. Four muscles lie entirely within the

tongue on each side, the superior and inferior longitudinal, the vertical, and the transverse muscles. These muscles, known as the intrinsic muscles, on contracting, change the length, width, and breadth of the tongue and protrude it. Moreover, when contracted on one side only, they cause the tongue's tip to bend in the opposite direction. A special branch of the hypoglossal nerve, known as the end branch, innervates these muscles. Other muscles—the pairs of hypoglossus, chrondroglossus, styloglossus, and genioglossus muscles—come from skeletal parts and are attached to the substance of the tongue on either side. They are known as the extrinsic muscles. With the exception of the genioglossus muscle, the extrinsic muscles are retractors of the tongue. The genioglossus muscle can protrude the tongue a small amount, but its main function is to build a fulcrum around which the intrinsic muscles function. The arteries of the tongue are derived mainly from the lingual (tongue) branches of the external carotid arteries, which supply blood to the exterior of the head, to the face, and to much of the neck. The lingual veins return the blood from the tongue to the internal jugular veins, which receive blood from the brain, the face, and the neck.

The extrinsic muscles



Human digestive system.

SALIVARY GLANDS

Besides the many minute glands that secrete saliva, there are three major pairs of salivary glands: the parotid, the submaxillary, and the sublingual glands.

Size, form, and location. The parotid glands, the largest of the salivary glands, in the adult weigh from 20 to 30 grams (from about 0.7 to one ounce) each. They are located at the side of the face, below and in front of each ear, in back touching the mastoid bone (which is behind the ear) and the sternomastoid muscle, the prominent muscle of each side of the neck, and in front shaped around the ascending portion of the bone of the lower jaw. The parotid glands are enclosed in sheaths that limit the extent of their swelling when inflamed, as in mumps.

The submaxillary glands, each of which in the adult weighs from eight to ten grams (from about 0.3 to 0.4 ounce), lie near the inner side of the lower jawbone, not far in front of the sternomastoid muscle. They are rounded in shape, with a flattened forward side.

The sublingual glands, which are elongated in shape, weigh from two to three grams (from 0.07 to 0.11 ounce). They lie directly under the mucous membrane covering the floor of the mouth beneath the tongue.

Structure. *Gross structure.* The salivary glands are of the type called racemose, from the Latin *racemosus* ("full of clusters"), because of the clusterlike arrangement of their secreting cells in rounded sacs, called acini, attached to freely branching systems of ducts.

The ducts of each parotid gland empty into successively larger ducts until they reach one large duct, as do the ducts of each submaxillary gland. The main duct of the parotid gland is known as Stensen's duct; that of the submaxillary gland, as Wharton's duct. Stensen's duct opens into the mouth at a point in the cheek opposite the crown of the second upper molar; Wharton's duct opens on the summit of the soft elevation known as the caruncula sublingualis, beside the frenulum of the tongue.

The sublingual glands, unlike the other two types, are drained by a number of small, short ducts rather than one large one. The short ducts, called the ducts of Rivinus, open on papillae along the side of the tongue. In many cases the forward part of the sublingual gland is drained by a single, large duct that opens alongside the opening of the submaxillary duct; *i.e.*, Wharton's duct.

Microscopic structure. The walls of the acini surround a small central cavity known as an alveolus. In the walls are one or two types of somewhat pyramidal secreting cells and, besides the secreting cells, some flat, star-shaped cells called myoepithelial, or basket, cells. The latter cells are thought to contract, like the similar myoepithelial cells of the breast, which by their contraction expel milk from the milk ducts.

Types of secreting cells

The secreting cells may be of the serous or the mucous type. The latter type secretes mucin, the chief constituent of mucus; the former, a watery fluid containing an enzyme, amylase, which is also known as ptyalin. The secreting cells of the parotid glands are of the serous type; those of the submaxillary glands, of both serous and mucous types, with the serous cells outnumbering the mucous cells by four to one. The acini of the sublingual glands are composed primarily of mucous cells.

The ducts of the salivary glands are lined with several different types of cells. Cube-shaped cells that are largely nuclei line the smallest ducts, those leading from the acini. Larger ducts, called intralobular ducts, are lined with columnar, or rod-shaped, cells that resemble the cells lining the kidney tubules, the minute tubes of the nephrons, or functioning elements, of the kidneys. The next grade of duct, the interlobular, has a lining composed of two layers; the superficial layer is of columnar cells and the deeper layer of flattened cells. Near the termination of the interlobular ducts the lining changes to stratified squamous epithelium—a covering made up of layers of scalelike cells.

Innervation. The nerves controlling the salivary glands in man belong both to the parasympathetic and to the sympathetic divisions of the autonomic nervous system (the autonomic nervous system has to do with reactions that are independent of voluntary control). The parasympathetic nerve supply regulates secretion by the acinar cells and causes the blood vessels to dilate. Functions regulated by the sympathetic nerves include secretion by the acinar cells, constriction of blood vessels, and, presumably, contraction of the myoepithelial cells.

The parasympathetic nerves governing the salivary glands originate in the inferior and superior salivary nuclei, groups of nerve cells in the medulla oblongata, a centre in the brainstem just above the spinal cord.

The sympathetic nerve supply to the salivary glands originates in nerve cells near the top of the spinal cord (in the lateral horn of the first and second thoracic segments of the spinal cord).

Nerve fibres that carry pain impulses from the salivary glands to the brain form part of the chorda tympani (a branch of the facial, or seventh cranial, nerve and of the glossopharyngeal, or ninth cranial, nerve).

PHARYNX

Functions of the pharynx

The pharynx, or throat, is the passageway leading from the mouth and nose to the esophagus and larynx. The pharynx permits the passage of swallowed solids and liquids into the esophagus, or gullet, and conducts air to and from the trachea, or windpipe, during respiration. The pharynx also connects on either side with the cavity of the middle ear by way of the eustachian tube and provides for equalization of air pressure on the eardrum

membrane, which separates the cavity of the middle ear from the external ear canal.

The pharynx has roughly the form of a funnel flattened from front to back. It is about 13 centimetres (five inches) in length and narrows from a width of about five centimetres (two inches) at the top to about one inch at its junction with the esophagus. The walls of the pharynx are attached by its connective tissue and muscles to the surrounding structures. The attachment in most areas is loose enough in organization to permit gliding of the pharyngeal wall, in the movements of swallowing, against the surrounding structures.

Three main divisions of the pharynx are distinguished: the oral pharynx, the nasal pharynx, and the laryngeal pharynx. The latter two are airways, whereas the oral pharynx is shared by both the respiratory and alimentary (digestive) tracts. On either side of the opening between the mouth cavity and the oral pharynx is a tonsil called a palatine tonsil because of its proximity to the palate. Each palatine tonsil is between two vertical folds of mucous membrane (behind the glossopalatine arch and in front of the pharyngopalatine arch). The glossopalatine arches, located on the sides of the pharynx and an imaginary line back of the front two-thirds of the tongue, are approximately where the mouth cavity and oral pharynx meet. The nasal pharynx, above, is separated from the oral pharynx by the soft palate. The laryngeal pharynx and the lower part of the oral pharynx are hidden by the bulging root of the tongue. An important feature of this obscured region is the epiglottis, or laryngeal flap, which acts as a deflector between the laryngeal pharynx and the lowermost oral pharynx.

The principal layers of the walls of the throat, from within outward, are (1) a mucous membrane, consisting of epithelium (covering), numerous mucous glands, and fibrous connective tissue, of which the outermost part is a tough membrane called the pharyngeal aponeurosis; (2) the pharyngeal muscles; and (3) connective tissue. The pharyngeal muscles are concerned in the mechanics of swallowing. The principal muscles of the pharynx are the three pharyngeal constrictors. These pharyngeal constrictor muscles overlap each other slightly, from below upward, and form the primary musculature of the side and rear pharyngeal walls.

The tonsils

The pharynx is the seat of the tonsils. The term tonsil may suggest only the palatine tonsils, mentioned above, but these are only components of a series of tonsils encircling the wall of the pharynx. On either side, in addition to the palatine tonsil, is the tubal tonsil, near the opening of the eustachian tube. On the roof of the nasopharynx is the pharyngeal tonsil, which is frequently so enlarged in children as to obstruct the airway. (The tonsils are then known as adenoids.) On the floor of the mouth cavity, in the back one-third of the tongue, is the lingual tonsil. The tonsils are masses of lymphoid tissue embedded in the mucous membrane; this tissue is composed principally of lymphocytes, a variety of white blood cell. The tonsils are usually smaller in adults.

ESOPHAGUS

The esophagus, that portion of the alimentary canal that extends from the pharynx to the stomach, is about 25 centimetres (ten inches) in length; the width varies from one and one-half to two centimetres. The esophagus contains the four typical layers of the alimentary canal—mucosa (or mucous membrane), submucosa, muscularis, and tunica adventitia. The mucosa is made up of stratified squamous epithelium containing numerous mucous glands. The submucosa is a thick, loose fibrous layer connecting the mucosa to the muscularis. The mucosa and submucosa form long longitudinal folds so that a cross section of the esophagus opening would be star-shaped. The muscularis is composed of an inner layer in which the fibres are circular and an outer layer of longitudinal fibres. The outer layer of the esophagus, the tunica adventitia, is composed of loose fibrous tissue that connects the esophagus with neighbouring structures. Except during the act of swallowing, the esophagus is normally empty, and its lumen, or channel, is essentially blocked by the longitudinal folds of the mucosal and the submucosal layers.

Course of the esophagus

The esophagus can be divided into cervical (neck), thoracic (chest), and abdominal portions. The cervical portion lies immediately in front of the vertebral column and behind the trachea. The thoracic portion of the esophagus is that part within the mediastinum, the space between the two lungs. Within the thoracic cavity the esophagus passes behind the fork of the trachea, where it divides with the right and left main bronchi, and the right pulmonary artery (the artery that carries blood low in oxygen from behind the heart to the right lung). In the lower portion of the mediastinum, the esophagus passes between the aorta and the pericardium, the membranous sac that encloses the heart. The esophagus passes out of the thoracic cavity through the diaphragmatic hiatus, the opening in the muscular partition between the chest and the abdomen. The abdominal portion of the esophagus is that short portion (two to three centimetres) that extends from the diaphragm to the stomach.

Another way of dividing the esophagus is into upper, middle, and lower thirds. The upper third of the esophagus is composed of striated muscle, the middle third of a mixture of striated and smooth (involuntary) muscle, the lower third only of smooth muscle.

The esophagus has two sphincters. (Sphincters are circular muscles that act like drawstrings in closing channels.) The upper esophageal sphincter is located at the level of the cricoid cartilage (a single ringlike cartilage forming the lower part of the larynx wall). This sphincter is called the cricopharyngeus muscle. The lower esophageal sphincter encircles the three to four centimetres of the esophagus that pass through the diaphragmatic hiatus and is thus located partly above and partly below the diaphragm. Both sphincters normally remain closed except during the act of swallowing.

The arterial blood supply to the esophagus is derived mainly from the esophageal branches of the aorta. The veins that drain the esophagus accompany the arteries. The esophagus is supplied by both sympathetic and parasympathetic nerves, which terminate in two networks called plexuses. Of these, Meissner's plexus is located in the submucosa, whereas Auerbach's plexus is located between the circular and longitudinal muscle coats.

STOMACH

The stomach serves as a reservoir and receives ingested food and liquids from the esophagus and retains them for admixing with the gastric juice in order that digestion can begin. It is located in the left upper part of the abdomen immediately below the diaphragm. In front of the stomach are the liver, part of the diaphragm, and the forward abdominal wall. Behind it are the pancreas, the left kidney, the left adrenal, the spleen, and the colon. When the stomach is empty, it contracts, and the transverse colon ascends to occupy the vacated space.

The size, shape, and position of the stomach vary extremely and depend upon the extent of its contents as well as upon the tension in the muscles of its walls. The average capacity of the stomach is approximately one quart (one litre). The organ is more or less concave on its right side, convex on its left. The concave border is called the lesser curvature; the convex border, on the left and below, the greater curvature. The opening from the esophagus into the stomach is the cardia, while the outlet from the stomach into the duodenum is the pylorus.

The parts of the stomach

The various parts of the stomach may be summarized as follows. The uppermost part, located above the entrance of the esophagus, is the fundus; it frequently contains a gas bubble, especially after a meal. The cardia is that portion of the stomach surrounding the opening from the esophagus. The largest part of the stomach is the body; it serves primarily as a reservoir for ingested food and liquids. The antrum, the lowermost part of the stomach, is somewhat funnel-shaped, with its wide end joining the lower part of the body of the stomach and its small end connecting with the pyloric canal, which empties into the duodenum. The pyloric portion (antrum plus pyloric canal) of the stomach tends to curve to the

right and slightly upward and backward and thus gives the stomach its J-shaped appearance. The pylorus, the narrowest portion of the stomach, is approximately two centimetres (0.8 inch) in diameter.

The stomach is lined with mucosa, or mucous membrane. When the stomach is empty, the lining is thrown into numerous longitudinal folds, known as rugae; these tend to disappear when the stomach is distended.

The surface of the gastric (stomach) mucosa is always covered by a thin layer of thick tenacious mucus that is secreted by the columnar cells in the surface epithelium. Beneath the surface epithelium various types of gastric glands are located that, in different parts of the stomach, vary in structure and in composition of their secretion. Thus, each area of the gastric mucosa is characterized by its glandular structure. When the gastric mucus is removed from the surface epithelium, small pits, called foveolae gastricae, may be observed with a hand magnifying glass. There are approximately 90 to 100 gastric pits per square millimetre (58,000 to 65,000 per square inch) of surface epithelium. Into each gastric pit from three to seven individual gastric glands empty their secretions.

The gastric mucosa contains five different types of cells. In addition to the tall columnar surface epithelial cells previously mentioned, there are three common cell types found in the various gastric glands and a rare cell type occasionally observed: (1) Mucoid cells—cells that secrete mucus—are common to all types of gastric glands. Mucoid cells are usually the only cell type found in the gastric glands in the cardiac and pyloric areas of the stomach. The necks of the glands in the body and fundic parts of the stomach are lined with mucoid cells. (2) Other cells, called zymogenic, or chief, cells, are located predominantly in the gastric glands in the body and fundic portions of the stomach. These cells secrete pepsinogen, from which the enzyme pepsin is formed. (3) Cells called parietal, or oxyntic, cells, in the glands of the body and fundic portions of the stomach, secrete hydrochloric acid and also most of the water found in gastric juice. (4) Another type of cell occasionally found in these glands is known as the argentaffin cell. Its exact function is not known.

Beneath the gastric mucosa is a thin layer of smooth muscle fibres called the muscularis mucosa, and below this, in turn, is loose connective tissue, the submucosa, which attaches the gastric mucosa to the muscles in the walls of the stomach.

The muscles of the stomach wall

The muscles of the gastric wall are arranged in three layers, or coats. The innermost layer of smooth muscle of the gastric wall, called the oblique muscular layer, is strongest in the region of the gastric fundus and progressively weaker as it approaches the pylorus.

The middle or circular muscular layer, the strongest of the three muscular layers, completely covers the stomach. The circular fibres of this coat are best developed in the lower portion of the stomach, particularly over the antrum and the pylorus.

The external coat, called the longitudinal muscle layer, is formed from the longitudinal muscle coat of the esophagus. The longitudinal muscle fibres are divided at the cardia into two broad strips. The one on the right, the strongest, spreads out to cover the lesser curvature and the adjacent posterior and anterior walls of the stomach. The longitudinal fibres on the left sweep from the esophagus over the dome of the fundus of the stomach to cover the greater curvature and continue on to the pylorus, where they join the longitudinal muscle fibres coming down over the lesser curvature.

At the pyloric end of the stomach the middle layer of muscle, the circular layer, becomes greatly thickened to form the pyloric sphincter. This muscular ring is slightly separated from the circular muscle of the duodenum by connective tissue. The outer muscle layer of the stomach, the longitudinal layer, continues on into the duodenum, forming the longitudinal muscle of the small bowel.

Arterial and venous supply

Many branches of the celiac trunk bring arterial blood to the stomach. The celiac trunk is a short, wide artery that branches from the abdominal portion of the aorta, the main vessel conveying arterial blood from the heart to the systemic circulation. Blood from the stomach is returned to the venous system into the portal vein, which carries the blood to the liver.

The nerve supply to the stomach is provided by both the parasympathetic and sympathetic divisions of the autonomic nervous system. The parasympathetic nerve fibres are carried in the vagus, or tenth cranial, nerves. As the vagus nerves pass through the opening in the diaphragm with the esophagus, branches of the right vagus nerve spread over the posterior part of the stomach, while the left vagus supplies the anterior part. Sympathetic branches from a nerve network called the celiac, or solar, plexus accompany the arteries of the stomach into the muscular wall.

SMALL INTESTINE

The small intestine, which is 670–760 centimetres (22 to 25 feet) in length, is the longest part of the digestive tract of man. It begins at the pylorus, the juncture with the stomach, and ends at the ileocecal valve, the juncture with the colon. The parts of the small intestine are the duodenum, jejunum, and ileum.

The three parts of the small intestine. The duodenum is 23–28 centimetres (nine to 11 inches) long and forms a horseshoe, or C-shaped, curve that encircles the head of the pancreas. Unlike the rest of the intestine, it is retroperitoneal (that is, it is behind the peritoneum, the membrane lining the abdominal wall). Its first part, known as the duodenal bulb, is the widest part of the small intestine. It is horizontal, passing backward and to the right from the pylorus, and lies somewhat behind the wide end of the gallbladder. The second part runs vertically downward in front of the hilum of the right kidney (the point of entrance or exit for blood vessels, nerves, and the ureters), and into this part of the duodenum, at the ampulla of Vater, pancreatic juices and bile flow. The third part runs horizontally to the left in front of the aorta and the inferior vena cava (the principal channel for return to the heart of venous blood from the lower part of the body and the legs), while the fourth part ascends to the left side of the second lumbar vertebra (at the level of the small of the back), then bends sharply downward and forward to join the second part of the small intestine, the jejunum. An acute angle, called the duodenojejunal flexure, is formed by the suspension of this part of the small intestine by the ligament of Treitz.

The jejunum and ileum

The jejunum forms the upper two-fifths of the rest of the small intestine; it, like the ileum, has numerous convolutions and is attached to the posterior abdominal wall by mesentery, a fold of serous—clear-fluid-secreting—membrane.

The ileum is the remaining three-fifths of the small intestine, though there is no absolute point at which the jejunum ends and the ileum begins. In broad terms, the jejunum occupies the upper and left part of the abdomen below the subcostal plane (*i.e.*, from a plane just above the floating ribs), while the ileum is located in the lower and right part. At its termination the ileum opens into the large intestine.

The small bowel mucosa. The mucosa lining the small intestine is a remarkable structure in that it provides a tremendous absorptive surface to the intestinal contents. Although the small intestine is only three to four centimetres (1.2–1.6 inches) in diameter and approximately 7.32 metres (24 feet) in length, it has been estimated that the total absorptive area of the human small intestine is approximately 4,500 square metres (5,400 square yards). This enormous absorptive surface is provided by the unique structure of the mucosa. Within the small intestine, the mucosa is arranged in concentric folds that have the appearance of transverse ridges. These folds, or ridges, are approximately five to six centimetres (two to 2.4 inches) in length and about 3.2 millimetres (about one-eighth inch) thick. They are known as plicae circulares. These mucosal folds are present throughout the small bowel except in the first portion, or bulb, of the duodenum. The mucosa of the duodenal bulb is characteristically flat and smooth, except for a few longitudinal folds. The plicae circulares are largest in the lower

part of the duodenum and in the upper part of the jejunum. They become smaller and finally disappear in the lower part of the ileum. It has been estimated that the small intestine of man contains approximately 800 plicae circulares and that they increase the surface area of the lining of the small bowel by five to eight times the outer surface of the small intestine.

Another feature of the small bowel mucosa that greatly multiplies its surface area is that of the tiny projections called mucosal villi. These mucosal villi usually vary from 0.5 to one millimetre (0.02–0.04 inch) in height. Their diameters vary from approximately one-eighth to one-third their height. The villi are covered by a single layer of tall columnar—rodlike—cells. Goblet cells, so-called because of their rough resemblances to empty goblets after they have discharged their contents, are also found scattered among the surface cells. The goblet cells are a source of mucin, the chief constituent of mucus.

At the base of the mucosal villi are depressions called intestinal glands, or the crypts of Lieberkühn. The cells that line the crypts continue up and over the surface of the villi. In the bottom of the crypts, the epithelial cells are filled with alpha granules, or eosinophilic granules, so-called because they take up the rose-coloured stain eosin. These cells have been termed the cells of Paneth.

In the duodenum the villi are dense and large and frequently are leaflike in shape. In the jejunum the individual villus measures between 350 and 600 microns in height (there are about 25,000 microns in an inch) and has a diameter of 110 to 135 microns. The appearance and shape of the villi vary in different levels of the small intestine. The inner structure of the individual villus consists of loose connective tissue containing a rich network of blood vessels, a central lacteal, or channel for lymph, smooth muscle fibres, and scattered cells of various types. The smooth muscle cells surround the central lacteal and provide for the pumping action of the villi.

A small central arteriole (minute artery) branches at the tip of the villus to form a capillary network (the capillaries are the smallest of the blood vessels); the capillaries, in turn, empty into a collecting venule, or minute vein, that extends to the bottom of the villus.

The microvilli

The most remarkable feature of the small bowel mucosa is the rough surface of the epithelial cells, the cells covering the villi. This surface, called a brush border, is now known to consist of individual microvilli. The microvilli are approximately one-tenth micron in diameter and one micron in height; each epithelial cell may have as many as 1,000 microvilli. The microvilli, now known to play an important role in absorption of intestinal contents, enlarge the absorbing surface approximately 25 times and contain a number of enzymes.

Beneath the mucosa of the small intestine, as beneath that of the stomach, are the muscularis mucosae and the submucosa. The submucosa consists of loose connective tissue and contains many blood vessels and lymphatics. In the submucosa of the duodenum are located the glands of Brunner, composed of acini (round sacs) and tubules that are twisting and have multiple branching. Brunner's glands empty into the base of the crypts of Lieberkühn in the duodenum. Their exact function is not known, but they do secrete a scanty clear fluid that contains mucus and a relatively weak proteolytic (protein-splitting) enzyme. In the submucosa of the jejunum, solitary nodules (lumps) of lymphatic tissue are located. There is more lymphatic tissue in the ileum, in aggregates of nodules known as Peyer's patches.

The arrangement of the muscular coats of the small intestine is uniform throughout the length of the intestine. There is an inner, circular layer that is thicker than the outer, longitudinal layer. The outermost covering of the small intestine is the peritoneum, a single layer of flattened epithelial cells.

The blood supply of the small intestine is from the superior mesenteric artery (a branch of the abdominal aorta) and from the superior pancreaticoduodenal artery (a branch of the hepatic artery). These vessels run between layers of the mesentery and give off large branches that form a row or series of connecting arches from which branches enter the wall of the small bowel. The blood from the intestine is returned by means of the superior mesenteric vein, which, with the splenic vein, forms the portal vein, which drains into the liver.

The small intestine has both sympathetic and parasympathetic innervation. The vagus nerves provide the parasympathetic innervation. The sympathetic is provided by branches from the superior mesenteric plexus, a nerve network close under the solar plexus, which follow the blood vessels into the small intestine and finally terminate in the two plexuses, or networks—Auerbach's, located between the circular and longitudinal muscle coats, and Meissner's, which is located in the submucosa.

LARGE INTESTINE

The large intestine serves as a reservoir for the liquids emptied into it, through the ileocecal valve, from the small intestine. The large intestine, or colon, may be divided into the cecum, ascending colon, transverse colon, descending colon, and sigmoid colon. The primary function of the colon is to absorb water and electrolytes (substances, such as salts, that in solution take on an electrical charge) from the ileal contents and to store fecal material until it can be evacuated by defecation.

Parts and function of colon

The cecum, the first part of the large intestine, is a sac with a closed end. It occupies the right iliac fossa, the hollow on the inner side of the ilium (the upper part of the hipbone). Guarding the opening of the ileum into the cecum is the ileocecal valve. Most textbooks of anatomy describe the ileocecal valve as consisting of two folds or flaps of mucous membrane on the cecal side and above and below the ileal opening. The ileocecal junction does have this appearance commonly after death. During life, however, the ileocecal junction appears much different in that the terminal portion of the ileum doubles into the cecum; from the cecal side the ileocecal valve greatly resembles the cervix, the projection of the uterus into the vagina. The circular muscle fibres of the ileum and those of the cecum combine to form the circular sphincter muscle of the ileocecal valve.

The ascending colon extends up from the cecum at the level of the ileocecal valve to the bend in the colon called the hepatic flexure, which is located beneath and behind the right lobe of the liver. The ascending colon is about 20 centimetres (eight inches) long. In back it is in contact with the rear abdominal wall and the right kidney. The ascending colon is covered by peritoneum except on its posterior surface.

The transverse colon is variable in position, depending largely on the distention of the stomach, but usually is located in the subcostal plane; that is, at the level of the tenth rib. On the left side of the abdomen it ascends to the bend called the splenic flexure, which may make an indentation in the spleen. The transverse colon is bound to the diaphragm opposite the eleventh rib by a fold of peritoneum.

The descending colon passes down and in front of the left kidney and the left side of the posterior abdominal wall to the iliac crest, the upper border of the hipbone. The descending colon is about 15–20 centimetres (six to eight inches) in length, and it is more likely than the ascending colon to be surrounded by peritoneum.

The sigmoid colon is commonly divided into iliac and pelvic parts. The iliac colon stretches from the crest of the ilium, or upper border of the hipbone, to the inner border of the psoas muscle, which lies in the left iliac fossa. Like the descending colon, the iliac colon is usually covered by peritoneum. It is about 15 centimetres (six inches) in length. The pelvic colon lies in the true pelvis (lower part of the pelvis) and forms one or two loops, reaching across to the right side of the pelvis and then bending back and, at the midline, turning sharply downward to the point where it becomes the rectum.

The layers that make up the wall of the colon are similar in some respects to those of the small bowel; there are distinct differences, however. The external aspect of the colon differs markedly from that of the small bowel because of features known as the haustra, taeniae, and appendices epiploicae.

Haustra and taeniae

The haustra, bulges or sacculations, are formed by constricting circular furrows of varying depths. The three taeniae are long, narrow bands of longitudinal muscle fibres, about one centimetre (0.4 inch) in width, that are approximately equally spaced around the circumference of the colon. Between the thick bands of the taeniae there is a thin coating of longitudinal muscle fibres.

The appendices epiploicae are collections of fatty tissue beneath the covering membrane. On the ascending and descending colon they are usually found in two rows, whereas on the transverse colon they form only one row.

The mucous membrane of the colon has a characteristic structure. It lacks the villi and the folds known as plicae circulares characteristic of the small intestine. It contains many solitary lymphatic nodules but no Peyer's patches. The surface epithelium is columnar, and there are many goblet cells. Characteristic of the colonic mucosa are deep tubular pits, increasing in depth toward the rectum. The submucosa contains numerous solitary lymphatic nodules, blood vessels, lymphatics, and submucosal nerve networks known as Meissner's plexuses.

The arterial blood supply to the large intestine is supplied by branches of the superior and inferior mesenteric arteries (both of which are branches of the abdominal aorta) and the hypogastric branch of the internal iliac (which supplies blood to the pelvic walls and viscera, the genital organs, the buttocks, and the inside of the thighs). The vessels form a continuous row of arches from which vessels arise to enter the large intestine. Venous blood is drained from the colon from branches that form venous arches similar to those of the arteries. These eventually drain into the superior and inferior mesenteric veins, which ultimately join with the splenic vein to form the portal vein (which carries venous blood to the liver).

The innervation of the large intestine is similar to that of the small intestine.

RECTUM AND ANUS

The rectum, which is a continuation of the sigmoid colon, begins in front of the midsacrum (the sacrum is the triangular bone near the base of the spine and between the two hipbones). It ends in a dilated portion called the rectal ampulla, which in front is in contact with the rear surface of the prostate in the male and with the posterior vaginal wall in the female. Posteriorly, the rectal ampulla is in front of the tip of the coccyx (the small bone at the very base of the spine).

At the end of the pelvic colon, the mesocolon, the fold of peritoneum that attaches the colon to the rear wall of the abdomen and pelvis, ceases, and the rectum is then covered by peritoneum only at its sides and in front; lower down, the rectum gradually loses the covering on its sides, until only the front is covered. At about the junction of the middle and lower thirds of the rectum, about 7.5 centimetres (three inches) from the anus, the anterior peritoneal covering is also folded back onto the bladder and the prostate or the vagina.

Near the termination of the sigmoid colon and the beginning of the rectum, the colonic taeniae spread out to form a wide external longitudinal muscle coat. At the lower end of the rectum muscle fibres of the longitudinal and circular coats tend to intermix. The internal circular muscle coat terminates in the thick rounded internal anal sphincter muscle. The smooth muscle fibres of the external longitudinal muscle coat of the rectum terminate by interweaving with striated muscle fibres of the levator ani, a broad muscle that forms the floor of the pelvis. A second sphincter, the external anal sphincter, is composed of striated muscle and is divided into three parts known as the subcutaneous, superficial, and deep external sphincters. Thus, the internal sphincter is composed of smooth muscle and is innervated by the autonomic nervous system, while the external sphincters are of striated muscle and have somatic (voluntary) innervation provided by nerves called the pudendal nerves.

The mucosal lining of the rectum is similar to that of the sigmoid colon but becomes thicker and better supplied with blood vessels, particularly in the lower rectum. In the rectal ampulla are two to three large crescentlike folds known as rectal valves. These folds, or valves, are caused by an invagination, or infolding, of the circular muscle and submucosa. The columnar epithelium of the rectal mucosa changes to the stratified squamous (scale-like) type in the lower rectum a few centimetres above the pectinate line, which is the junction between squamous mucous membrane of the lower rectum and the skin lining the lower portion of the anal canal.

The rectal valves

Arterial blood is supplied by branches from the inferior mesenteric artery and the right and left internal iliac arteries. Venous drainage from the anal canal and rectum is provided by a rich network of veins called the internal and external hemorrhoidal veins.

ASSOCIATED GLANDS AND STRUCTURES

Pancreas. The pancreas is both an exocrine (ducted) and endocrine (ductless) gland. The exocrine functions consist of providing the digestive system with water, bicarbonate, and enzymes. The endocrine functions consist primarily of providing the bloodstream with two hormones—insulin and glucagon—necessary for carbohydrate regulation and metabolism.

In man, the pancreas is a large gland located behind the stomach. The largest portion of the gland, known as the head, is within the curve of the duodenum and firmly attached to it. An extension of the head of the gland downward and to the left, a portion called the uncinate process (*i.e.*, hooked projection) lies immediately above the transverse portion of the duodenum. Extending to the left from the head of the gland there is a short neck and then the body and tail of the pancreas. The tail of the pancreas extends to the spleen.

The pancreas is completely retroperitoneal (behind the peritoneum). In front, it is covered almost in its entirety by the stomach. Posteriorly, the head is in contact with the inferior vena cava and the aorta. The body of the pancreas covers the splenic artery and vein, while the tail of the pancreas covers the upper portion of the left kidney. The pancreas is drained by a large main duct, the duct of Wirsung, which runs the entire length of the gland and, in the wall of the second part of the duodenum, joins usually with the common bile duct at a sac called the ampulla of Vater. This empties its mixture of bile and pancreatic juices into the duodenum. A smaller accessory pancreatic duct, the duct of Santorini, branches from the main pancreatic duct at a level near the junction of the head and neck of the gland and empties directly into the second portion of the duodenum at a point near the ampulla of Vater.

Like the salivary glands, the pancreas is a racemose (cluster-filled) gland. When studied microscopically, the pancreas is seen to be made up chiefly of cell groups, called acini, that are hollow and tend to be spherical or oval in shape. Groups of acini form primary lobules. These in turn combine to form larger secondary lobules. The secondary lobules are completely surrounded by connective tissue and can be dissected out as separate structures connected to the rest of the gland by ducts, nerve fibres, lymphatic vessels, and blood vessels.

Acini and lobules

The cells in each acinus tend to be pyramidal in shape, with the apex of the pyramid directed toward the hollow (lumen) in the centre of the sphere. The cells are large, with a well-developed nucleus, and have an abundant granular cytoplasm. The granules in the cytoplasm, called zymogen granules, tend to be more abundant in the part of the cell near the lumen of the acinus.

Besides the clusters of acini joined by ducts, there are, scattered through the pancreas, many thousands of microscopic groups of cells, known as the islets of Langerhans, that serve the endocrine functions of the pancreas. In the islands there are three types of cells: alpha cells, the sources of glucagon; beta cells, the sources of insulin; and D-cells, of unknown function.

Sympathetic nerve fibres accompany the blood vessels to all parts of the pancreas. Parasympathetic innervation is provided from both the vagus nerves.

Liver, gallbladder, and bile ducts. *Liver.* All invertebrate animals possess a liver, and the livers of all adult vertebrates are essentially alike, though among verte-

brate animals some 20 modes of liver development have been found. The structure, functions, and disorders of the human liver, gallbladder, and bile ducts are covered at length in the article LIVER, HUMAN.

In man, the liver is the largest gland in the body; it is located in the upper right part of the abdominal cavity beneath the right diaphragm. Its weight is approximately one and one-half kilograms (3.3 pounds), representing about one-fortieth of the total body weight. In the living state, the liver is soft and compressible. Surrounding organs in contact with the liver push in its surface and leave impressions. When the liver is removed from the body, however, these depressions disappear.

Relationship of the liver to adjacent structures

The anterior, or forward, surface of the liver is divided into a large right lobe and a smaller left lobe by a thick ligament, the falciform ligament, which attaches the anterior surface of the liver to the abdominal wall. The upper and posterior surfaces of the liver are attached to the diaphragm and to the rear abdominal wall by a ligament called the coronary ligament. The inferior surface of the liver has many depressions from adjacent organs, the right kidney, the hepatic flexure of the colon, and the duodenum. The inferior surface of the smaller left lobe of the liver usually shows a depression produced by the fundus of the stomach.

The liver has a wide variety of functions, including the destruction of worn-out red cells, the production of the proteins that make up most of the clotting factors in the blood, the storage of carbohydrates and some fats and proteins, the production of urea, and the detoxification or modification of some drugs and poisons.

Only one activity of the liver seems directly concerned in the functioning of the digestive system. This is the production of bile salts, which aid in the digestion and absorption of fats in the small intestine. The bile salts are produced by the cells called variously hepatocytes, liver cells, and hepatic cells. They are arranged in close conjunction with microscopic blood and bile channels, in plates forming complex honeycomb-like structures.

Bile ducts and gallbladder. The gallbladder, and the system of ducts that carry bile from the liver to the gallbladder for concentration and storage and from the liver and the gallbladder to the duodenum, are described in LIVER, HUMAN.

Ducts from the right and left lobes of the liver unite to form the common hepatic duct. From the side of this a duct, called the cystic duct, leads to the gallbladder.

The common bile duct, formed by the union of the hepatic and cystic ducts, is like the vertical portion of the letter Y. It passes down through the head of the pancreas and then penetrates obliquely the muscular wall of the duodenum. Within the wall of the duodenum the common bile duct and the main pancreatic duct join and empty into the duodenum through the ampulla of Vater. The ampulla of Vater is surrounded near its outlet into the duodenum by a ring of muscle fibres, the sphincter of Oddi. (This muscle closes the passage through the ampulla of Vater and causes bile to back up into the gallbladder except when there is fat in the duodenum.)

The human gallbladder has a capacity of approximately 50 millilitres (three cubic inches). The wall of the gallbladder is composed of a thin layer of smooth muscle fibres and fibroelastic tissue. The gallbladder mucosa is composed of three types of cells—tall columnar cells, the most common, and two other types called pencil cells and basal cells.

The columnar epithelium of the gallbladder has many characteristics of an active absorbing membrane, including a border of microvilli on the surfaces of the cells next to the cavity of the organ, like the microvilli on the cells lining the small intestine. The lateral walls of the adjacent cells are interlocked into each other by folds of plasma membrane (the membrane enclosing the cell). This interlocking is thought to prevent material in the gallbladder cavity from penetrating between the cells, leaving the main route of absorption through the cells themselves.

BIBLIOGRAPHY. Further material on the human digestive system may be found in the following works: L.B. AREY, *Developmental Anatomy*, 7th ed. (1965); CHARLES MAYO GOSS (ed.), *Gray's Anatomy*, 28th ed. (1966); R.J. LAST, *Anatomy, Regional and Applied*, 4th ed. (1966); F.H. NETTER, *The Ciba Collection of Medical Illustrations*, vol. 3, pt. 1–3, ed. by E. OPPENHEIMER (1966).

(N.C.H.)

Digestive System Diseases

The ingestion of food and fluid is a prerequisite for man's growth; his digestive system represents one of his most ancient and essential vital pathways for survival. The system has evolved in such a way that great discriminatory powers are concentrated at the mouth, governing the amount and type of material ingested, and at the lower end providing for orderly expulsion of what is no longer needed. In between these two highly sensitive areas the tract has feebly developed sensors to the brain, but highly developed neurohumoral reflexes—reflexes dependent upon chemical substances released at nerve endings—which provide for the proper mixing of foodstuffs with those enzymes, hormones, emulsifiers, and electrolytes whose interaction constitutes what is called digestion and absorption. This article considers the derangements that may occur in the human digestive tract as a result of congenital, infectious, environmental, cancerous, or inflammatory processes.

MOUTH AND PHARYNX

This uppermost section of the digestive tract extends from the lips to the point at which the pharynx, or throat, meets the esophagus, a muscular tube extending to the stomach. It is designed to smell, taste, masticate, and swallow materials, processes in which the individual gains experience as he ages. It may be damaged because it is poorly provided with sensors—as in the blind, or in those unable to smell or taste properly—or because (from inexperience) it allows matter too hot or too cold to come into contact with its mucous membranes, or because it is deceived by poisons masquerading as nutrients—*e.g.*, poisonous plants or meats carrying botulinus toxin.

The most important of the inborn defects involve failures in fusion of the palate—the bones and soft tissues of the roof of the mouth—thus impairing the ability to produce a closed, high-pressure cavity behind the lips and teeth. Other defects relate to defective apposition of the teeth and the jaws, resulting in inefficient mastication, and to the absence of one or more of the salivary glands, which may lessen the amount and quality of saliva that they produce. Neurological defects that provide inadequate motor power to the muscles of the tongue and the pharynx can seriously impair powers of mastication and even of swallowing; sensory-innervation defects may not allow the usual reflexes to mesh smoothly or may permit harmful ingestants to pass by undetected.

The mouth is in constant contact with the external environment, and through kissing and the use of shared feeding utensils may provide an easy source of entry for infectious agents into the abundant lymphoid tissues that form the tonsils, or through the esophagus and stomach into the small intestine from which they are disseminated. These agents can produce characteristic signs in the mucous membranes of the mouth and throat such as the Koplik spots that are the first signs of measles, the inflammation of the inside of the cheek at the point where Stensen's duct opens into the mouth in mumps, or the various ulcerations of the palate seen in chickenpox or smallpox.

Infectious processes

In addition to the mouth lesions, some infections produce their major or most dramatic manifestations in the mouth and throat. Among the serious disorders of this type are moniliasis and histoplasmosis, fungous infections that produce ulcerative breakdown of the mucous membranes and often intense enlargement of the lymph glands of the tonsillar and tongue areas. Tuberculosis and syphilis produce a variety of reactions in this region, depending on the stages of the infection. These vary from white mucous patches on the mucous membrane of the cheek to deep ulceration of tongue, lip, or palate; or to

accumulations of nodules, or small lumps. Infection with the common pyogenic bacteria produces acute redness and swelling of the posterior pharynx, from which the surface exudate is found to be loaded with streptococcus or staphylococcus organisms. Swollen, bleeding, malodorous gums are often the site of invasion by the organisms that cause "trench mouth" in mouths rendered insufficiently resistant by general ill health, malnutrition, or poor hygiene. Among the commonest diseases of this area is the multiple vesicular and ulcerative disease, aphthous stomatitis (canker sores), affecting the mucous membrane of the mouth and often the adjacent skin of the lips. Many skin diseases, both acute and chronic, have associated eruptive phases in the mucous membranes of mouth and pharynx.

As the mouth normally contains a variety of bacteria living in peaceful coexistence, it is apparent that adequate resistance to them can be maintained by properly nourished mucous membranes. Whenever the nutritional status of this barrier mucosa is lessened by whatever cause, disease becomes apparent as inflammation and easy bleeding of gums and mucous membranes, swelling and loss of papillae of the tongue, changes in colour of lips and tongue, and breakdown of the thin skin of the lips. These and other changes may be brought out more sharply whenever trauma—as in toothbrushing or denture irritation—is exerted or when infections damage the surface or increase metabolic needs. The best documented evidences of nutritional deficiency in the mouth and throat are: the easy bleeding of gums often seen in ascorbic-acid deficiency, the smooth, red tongue of pernicious anemia, the bright red tongue noted in niacin deficiency, and the inflammation at the corners of the mouth in riboflavine and perhaps other B-vitamin deficiencies. In Addison's disease the mucous membrane of the cheeks becomes infiltrated with a brownish pigmentation.

Certain toxins found in foods can produce devastating effects on the function of the oropharynx without leaving visible lesions. Botulinus toxin causes such damage to the nervous system centres governing the 9th, 10th, and 12th cranial nerves that swallowing is usually badly deranged, and coordination of respiration with swallowing may be impossible.

Ingestion of lead characteristically results in a dark line along the margins of gums rendered susceptible by poor hygiene; mercury inhalation or ingestion leads to an ulcerative inflammation of the gums and mouth, often of great severity. Such disorders are now rare, and instead it is more reasonable to attribute gum hypertrophy and inflammation to certain chronically ingested therapeutic agents, in particular anticonvulsants and antibiotics. Again, these lesions are most frequent and most severe in mouths subjected to poor hygiene.

A black, hairy tongue is most often a result of antibiotic sensitivity, but it is also seen in those who use excessive amounts of peroxide mouthwashes, smoke or chew excessively, or constantly chew on various medicaments. The mouth can be burned by caustics and by exposure to extremes of temperature.

Tumours involving the blood-forming organs often interfere with the body's ability to control bleeding, and such bleeding is often first manifest in the mouth and throat, either as diffuse oozing from the gums or as multiple ulcerations of the entire buccal mucosa and tongue. Any disorder that causes a decreased effective circulation of blood platelets may result in the same kind of oozing from the gums. The serious disorder known as agranulocytosis, usually the result of a sensitivity to drugs, is suspected when ulcerative lesions of the mouth fail to heal and become secondarily infected. In this condition there is a sudden failure to produce white blood cells (leucocytes), so that the normal resistance of mucous membranes to their indigenous bacterial population cannot be maintained. In addition to these oral manifestations of diseases seated elsewhere, tumours may arise in the mouth and throat. Of these, cancers of the salivary glands and of the mucous membranes of the cheeks may be first observed from their effect on the mouth, where they may cause pain, bleeding, or difficulty in swallowing. The lymphomas and other tumours of lymphoid origin may first manifest themselves in the tonsillar or pharyngeal lymph nodes; these can usually be seen by direct inspection of the mouth and pharynx. Cancer of the tongue and of the bony structures of the hard palate or sinuses may project into the mouth or may burrow deep into the surrounding tissues. A precancerous lesion most often picked up by dentists is leukoplakia, a whitish discoloration of the mucous membrane of the mouth.

Primary disorders of teeth and gums

Any disorders of the bony framework of the mouth, including the lower and upper jaws and the hard palate, may produce difficulties with mastication or swallowing or may injure the oral mucous membranes. Differentiation of benign from malignant lesions in the oral cavity is the special area of expertise of dentists, but there is an overlap in the fields covered by otolaryngologists, who treat disorders of the ears and upper respiratory tract, by plastic surgeons, and by neurologists in solving complaints of persons suffering from dysfunction of this highly complex portion of the body. In the same way, pain or disability originating in the jaws, gums, and teeth may seem, when first complained of, to be pain associated with other disorders. Within relatively recent medical history teeth were suspected as the cause of many chronic disorders, especially arthritis, and were extracted for that reason rather than for any intrinsic dental disorder. Now, although careful inspections are made of the teeth and jaws of patients complaining of various chronic disorders, there is less willingness to extract teeth not obviously diseased.

Inflammation of the posterior wall of the mouth and of the tonsils and adjoining tissue on each side of the oropharynx, the central area of the pharynx, is very common especially in the young. Mention has already been made of streptococcal and staphylococcal infections of this area, but it should be emphasized that as many diffuse inflammatory responses of these tissues result from viral infections as from invasion by pyogenic bacteria. In viral pharyngitis the tissue is usually less violently red and swollen with fluid than is streptococcal pharyngitis and is less often covered by a whitish crust of even consistency. In acute streptococcal tonsillitis, both tonsils are swollen, pitted, and covered with whitish exudate; the soft palate is swollen and red, and there is much exudate seen streaming about the posterior pharyngeal wall. Other tonsillar tissue in the upper part of the pharynx and at the root of the tongue may be similarly involved. In diphtheritic pharyngitis the membranous exudate is more diffuse than in other types of pharyngitis, tougher, and extends over a much larger part of the mucous membrane of the mouth and nose. One of the complications of any tonsillitis or pharyngitis may be a peritonsillar abscess adjacent to one tonsil; this abscess is associated with bulging of the posterior pharyngeal mucosa over it. Surgical evacuation of such abscesses is usually required for healing.

Summary. The mouth and pharynx, being rather easily visualized by the physician, can be well characterized when diseases involve them. Symptoms of such disorders are usually localized by the patient, since this portion of his body is very heavily supplied with sensing devices, is geared for rapid action in eating and rejecting, and must carry out carefully coordinated actions, such as the simple ones of swallowing and breathing. As the first barrier that must be passed by ingestants or fluids, the area often suffers a variety of injuries from the environment, but its efforts are generally effective in reducing the variety of hostile agents that could attack the rest of the gastrointestinal tract.

THE ESOPHAGUS

This organ, a far less complex structure than the oropharynx, is designed to deliver to the stomach those materials that have successfully passed the pharynx. The esophagus does nothing to alter the physical or chemical composition of the material it receives, and it is poorly

equipped to reject materials that have got past the intricate sensors of the mouth and throat. Consequently, it is peculiarly vulnerable to mucosal injury from ingestants, as well as to materials that return into its lower segment from the stomach. Although its muscle coats are thick, it is not protected with a covering of serous membrane, as are neighbouring organs in the chest.

Inborn defects of the esophagus may be very serious; they are most often seen in infancy, primarily as a failure to develop normal passageways, often occurring between esophagus and trachea. Babies born with these openings cannot survive without early surgery. The lower end of the esophagus is subject to various developmental anomalies resulting in shortening of the organ so that the stomach may be pulled up into the thoracic cavity. Anomalies of the diaphragm may contribute to such an outcome.

Inflammatory disorders

Inflammatory disorders of the esophagus result from ingestion of noxious materials, as in lye or acid burns, from lodgment of foreign bodies, or from a complex of events associated with reflux of gastric contents from the stomach into the lower esophagus. All types of trauma produce damage to the mucous membrane of the esophagus. Inflammation resulting from surface injury by caustics is called corrosive esophagitis. When the problem is associated with reflux, the term peptic esophagitis is applied to the inflammation, which involves both the mucous membrane and the submucosal layer in a generally mild process. A number of other diseases may cause inflammation of the esophagus; *e.g.*, scleroderma, a disease in which the smooth muscle of the organ degenerates and is eventually replaced by fibrous tissue; and generalized moniliasis, in which the esophagus is often involved in a septic process characterized by many small abscesses and ulcerations throughout its entire length.

A straight tube of smooth muscle, the esophagus responds to localized inflammation by narrowing its passageway and, depending on the severity of the inflammation, producing a localized stricture. Such strictures have been known to be a consequence of ingesting lye or acid. They are readily seen by X-ray and by direct viewing with the esophagoscope, an instrument designed for this purpose. Treatment, after control of the infection, is surgical. With strictures brought about by interference with the emptying ability of the esophagus after inflammation caused by stomach juices, treatment is more complicated because it must involve control of the still-present causative processes, including acid-peptic stomach action and various structural variants of the lower esophagus.

An interesting disease of the esophagus is achalasia, formerly called cardiospasm. In this disorder a primary disturbance in the peristaltic action of the esophagus results in failure to empty the organ of its ingested and swallowed contents. The lower sphincteric portion of the esophagus does not receive its customary signal to relax and, over the months and years, may become hypertonic. A vicious cycle is thus set up in which the main portion of the esophagus slowly becomes distended, holding a column of fluid and food that it cannot propel downward, and the lower exit valve mechanism stays closed because of a failure in its information-feedback system. In most patients with this disorder there is a shortage of the ganglion cells of the intermuscular plexus (Auerbach's plexus) or they are diseased, so that coordinated peristalsis becomes impossible. In Chagas' disease, the infecting trypanosomes invade the neural tissue and directly destroy the ganglion cells. These organisms are not present in the temperate zones of the world, however, and the reason for ganglion cell degeneration in achalasia is unknown. Effective treatment is achieved by destroying the ability of the lower segment of the esophagus to contract; this may be done by dilating the esophagus with an inflated balloon or by surgery.

Tumours

Tumours of the esophagus may be benign or malignant. Generally, benign tumours originate in the submucosal tissues, and principally are leiomyomas, or tumours composed of smooth muscle tissue, or lipomas, which are of adipose tissue. Malignant tumours are either epidermoid cancers, made up of unorganized aggregates of cells, or adenocarcinomas, in which there are glandlike formations. Cancers arising from squamous tissues are found at all levels of the organ, whereas adenocarcinomas are more common at the lower end where a number of glands of gastric origin are normally present. Tumours generally make their presence known by producing difficulty in swallowing, particularly of solid foods. They are much more common in men than in women, and seem to vary widely in their worldwide distribution. Some correlations with the alcohol or tobacco consumption of the population have been established, suggesting that continued ingestion or inhalation of these materials may predispose to malignant change, as, apparently, does previous stricture formation. Treatment by surgery or by irradiation is largely palliative at present.

Disturbances of the esophagogastric junction, the point where esophagus and stomach meet, are both common and important. In many persons material from the stomach flows back into the esophagus, only to be immediately swept down into the stomach once again. New methods of measuring the pressure in the lumen or passageway of the esophagus and stomach have led to the belief that there is an effective barrier to such reflux from the stomach, and that this barrier is inherent in the musculature of the last few centimetres of esophagus. It is not clear whether anatomical displacement of the junction is in itself enough to efface the pressure barrier, but there is little question that an intrinsic mechanism of the esophagus does function to prevent gastric reflux and that in many patients in whom the radiologist can demonstrate a definite hiatus hernia—*i.e.*, a protrusion of a portion of upper stomach through the opening in the diaphragm—there is no reflux of gastric juice. On the other hand, most patients with reflux do have some evidence of hiatus hernia. The principal symptoms of the reflux syndrome are heartburn, a moving burning sensation under the breastbone, made worse by bending over or lying down and usually relieved rapidly by antacid medication. Some of such patients show reddening of the mucous membrane in the lower esophagus, but many do not.

Biopsies of the area may or may not show inflammation, and rarely is there good correlation between symptoms and intensity of inflammation. Treatment by hourly use of antacids, by small meals, by raising the head of the bed for sleeping, and by weight reduction in the usually obese patient has usually been quite successful. In some patients these measures do not suffice, or there is in addition some difficulty in swallowing; it is in such patients that surgical correction of the hernia is indicated. Many antispasmodic drugs, by interfering with the normal motor function of the esophagus, may make the symptoms worse; furthermore, since they slow down gastric emptying, the esophagogastric junction may then more likely be breached by reflux of juice from the distended stomach.

Diverticula

Diverticula of the gastrointestinal tract—pouches in the walls of the structures in the digestive system—occur wherever weak spots may exist between adjacent muscle layers. In the upper esophagus there is such an area: the striated-constrictor muscles of the pharynx merge with the upper smooth muscle of the esophagus just below the larynx. Some males over age 50 will show protrusion of a small sac of pharyngeal mucous membrane through the space between these muscles. As aging continues, or if there is motor disturbance in the area, this sac may become distended and actually fill with food or saliva. It usually projects to the left of the midline, and may demonstrate its presence by producing bubbling and crunching sounds during eating; often the patient can feel it in the left side of the neck as a lump, which can be reduced by pressure of the finger. Sometimes the sac may get so large it compresses the esophagus adjacent to it, producing a true obstruction. The treatment is surgical. Small diverticula just above the diaphragm sometimes are found after introduction of surgical instruments into the esophagus.

The stomach juices are much less likely to ulcerate the esophagus than to cause inflammation. Ulcers occur only in esophageal tissue next to glands of gastric origin in the esophagus, glands that secrete acid and pepsin. Because the esophagus is not protected by serous membranes, such ulcers may penetrate to neighbouring structures, most commonly the aorta, and can thus produce fatal hemorrhages. A serious injury to the esophagus is spontaneous rupture, a catastrophe that can occur in patients who have been vomiting or retching and often in debilitated elderly persons with chronic lung disease. Emergency surgical repair of the perforation is required.

THE STOMACH AND DUODENUM

In areas of the stomach and adjacent small intestine exposed to the action of gastric secretion, the mucous membrane may become ulcerated. If this breakdown of the mucosa is acute in onset and superficial in extent—*i.e.*, if it does not extend into the muscular coat of the stomach or intestine—it is called an erosion or acute superficial ulceration. If the damage extends more deeply into the wall of the organ, with resultant fusion of mucosa, submucosa, and muscularis into a chronic inflammation, it is known as peptic ulceration.

Under certain not completely understood circumstances the action of the gastric juice is directed at the tissues lining the stomach and intestine, resulting in either erosion or peptic ulceration. The present belief is that there is a rather delicate balance between the aggressive action of acid-pepsin and the resistance to this action maintained by healthy mucous membranes. The membranes of the stomach consist of epithelial cells covered with a flowing stream of mucus. The gastric glands secrete into the lumen of the stomach a number of products, including acid, pepsin, and various soluble forms of mucus. (See DIGESTION, HUMAN.) It is apparent that a complex neurohumoral regulatory system controls not only the secretion of gastric juice but also the compensatory defense mechanisms of the mucous membrane. Studies of persons who have peptic ulcers show that they differ in some ways from healthy persons, but that none of these differences is really unique. Thus persons with stomach ulcers are more likely to belong to blood group A, whereas those with duodenal ulcers are more likely to belong to blood group O. Persons with ulcers in the stomach have total rates of gastric acid secretion no different from a normal population, whereas those with duodenal ulcers as a group have rates of acid production greater than those of the general population.

Peptic ulcer

Peptic ulcer of the stomach occurs about equally in men and women, tends to occur above age 35, and to be a complicating factor in the lives of persons suffering from other disorders—in particular diseases of the nervous system and the lungs. Duodenal ulcer is much more common in men, generally makes its presence known from age 15 to 25, and does not often arise as a complication of other disorders. Persons with these diseases complain usually of pain in the upper portion of the abdomen and in the midline; the pain usually is absent on arising in the morning, tends to become troublesome in the late morning, is relieved by eating, only to return about one or two hours later after the stomach has emptied. Such episodes may occur infrequently and last only a few days or may become persistent for many weeks. As the ulcer becomes more active, inflammatory reaction around it usually interferes somewhat with the emptying of the stomach or duodenum producing retention of gastric contents. If this progresses, the patient may begin to vomit large quantities of retained food and to complain of unpleasant sour taste and more severe pain shortly after eating. About one-quarter of all patients with peptic ulcer will have blood in the stools at some time in the course of the disease; acute and massive bleeding occurs when the burrowing of the base of the ulcer strikes an artery of some size. In a few patients massive vomiting of blood may be the manner in which the ulcer first makes itself known. The burrowing of the ulcer deep into the wall of stomach or duodenum is known as penetration, and its ultimate result is perforation, permitting the leakage of stomach contents into the general peritoneal cavity. When this happens, either a local peritonitis with walling-off of the leak occurs, or a more general peritonitis ensues. Such patients are extremely ill and need emergency surgical care.

Peptic ulcer is a common disease in the Western World, being found in about 10 percent of all men, and in about two percent of women over the age of 50. It is generally a benign process that responds to conservative management with frequent feeding and the use of antacid medications. It is recurrent and episodic, often related to increased emotional tension, fatigue, or at a time when a person has an acute attack of another disease. In the gastric ulcer patient the differential diagnosis is complicated by the possibility of malignancy in the area of the ulcer, and thus the management of such cases is more often carried out in hospitals where direct viewing of the stomach lining with a gastroscope, study of cells in gastric juice, and frequent study of the stools for blood loss can be performed. Gastric ulcers can bleed massively and often interfere with eating to a degree not often noted with duodenal ulcer; thus gastric ulcer is treated by direct surgical excision more often than is duodenal ulcer. In about 15 percent of all cases, ulcers in the duodenum do not respond to medical treatment and it may be necessary to add to this treatment surgical measures to remove the vagus nerve fibres, which normally stimulate secretion of acid by the stomach. Treatment of the secretory portion of the stomach with X-rays can accomplish this in patients thought not to be good risks for surgery. The overall death rate for peptic ulcer in the U.S. and the Western countries is about four per 100,000. This is a low case-fatality rate for a common disorder, producing much hospitalization and a total cost in work time lost and direct costs estimated at $1,000,000,000 annually in the U.S. alone.

Gastritis

Diffuse or patchy inflammations not ascribable to invasion by bacteria, viruses, or parasites are quite common in the stomachs of persons, whether apparently healthy or complaining of a variety of digestive symptoms. The inflammation of stomach mucosa is known as gastritis. It is described in different terms depending on whether it is seen through a gastroscope, by X-ray, or by microscopic study of a biopsy obtained from the stomach. Except for atrophic gastritis and for acute erosive gastritis, in which large portions of the mucous membrane have been destroyed by caustic agents, drugs, or by unknown causes, none of the other types of inflammation can be said to be associated with any consistent symptoms or long-term complications. Thus at the present time the general term gastritis has no medical meaning and should not be used to denote a clinical disorder. It is now thought that at least part of the atrophic process, relative to the mucous membrane, is immunologic—that is, various antibodies to specific cells or products of the stomach glands can be demonstrated in the biopsies, in the gastric juice, and in the circulating blood. In pernicious anemia, in which atrophy of the gastric mucosa is usual, there are present antibodies to the material intrinsic factor, a mucoprotein necessary for the absorption of vitamin B_{12}, a substance essential for the manufacture of red blood cells.

Malignant tumours of the stomach are common throughout the world but show remarkable variations, probably involving both genetic and environmental factors, in incidence from country to country. Cancer of the stomach often occurs in older persons whose stomachs are capable of making only small quantities of acid. Whether this indicates that the same process that results in acid-secretory depression is also neoplastic or that gastric cancer is inhibited normally by acid-peptic secretion cannot be answered at present. Gastric cancer affects men more than women and currently accounts for about 20 percent of all deaths from cancers of the gastrointestinal tract in the U.S. In Japan it accounts for nearly 70 percent of such cancers.

Other malignant tumours that involve the stomach are

tumours ordinarily made up of lymphoid and connective tissue respectively. Benign tumours, especially leiomyomas, are common and may, when large, cause massive hemorrhage. Polyps of the stomach are not common except in the presence of gastric atrophy. Treatment for all these tumours, benign or malignant, is surgical.

Symptoms produced by tumours of the stomach are highly variable, and there are no truly characteristic evidences of disease in the early stages of the tumours. Most often the patients have loss of appetite, some weight loss, and symptoms ascribable to an anemia, which is frequently present and is due to blood loss into the stools, constant but usually so minimal as to escape notice by the patient. As the tumour grows, or if it is situated in the lower part of the stomach, obstructive symptoms brought on by eating may be noted, and tumours high in the stomach may obstruct the esophageal entry into the stomach, producing difficulty in swallowing. Pain is usually mild but, on the other hand, may be the most impressive feature. Stomach cancers often spread to neighbouring lymph nodes or to the liver very early, accounting for the very low percentage of surgical cures.

Diseases of the duodenum

The duodenum, aside from being the site of duodenal peptic ulceration, is otherwise not an important seat of disease. It is, however, often involved by diseases of its neighbours, in particular the pancreas and the biliary tract. Primary cancer of the duodenum is an infrequent problem. Benign tumours, particularly polyps and carcinoids, are more frequent. Cancers of the common bile duct or of the pancreas are important causes of death and may make their presence known by what they do to the duodenum, particularly in terms of obstruction and pain. It is because of their encroachment on the duodenum that these entities often are diagnosed by upper intestinal X-ray studies. Benign anomalies of the organs of this area, like an encircling ring of pancreas, may also encroach upon the duodenum. Parasites, particularly roundworms and tapeworms, are often found anchored in the duodenum. In inflammations of the pancreas, the neighbouring duodenum is often involved in such a way as to produce impairment of motility, and occasionally ulceration with hemorrhage.

THE SMALL INTESTINE AND APPENDIX

Congenital malformations of the small intestine consist of various duplications, which are infrequent, and of Meckel's diverticulum, which is common. The latter occurs when the duct leading from the navel to the small intestine fails to atrophy. In the fetus the duct serves as a principal channel for nourishment from the mother. The diverticulum in the child or adult may range from a small nubbin to a tube a foot or more in length, and it contains cells derived from the stomach glands that can secrete acid and pepsin. If such secretion occurs near the normal intestinal mucosa, which is totally nonresistant to such acid-peptic aggression, the mucosa will ulcerate and often bleed. Thus a peptic ulceration can develop at a site far distant from stomach or duodenum, can give rise to pain, bleeding, or obstruction, and must be treated surgically. A third congenital problem in the small intestine is the presence of multiple diverticula—outpouchings of mucosa and serosa. These are usually seen in elderly persons, although occasionally one may be the site of acute inflammation in a young adult. In the elderly, bacteria flourish in these diverticula which have no motor activity and cannot empty themselves. The bacteria eat nutrients otherwise destined for the body's own economy and may produce serious malabsorption.

The commonest of acute inflammations of the small intestine is acute enteritis, usually viral. This extraordinarily common ailment is manifested as acute malaise, fever, and diarrhea, often with severe cramps and pain. If there is vomiting as well, one calls the illness gastroenteritis. The disease is ordinarily short-lived but may produce serious disturbances of dehydration in infants and cause elderly subjects suffering from other disorders to become seriously ill. Species of the genus *Salmonella* cause acute bacterial infections of the small bowel; typhoid fever is the best known of these. Stool culture usually establishes the diagnosis of bacterial infections, which may require specific antibiotic treatment. Many of these infections, however, are self-limited and require no treatment other than replacement of fluids and salts.

Chronic inflammations of the small intestine include tuberculosis, now fortunately infrequent, and regional enteritis (Crohn's disease), currently on the increase. These disturbances are difficult to diagnose early, since their initial symptoms are often vague. Generally fever of low grade is present; there is a tendency toward loose stools and weight loss; and there may be episodes of crampy abdominal pain due to obstruction of the lumen and interference with normal muscular activity by the inflammation of the bowel wall. Diagnosis is usually suggested by the X-ray appearance of the intestine. Differentiation of tuberculosis from Crohn's disease is not always simple but usually can be accomplished using standard techniques. There is specific drug therapy for tuberculosis. In Crohn's disease anti-inflammatory, nonspecific drugs are quite helpful. Surgical excision or bypass of the offending segments of bowel may be necessary.

Appendicitis

Appendicitis, inflammation of the vermiform appendix, may be caused by infection or partial or total obstruction. It is still a major cause of intra-abdominal pain, principally attacking those under 35. Fortunately, it is easily diagnosed, and, with current surgical practice, the mortality from the acute process in young adults and children has dropped to very low levels. Widespread use of antibiotics for upper respiratory and other diseases may have actually lessened the incidence of the acute form of this disorder, so that more cases of later-developing appendiceal abscess are now being reported. Parasitic worms also can contribute to its incidence. In a woman of the childbearing age the differential diagnosis of appendicitis is still difficult, and many such patients are operated on because the necessary distinction cannot be made otherwise.

Intestinal obstruction may occur from pressure on the lumen of the bowel by tumour, inflammation, or local areas of adhesion, or from twists of the intestine on the folds of membrane that support it, or as a result of the devitalization of the intestinal wall by diminished blood supply. Such episodes of obstruction may develop slowly or with alarming rapidity. Recognition depends on clinical acumen plus laboratory assistance, of which the X-ray film is of the greatest importance. In all instances the condition is a grave one and demands immediate treatment. When generalized vascular disease—*i.e.*, heart failure or severe degenerative disease of the major vessels—exists, management may be best exercised by inserting tubes into the intestine and leaving them there; these relieve the distended segments of gas and fluid. Otherwise, the pressure is relieved by surgical treatment, and the offending segment or tumour is removed when possible.

Malabsorption occurs because the small intestine is unable to carry out its specialized functions of transporting properly prepared digestive materials from the lumen of the bowel into the lymphatics or mesenteric veins, from which they are distributed to the rest of the body. Such defects in transport occur either because the intestinal absorptive cells lack certain enzymes, whether by birth defect or by acquired disease, or because they are hindered in their work by other disease processes that infiltrate the tissues, disturb motility, permit bacteria to overpopulate the bowel, or block the pathways over which transport normally proceeds. A special case is that of inadequate bowel surface due to surgical resection, an increasingly common problem as life is maintained longer in persons with degenerative diseases.

THE LARGE INTESTINE

Disturbances of motility

The function of the large intestine, or colon, is to transport the fluid contents it receives from the small intestine, while dehydrating them sufficiently to permit

the passage of solid feces under controlled conditions. The proximal, or right, side of the colon is peculiarly fitted to carry out this absorptive function, during which sodium is reabsorbed, potassium and bicarbonate are secreted into the lumen, and about 400–500 millilitres (about a pint) of water is withdrawn.

Disturbance of motor and absorptive function is produced by a number of conditions involving the large intestine. Imperfect fetal development may result in an anus that has no opening; this defect may require major plastic surgery to correct. Abnormal rotation of the colon is fairly frequent, and may lead to no, or to serious, disorders. Long mesenteries supporting the cecum (right-sided tip of the colon) or the sigmoid colon (that section of colon just above the rectum) may permit recurrent twisting, which affects the blood supply to the involved loop, and the loop may itself be completely obstructed by the rotation. Such difficulties are usually seen in elderly patients and particularly in those with a long history of constipation.

A disease which has some analogy with achalasia of the esophagus is idiopathic (cause unknown) aganglionic megacolon, or Hirschsprung's disease. In this condition there is defective neuromuscular transmission in the Auerbach's plexus of autonomic nerves, usually only in the lower segment of the bowel, but occasionally involving most, or even all, of the organ. In areas thus poorly innervated, peristaltic waves do not progress, and a "functional obstruction" occurs. The area of normal bowel above the obstruction tries harder to push the fecal contents distally, and eventually thickening of the muscle of the normal segments occurs. The entire colon thus slowly becomes more and more distended and thick-walled. This kind of colonic obstruction must be differentiated by X-ray from "acquired megacolon," usually a psychiatric problem due to inhibition of defecation. In the latter there is no obstructed segment of colon but only a hugely dilated upper rectum and a tight anus. Various surgical procedures have been devised to help patients with Hirschsprung's disease. Psychiatric measures usually can completely relieve acquired megacolon.

Diverticulosis and diverticulitis

Arteries penetrate the muscular walls of the colon from its outside covering, the serosa, and distribute themselves in the submucosa. The channels in which these arteries lie may be regarded as potential hernia tunnels. With aging, and perhaps in predisposed persons, these channels become larger. If the peristaltic activity of the colon maintains a high pressure within its lumen, as in patients straining at stool, the mucous membrane of the colon may slowly be driven into these channels and eventually follow the arteries back to their site of colonic entrance in the serosa. At such time the outward-pushing mucosa will become a budding sac, or diverticulum, on the antimesenteric border of the colon but with a connection to the lumen. In the Western world (but not in the Orient) multiple colonic diverticula occur in as many as 30 percent of the over-50 population. In the presence of the irritable colon syndrome a different form of diverticulosis is seen accompanied by a peculiar hypertrophy of colonic muscle. Diverticulosis is common in populations consuming a low-roughage diet, and rare in those with a high-roughage intake.

The principal dangers of diverticulosis are massive hemorrhage and inflammation (diverticulitis). Hemorrhage results from trauma of hard stools against those small arteries of the colon which are exposed and have lost their support due to diverticula. As the arteries age, they become less elastic, less able to contract after bleeding begins, and more susceptible to damage. Diverticulitis, on the other hand, occurs because the narrow necks of the diverticula may become plugged with debris or inedible foodstuff, and bacterial proliferation in the blind sacs proceeds unihibited by the usual motor activity that keeps the bowel clean. When such abscesses enlarge, the bowel wall next to them becomes inflamed and irritable, muscle spasm occurs, and the patient has abdominal pain and fever. If the abscesses continue to enlarge they may rupture freely into the peritoneum, giving rise to peritonitis. More commonly they fix themselves to neighbouring organs and produce localized abscesses, which may prove very difficult to handle surgically. Mild diverticulitis responds well to conservative measures and to antibiotics; massive hemorrhage often requires emergency surgery; recurrent diverticulitis now is thought to warrant resection.

Infections of the colon

The colon may become inflamed because of invasion by pathogenic bacteria or parasites. A variety of species of *Shigella*, for example, attack the mucous membrane of the colon producing a very intense but rather superficial hemorrhagic response. In infants and in the elderly the amount of fluid and protein lost by the intense inflammatory response may be fatal, but ordinarily such symptoms are readily survived by otherwise healthy persons. *Salmonella* species, responsible for severe generalized infections originating from invasion of the small intestine, may cause damage in the lymph follicles of the colon but do not produce a diffuse colitis. The virus of lymphopathia venereum, on the other hand, can cause a diffuse superficial colitis. The most important parasite invading the colon and producing human disease there is the protozoan *Endamoeba histolytica*. This widely distributed parasite enters via the mouth and ordinarily lodges in the cecum and ascending colon of man. This usually results in irritability of the right colon and failure to absorb water properly, so that intermittent watery diarrhea ensues. The amoebae undermine the mucosal coat and may create large ulcerations that bleed impressively. Thus the stools contain blood, often much blood, but there is little pus or other evidence of reaction by the colon to the invading organism. In more generalized amoebic colitis, the rectum and sigmoid colon become invaded, and on direct visualization through the sigmoidoscope manifest their presence by numerous discrete ulcerations separated from each other by relatively normal-appearing mucous membrane. After identification of the parasites by direct smears from the margins of the ulcers, or from the stools, treatment with a combination of amoebicidal drugs plus a broad-spectrum antibiotic is carried out.

The colon can become the site of diffuse inflammation in which no invading organisms can be identified as causal. When this inflammation is confined to the mucosa and submucosa, the mucous membrane becoming diffusely reddened and bleeding at the slightest touch of a cotton swab, the term ulcerative colitis, either acute or chronic, is used to describe it. If the reaction is a proliferative one, the wall of the colon becoming thickened and all coats of the wall inflamed, the mucosa least of all, the term Crohn's disease of the colon is applied. These two diseases are not common but are disabling. As there is no specific etiology, a combination of anti-inflammatory drugs and other approaches to the patient must be used to help him over the worst episodes. Often it is necessary to resort to surgical techniques for partial or complete removal of the diseased bowel. The entire colon can be removed, and the small intestine brought out to the skin as an ileostomy, an opening to serve as a substitute for the anus, with generally acceptable results. When ulcerative colitis occurs in children, a high incidence of cancer of the colon is noted. In adults this relationship is less striking although still present.

Chronic overusage of laxatives can also produce mild inflammatory changes in the colon. Persons who have had constipation for many years often have elongated colons, occasionally with the mucous membrane bearing dark brown to black pigmentary deposits (melanosis coli), which have no pathological significance.

Tumours of the colon are usually polyps or cancers. The tendency of some persons to form polyps is nowhere more strikingly exemplified than in the rare disorder known as congenital polyposis, in which the colon may be studded with hundreds or thousands of small polyps. The kind of colon that can produce so many polyps eventually produces cancers as well; therefore, these patients should have their colons removed surgically as soon as the diagnosis is made. Another peculiar form

of polyp is the villous adenoma, often a slowly growing, fern-like structure that spreads along the surface for some distance. It possesses the potential for local recurrence after being locally resected, or it may develop into a full-blown cancer. Cancer of the colon is in the Western world a more common tumour than is cancer of the stomach, and it occurs in both sexes about equally. Symptoms are highly variable, the main unifying feature being blood in the stools, but this may be only detectable by chemical testing. Cancers compress the colonic lumen to produce obstruction, they attach to neighbouring structures to produce pain, and they perforate to give rise to peritonitis. They also may metastasize to distant organs before causing local symptoms. Nevertheless, the outlook for patients with this tumour is considerably better than is the case with cancer of the stomach. About half the patients who have colonic cancer removed surgically will live at least five years. Some patients must have colostomy, an opening from the colon to the skin, but this is usually easily managed.

Anorectal problems

Problems related to defecation are more common in the Western world than elsewhere. Whether this distribution is related to diet, exercise, or to social customs inhibiting the natural gratification of the defecation urge is not clear. In any case, anal problems plague Western persons from childhood on. These usually take the form of fissures—cuts or cracks in the skin or mucous membrane at the junction of the anal mucous membrane with the skin between the thighs. Sometimes these fissures become chronically infected and resistant to simple sitz baths and local medication. Under such circumstances they require surgical correction. Anal fistulas occur sometimes as complications of serious bowel disease, as in tuberculosis or Crohn's disease of the bowel, or in certain parasitic diseases. This complication requires surgical treatment. A more general form of disturbance is the enlargement of veins of the rectum and anus. The dilated veins become either external or internal hemorrhoids, or both. Probably half the adult population in the West have such venous enlargement, but only a small number of persons suffer significant symptoms from their presence. Hemorrhoids protrude, are associated with anal itching and pain, and they bleed, especially when in contact with hard stools. Most of the time these symptoms can be controlled by conservative measures, but occasionally they persist or cause so much acute distress that surgical removal of the enlarged and dilated veins is necessary.

In the past, poor obstetrical and postnatal care resulted in severe loss of support of the pelvic floor in parturient women, and prolapse of the rectum, as well as of the uterus and bladder, was common. Present day obstetrical practice has largely eliminated such complications; rectal prolapse is still seen in conditions such as fibrocystic disease of the pancreas in children or in aged persons with certain neurological disorders.

Abscesses in the perianal area, complicating any of the above conditions, are common. Hygiene is difficult to maintain in the elderly, the obese, or the feeble confined to chairs and bed. Careful control of skin bacteria and avoidance of trauma to the perineum are difficult to assure even under optimal conditions. Fungous infections of this moist and poorly cleansed area are common, permitting maceration of tissue and invasion by skin and colonic bacteria. In the diabetic patient, who is so susceptible to skin infection, scrupulous perianal hygiene is of great importance.

Primary cancer of the anus is fortunately a rare disease. Chronic involvement of perianal skin by the wart-like lesions of syphilis or the scarring of the venereal disease lymphopathia venereum are well-known but less frequent than formerly. Crohn's disease of the colon and intestine has a predilection for producing anal involvement with stricture.

THE DIGESTIVE GLANDS

The salivary glands contribute their secretions to mix with the food being masticated. Consequently, any interference with their function by inflammation, tumour, or calculous obstruction will lessen the amount of secretion available. The patient may have a dry mouth, swelling of one of the glands, or pain in the region of the duct and the gland supplying it. The virus disease, epidemic parotitis, or mumps, involves one or both parotids. The other glands are more often inflamed because a small stone has become impacted in the orifice of the secretory duct, and the resultant stasis may have allowed bacteria to proliferate in the stagnant stream. Thus abscesses in the blocked duct-gland system are common under these conditions and must be surgically repaired. The parotid glands, the salivary glands below and in front of the ears, which are very active in the synthesis of enzymes, are also subject to peculiar enlargement or swelling in the presence of severe malnutrition and other debilitating diseases, especially sarcoidosis, a chronic inflammatory disease of unknown cause. A disorder of connective tissue metabolism known as Sjögren's syndrome is associated with generalized dryness of the mouth as a result of involvement of these glands.

The salivary glands are also subject to neoplastic changes. The commonest tumour, the mixed tumour of the parotid, is a relatively slow-growing malignancy that produces a very hard swelling at the angle of the jaw below the ear. Various lymphomas may also invade the parotid and submandibular glands; in leukemias more than one gland may be infiltrated.

Liver diseases

The liver is the largest single gland in the body, accounting for one-fortieth the body weight. The interaction between gut and liver is an important bodily regulatory function, and diseases in either organ can affect the function of the other. Disease processes involving the gut can cause the leakage of such quantities of protein into its lumen that the liver's synthetic abilities for making albumin are exceeded, leading to a generalized deficiency in this circulating plasma protein. Primary inflammations of the liver are often preceded by a period of nausea, diarrhea, and abdominal discomfort. Both acute and chronic inflammations of the liver result in a defective output of bile, leading to poor absorption of fat from the gut, and the appearance of excessive fat in the stools. Stools become light in colour if bile is prevented from entering the duodenum either because of liver disease or of blocks in the biliary ducts. Primary infections of the intestinal tract can permit their causative agents to get into the portal blood flow and travel to the liver, where they set up habitation and produce abscesses and various reactive processes. The commonest examples of this are the amoebic abscess of the liver and the purulent liver abscess secondary to appendicitis or diverticulitis. Cancers of the stomach, pancreas, and colon ultimately metastasize to the liver, which they enlarge and often make tender and painful. A consequence of chronic liver disease, as in cirrhosis, is an obstruction to the inflow from the veins draining the intestines. This leads to enlargement of these veins and of the spleen; the most serious complication of this process is rupture of veins into the stomach or esophagus, producing usually severe blood loss, with vomiting of blood and the passage of black or bloody stools (see also LIVER, HUMAN).

Diseases of the gallbladder and bile ducts

The gallbladder in the human is so frequently removed because of disease that ample proof of its dispensability is available. The principal factors underlying gallbladder disease in man are concerned with the results of the concentration of the bile that takes place in the gallbladder. The bile contains in an aqueous medium a number of substances that are difficult to keep water-soluble. These include cholesterol, lecithin, and bile salts. The proper combination of these three entities is importantly concerned with rendering cholesterol soluble as it flows from the liver to the gallbladder and down the common bile duct. Disturbance in the ratio of cholesterol to bile salts plus lecithin may permit cholesterol to precipitate. The cholesterol concentration in bile may rise, as seems to happen in late pregnancy, or the liver may not be

able to manufacture sufficient bile salts or lecithin. Since the gallbladder usually empties after every meal, and perhaps more often, very small imbalances of this type may, over the years, lead to the slow deposition of cholesterol in the wall of the gallbladder and impair its function. Small amounts of precipitated cholesterol plus other materials probably underlie the formation of the cholesterol-calcium gallstones seen in human gallbladder disease. The disease occurs much more frequently in women than in men, but by age 70 its incidence in both sexes exceeds 15–20 percent in Western populations. The other main type of gallstone, the bilirubin stone, is unrelated to this process, being rather the product of excessive breakdown of hemoglobin, and most often related to hemolytic diseases.

In any case, the formation of gallstones, cholelithiasis, is related both to inflammation of the gallbladder, cholecystitis, and to the presence of gallstones in the common bile duct, choledocholithiasis. Acute cholecystitis can be a severe abdominal catastrophe, with gangrene and rupture of the gallbladder, or it can be a mild chronic problem diagnosed by recurrent episodes of discomfort under the right portion of the lower ribs. Gallstones, if small, can be pushed by contraction of the wall of the gallbladder into the conical neck of the organ, where they may become impacted. The gallbladder then is obliged to contract forcibly against an obstruction; this produces pain of recurrent or colicky type, often accompanied by vomiting. The pain is frequently referred to the back near the lower tip of the shoulder blades and occasionally to the tip of the right shoulder. The gallbladder may become chronically enlarged if the obstruction continues; this is called hydrops of the gallbladder if the contents of the organ remain uninfected; when infected, empyema of the gallbladder is produced, and usually fever and chills accompany the finding of the enlarged organ. In general, the treatment of these complications is entirely surgical.

Should the gallstones leave the gallbladder, they may become impacted at the lowest portion of the common bile duct in what is called its ampulla. This area occurs at the junction of the duct with the medial wall of the descending, or second, portion of the duodenum, near the entrance of the principal duct from the pancreas. A stone impacted in this narrow opening may completely or incompletely obstruct the flow of bile from the liver, producing jaundice, often with itching and chills. The liver may be unable to secrete the bile against the obstructing pressure, and the jaundice may rapidly increase. On the other hand, the stone may crack or disintegrate and be passed from the common duct into the duodenum, the bile flow resume, and the jaundice disappear over a period of a week or so. Proper diagnosis of this chain of events depends upon clinical observation, certain liver function tests, and radiologic demonstration of the gallbladder and the ducts.

Long-standing obstruction may result in scarring of the ducts or strictures. Sometimes the surgical intervention—necessary to remove the gallbladder or the offending gallstones—may lead to scarring and stricture. Strictures, or narrowed portions of these hollow tubes, present opportunities for more obstruction, either by newly formed gallstones or by inflammatory exudate and mucus. The strictures may require surgical correction. Administered antibiotics of many kinds are concentrated by the liver and excreted in the bile, so that serious infections in this area are not the problem they once were.

Tumours of the gallbladder and of the bile ducts are serious but infrequent problems. Cancer of the gallbladder occurs almost entirely in persons who suffered from cholecystitis and is more common in women than in men. It is hard to diagnose and its surgical treatment is not very satisfactory. Primary cancer of the bile ducts makes itself known by producing jaundice relatively early. Since it tends to be a slowly growing tumour, reasonably favourable results may be anticipated from surgical and radiation treatment.

The pancreas, a flat structure lying across the back of the peritoneal cavity, secretes enzymes of digestive importance that are very powerful. Their liberation in areas outside the gastrointestinal tract may lead to serious destruction of normal tissue. Thus diseases of the pancreas may either deprive the intestinal lumen of enzymes needed to digest food properly, or may liberate them into the blood or peritoneal cavity, where they may digest normal fat and protein structures. Most importantly, inflammation within the substance of the pancreas itself may allow activated enzymes to attack the blood vessels of the organ, with the consequent production of hemorrhage into the substance of the gland.

The various possibilities are included within the concepts of pancreatitis, which may also be mild with swelling of the gland and none of the hemorrhagic or widespread destruction noted above. Attacks of edematous pancreatitis, often associated with gallbladder disease or acute alcoholic excesses, produce abdominal pain, fever, nausea, vomiting, and sometimes mild diabetic features. In acute hemorrhagic pancreatitis there is acute pain, shock, and hemorrhage within the gland. It is often fatal, since it occurs principally in persons with underlying vascular or renal disease. In alcoholic subjects repeated attacks of mild to severe nonhemorrhagic pancreatitis are associated with a chronic deterioration of pancreatic function. Sometimes the gland becomes laden with calcium deposits, sometimes a permanent diabetic state is produced, and sometimes the reduction of digestive enzymes produces a serious loss of nutrients in the frequent loose stools passed by such persons. Diagnosis of these various disorders is made by utilizing the clinical signs and symptoms, employing certain laboratory tests for pancreatic enzymes in the blood and duodenal juices, and by chemical analysis of the stools, which characteristically show increased amounts of unabsorbed fats and proteins. X-ray examination of the stomach and duodenum may also be helpful. The pancreas lies behind the stomach and is bounded by the duodenum; thus disturbance in its shape or configuration may be reflected by pressure on these neighbouring structures.

Diseases of the pancreas

Although most cases of pancreatitis are the consequence of poorly understood metabolic derangements, direct trauma to the organ occasionally occurs, and may produce an inflammatory pancreatitis. In addition, several familial forms of the disease have occurred. An entirely different disorder of this organ is that associated with mucoviscidosis, a generalized disorder of the exocrine glands, that is, glands with ducts such as the mucous and sweat glands. This disease, also known as fibrocystic disease of the pancreas, occurs in infants and children in an inherited pattern characterized as autosomal recessive. As such, it is the most frequent lethal genetic disorder among the non-Negroid populations of the world. The secretions of the pancreas become thickened and plug the pancreatic ducts, leading to destruction of the organ. Similar processes cause chronic lung disease, with the bronchial lumina obstructed and then dilated behind the obstruction. Inability to control the loss of salt in the sweat leads to death in hot weather. Diagnosis in children is made by measuring the high concentrations of sodium and chloride in the sweat, and by demonstrating absence or great reduction in the amount of pancreatic enzymes in the duodenal juice.

Treatment of acute inflammatory diseases of the pancreas is principally supportive. Occasionally surgical intervention is necessary to drain blood clots or abscesses, or to remove diseased gallbladders when they coexist. Chronic pancreatitis is treated by trying to persuade the patients to discontinue alcohol consumption. There are available very good preparations of concentrated pancreatic enzymes, which can be fed to patients with poor enzyme output; such treatment can partially replace the deficiencies, make the patient eat better, and decrease the loss of foodstuffs into the stools.

Fibrocystic disease is treated with pancreatic replacement substances, with antibiotics for the pulmonary disease, and with various techniques for mechanically improving lung ventilation. The outlook is still poor for sur-

vival, although now fewer children are dying of the disease at very early ages.

Cancer of the pancreas is one of the most rapidly increasing cancers in the U.S., having reached the same incidence as cancer of the stomach. The reasons for this are not clear, but there seems to be some association with better nutrition, or perhaps affluence, in the populations subject to the disease. On the other hand, cancers of all the digestive glands are thought to be more common in Africa, possibly as a result of protein deficiency in childhood.

BIBLIOGRAPHY. H.L. BOCKUS *et al., Gastroenterology*, 2nd ed., 3 vol. (1963–65), a comprehensive and well-illustrated textbook with emphasis on diagnostic methods and treatment; M. PAULSON (ed.), *Gastroenterologic Medicine* (1969), deals comprehensively but not systematically with digestive diseases approached from a variety of disciplines; F.A. JONES *et al., Clinical Gastroenterology*, 2nd ed. (1968), a shorter, easy-to-read textbook that covers all the important disease entities and is critical in its approach to diagnosis and therapy.

(A.I.M.)

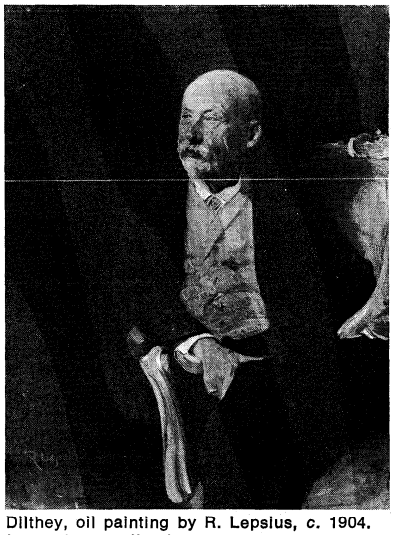
Dilthey, Wilhelm

Wilhelm Dilthey was a German philosopher of the late 19th and early 20th century whose chief contribution to philosophy was the development of a distinct methodology for the humanities. He objected to the pervasive influence of the natural sciences and developed a philosophy of life that perceived man in his historicity; *i.e.*, in his historical contingency and changeability. Dilthey established a comprehensive treatment of history from the cultural viewpoint that has been of great consequence, particularly to the study of literature.

Life and works

Dilthey was born on November 19, 1833, in Biebrich on the Rhine, the son of a Reformed Church theologian. After he finished grammar school in Wiesbaden, he began to study theology, first at Heidelberg, then at Berlin, where he soon transferred to philosophy. After completing exams in theology and philosophy, he taught for some time at secondary schools in Berlin but soon abandoned this to dedicate himself fully to scholarly endeavours.

Archiv fur Kunst und Geschichte



Dilthey, oil painting by R. Lepsius, c. 1904. In a private collection.

During these years he was bursting with energy, and his investigations led him into diverse directions. In addition to extensive studies on the history of early Christianity and on the history of philosophy and literature, he had a strong interest in music, and he was eager to absorb everything that was being discovered in the unfolding empirical sciences of man: sociology and ethnology, psychology and physiology. Hundreds of reviews and essays testify to an almost inexhaustible productivity.

In 1864 he took his doctorate at Berlin and obtained the right to lecture. He was appointed to a chair at the University of Basel in 1866; appointments to Kiel, in 1868, and Breslau, in 1871, followed. In 1882 he succeeded R.H. Lotze at the University of Berlin, where he spent the remainder of his life.

During these years Dilthey led the quiet life of a scholar, devoid of great external excitement and in total dedication to his work. He searched for the philosophical foundation of what he first and rather vaguely summarized as the "sciences of man, of society, and the state," which he later called *Geisteswissenschaften* ("human sciences")—a term that eventually gained general recognition. In 1883, as a result of these studies, the first volume of his *Einleitung in die Geisteswissenschaften* ("Introduction to Human Sciences") appeared. The second volume, on which he worked continually, never did appear. This introductory work yielded a series of important essays; one of these—his "Ideen über eine beschreibende und zergliedernde Psychologie" (1894; "Ideas Concerning a Descriptive and Analytical Psychology")—instigated the formation of a cognitive (*Verstehen*), or structural, psychology. During the last years of his life, Dilthey resumed this work on a new level in his treatise *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften* (1910; "The Structure of the Historical World in the Human Sciences"), which was also left unfinished.

Opposed to the trend in the historical sciences to approximate the methodological ideal of the natural sciences, Dilthey tried to establish the humanities as interpretative sciences in their own right. He considered as fundamental to this notion the interaction among personal experience (*Erleben*), its realization in creative expression, and the reflective understanding of this experience. He did not view the individual as isolated but always in the context of his environment. He emphasized that the essence of man cannot be grasped by introspection but only from a knowledge of all of history; this understanding, however, can never be final because history itself never is: "The prototype 'man' disintegrates during the process of history." For this reason, his philosophical works were closely connected to his historical studies. From these works later arose the encompassing scheme of his *Studien zur Geschichte des deutschen Geistes* ("Studies Concerning the History of the German Mind"); the notes for this work make up a complete coherent manuscript, but only parts have been published.

Dilthey did not have the ability for definitive formulation; he was suspicious of rationally constructed systems and preferred to leave questions unsettled, realizing that they involved complexity. For a long time, therefore, he was regarded primarily as a sensitive cultural historian who lacked the power of systematic thought. Only posthumously, through the editorial and interpretative work of his disciples, did the significance of the methodology of his historical philosophy of life emerge. Dilthey died on October 1, 1911, in Seis on the Schlern.

Assessment

It is difficult to assess Dilthey the man. Even his more intimate disciples confessed to have known very little of his deeper feelings. Only the few invited to collaborate in his later years became somewhat familiar with him. They shared almost his entire day, reading to him, taking dictation, even drafting complete passages for him. They learned while being involved in various aspects of his work. And yet, each one perceived only one facet; nobody had full comprehension of the whole. Even to them, Dilthey remained the "strange mysterious old man."

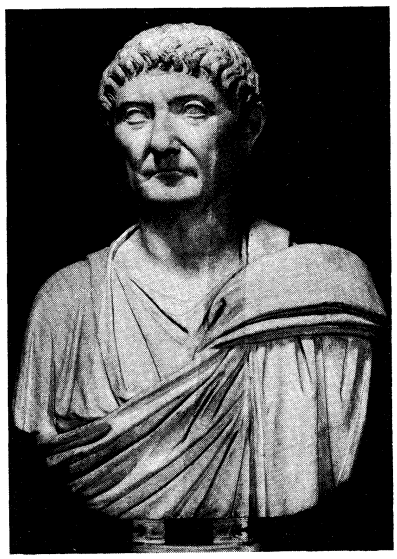
BIBLIOGRAPHY. A detailed biography is lacking, although HANS-HERMANN GROOTHOFF and ULRICH HERRMANN, "Wilhelm Dilthey: Persönlichkeit und Werk," in *Die Pädagogik Wilhelm Diltheys*, pp. 334–365 (1971), is a carefully executed recent work. A complete bibliography appears in ULRICH HERRMANN, *Bibliographie Wilhelm Dilthey* (1969). Dilthey's works are collected in his *Gesammelte Schriften*, 16 vol. (1923–72), with more volumes in preparation. Works on Dilthey in English include HERBERT ARTHUR HODGES, *Wilhelm Dilthey: An Introduction* (1944, reprinted 1969), with translations of selected excerpts; and *The Philosophy of Wilhelm Dilthey* (1952); and H. PETER RICKMAN, *Meaning in History: Wilhelm Dilthey's Thoughts on History and Society* (1961), which contains a general introduction and selected writings.

(O.F.B.)

Diocletian

Diocletian, Roman emperor from AD 284 to 305, restored efficient government to the empire after the near anarchy of the 3rd century. His reorganization of the fiscal, administrative, and military machinery of the empire laid the foundation for the Byzantine Empire in the East and temporarily shored up the decaying empire in the West. His reign is also noted for the last great persecution of the Christians.

Alinari



Diocletian, bust in the Capitoline Museum, Rome.

Diocletian's biography has been obscured by legends, rhetoric, the dubiousness of documents, and the hostility of his adversaries. He died at the age of 68 in his villa at Salonae in Dalmatia in AD 313 or 316. As he is said to have died "after having returned to his homeland," it is possible he was born in that town in 245 or 248; little is known of his origins. His father was a scribe or the emancipated slave of a senator called Anullinus. Diocletian's complete name, found in official inscriptions, is given as Gaius Aurelius Valerius Diocletianus. He received the name Diocles first, then the name Valerius, after the name of his daughter, Valeria, who married Galerius in 293. The gens name Aurelius did not appear until March 1, 286; that is, until after his accession. Nothing is known of his wife Prisca other than what the contemporary Latin Christian writer Lactantius Firmianus says in his *De mortibus persecutorum*, which is of debatable veracity. Diocles, having adopted the name Diocletianus, entered history like so many of those emperors who emerged from the shadows through force of arms, brought to power by the army. The little that is known of his appearance is based on coin effigies and on sculptures. From these it appears that he was tall and thin, with a large forehead, a short, strong nose, a hard mouth, and a determined chin.

Personal appearance

Rise to power. Up to the time of his accession, Diocletian had lived most of his life in military camps. These may have been either in Gaul, as reported in the *Historia Augusta*, or in Moesia. Or he may have been a member of the Roman emperor Carinus' bodyguard. The only definite fact known about Diocletian during this period is that he was among those army chiefs whom Carinus gathered, together with the Illyrians, to fight against the Persians. In 284, during that campaign, Numerian, Carinus' brother and co-emperor, was found dead in his litter, and his adoptive father, the praetorian prefect Aper, was accused of having killed him in order to seize power. When Diocletian, acclaimed as emperor by his soldiers, appeared for the first time in public dressed in the imperial purple, he declared himself innocent of Numerian's murder. He designated Aper as the criminal and killed him personally. Here again, rhetoric has obscured the real events. Aper's guilt was accepted by contemporaries, but it was also true that a prediction had been made to Diocletian previously, telling him that he would become emperor on the day he killed a boar (Latin *aper*). And it was true, too, that he did not wish to wait much longer for the boar to come. In reality, Numerian had died either a natural death or from a stroke of lightning. By eliminating Aper, however, Diocletian rid himself of an eventual competitor and, retroactively, provided his act with sacred meaning.

Acclaimed emperor on November 17, 284, Diocletian had real power only in those countries dominated by his army (*i.e.*, in Asia Minor and possibly Syria). The rest of the empire was obedient to Numerian's brother, Carinus. After having put down a revolt by Julianus, a troop commander in Pannonia, whom he attacked and killed near Verona, Carinus next attacked Diocletian. An indecisive battle near the confluence of the Margus (modern Morava) and Danube rivers, not far from present-day Belgrade, would have been a defeat for Diocletian had Carinus not been assassinated by a group of soldiers. Thus, in midsummer of 285, Diocletian became master of the empire.

Reorganization of the empire. At the beginning of 286 Diocletian was in Nicomedia. In the interim, he and his lieutenants had calmed the stirrings of revolt among Roman troops stationed on the frontiers. From that point on, he dedicated himself to restoring civil order to the empire by removing the army from politics.

Division of responsibility

Although he came from the army's ranks, Diocletian was not, properly speaking, a soldier. He had scarcely come to power when he made an unexpected decision—to share the throne with a colleague of his choice. The empire was too great for one man to administer; nearly every week, either in Africa, or somewhere on the frontier that extended from Britain to the Persian Gulf, along the Rhine, the Danube, the Pontus Euxinus (Black Sea), and the Euphrates, he was forced to suppress a revolt or stop an invasion. Diocletian, who was more attracted to administration, required a man who was both a soldier and a faithful companion to take responsibility for military defense. In 286 he chose Maximian, an Illyrian, the son of a peasant from the area around Sirmium. A little later, though still keeping Rome as the official capital, he chose two other residences. Maximian, who was responsible for the West, was installed at Milan in northern Italy, in order to prevent German invasions. Diocletian established himself at Nicomedia, in western Anatolia, and close to the Persian frontier in order to keep watch on the East. Six years later, in 293, having taken the title of "Augustus" and given it to Maximian as well, he added two more colleagues: Galerius, a former herdsman, and Constantius I Chlorus, a Dardanian nobleman according to the legend of his house, but a rather rude countryman also. These additional collaborators were each given the title "Caesar" and attached to an Augustus, Constantius to Maximian (with a residence in Trier), and Galerius to Diocletian himself (with a residence in Sirmium). Thus, while the empire remained a *patrimonium indivisum* (undivided inheritance), it was nevertheless divided administratively: Diocletian, residing in Nicomedia, watched over Thrace, Asia, and Egypt; Galerius, residing in Sirmium, watched over Illyria, the Danubian provinces, and Achaea; Maximian, residing in Milan, over Italy, Sicily, and Africa; and Constantius I Chlorus, residing in Trier, over Gaul, Spain, and Britain. In order to strengthen the union of the colleagues, each Augustus adopted his Caesar. The relationships were further cemented when Galerius married Valeria, Diocletian's daughter, and Constantius I Chlorus repudiated his wife Helena, mother of the future emperor Constantine, in order to marry Theodora, Maximian's stepdaughter. The Empire now had four masters, celebrated by the authors of the *Historia Augusta* (a collection of biographies of Roman emperors and caesars, published in the 17th century) as the *quattuor principes mundi* ("four princes of the world"), and Diocletian consecrated this human unity by forming a religious bond. Because he believed that he had come to power through divine will, as revealed by the "fateful"

boar, he regarded himself and Maximian as "sons of gods and creators of gods." After 287, he called himself Jovius (Jove) and Maximian was named Herculius (Hercules), signifying that they had been chosen by the gods and predestined as participants in the divine nature. Thus, they were charged with distributing the benefits of Providence, Diocletian through divine wisdom, and Maximian through heroic energy. Later designated as *dominus et deus* on coins and inscriptions, Diocletian surrounded himself with pomp and ceremony and regularly manifested his autocratic will. Under Diocletian, the empire took on the aspects of a theocracy.

Diocletian's reforms were successful; they put an end to domestic anarchy, and elsewhere they allowed Maximian to defeat the revolt in Gaul of the Bagaudae, bands of peasants who found the tribute oppressive. Then, with peace scarcely restored after a campaign against the Germans, Maximian had to battle Carausius, who, having fought for the empire in Britain against the Frankish and Saxon pirates, revolted and named himself emperor in Britain in 287. Carausius reigned in Britain for nearly ten years until Constantius I Chlorus succeeded in returning Britain to the empire in 296. Scarcely had troubles in Mauretania and in the Danubian regions been settled when Egypt declared itself independent under the usurper Achilleus. Diocletian reconquered the country in 296. Finally, in 297, he had to fight Narses, king of Persia, who had invaded Syria. Since he was still occupied in Egypt, he assigned this operation to Galerius, who, after a protracted campaign, finally won victory for the Romans. Tiridates, the king of Armenia and a protégé of the Romans, was able to return to his throne; the Tigris became the eastern border of the empire; and peace reigned in that part of the world until the reign of Constantine I (306–337).

Frontier revolts

Domestic reforms. Perhaps more important for the maintenance of the empire was Diocletian's program of domestic reform. He was not a complete innovator in this area, for his predecessors had made some tentative attempts in the same direction; the emperor Gallienus had excluded senators from the army and separated military from civil careers. The Senate had progressively been deprived of its privileges. Diocletian, however, systematized these arrangements in such a way that all his reforms led toward a kind of centralized and absolute monarchy that put effective means of action at his disposal. Thus, Diocletian designated the consuls; the senators no longer collaborated in the making of laws; the imperial counsellors (*consilia sacra*) were distributed among specialized offices, and their functions were strictly defined so that the power of the praetorian prefects (personal bodyguards to the emperor) was limited; the specialization of administrative work grew; and the number of bureaucrats increased. This was the beginning of the bureaucracy and technocracy that was eventually to overrun modern societies.

Such organization made it possible for administration to rely less on individual human beings and more on the application of legal texts. In fact, it was during Diocletian's reign that the Gregorian and Hermogenian codes, of which only fragments remain, were rewritten. But 1,200 extant rescripts show another aspect of the emperor's personality. A conservative, Diocletian was concerned with the preservation of the ancient virtues: the obligation of children to feed their parents in old age; of parents to treat their children justly; for spouses to respect the laws of marriage; for sons not to bear witness against their fathers, or slaves against their masters; and for private property, creditor's rights, and contract clauses to be protected. He forbade the use of torture if truth could be discovered otherwise and encouraged governors to be as autonomous as possible.

The army was also reorganized and brought back to the old discipline. Sedentary troops (local troops) were sent to the frontiers, and the ready army (main movable army) was made domestic. Troop strength was increased by a fourth, not multiplied by four as Lactantius claims. Here again, Diocletian's reforms were infused with a sense of human realities; he exempted soldiers from duty after 20 years of service, and if he limited the price of commodities so as to reduce the cost of living, it was mainly to make life easier for the troops. If one is to believe Lactantius, Diocletian divided the provinces "so as to make himself more feared," but in fact it was to bring the governors closer to those they administered and, by fragmenting their power, to diminish their territorial strength. He undertook to facilitate economic development through a recovery of agriculture and a program of building.

Such policies were expensive, as were wars and the legacy of an unstable financial situation. Diocletian's fiscal solutions are still debated; they constitute a very difficult problem. Two new taxes were instituted, the *jugum* and the *capitatio*, the former being the tax on a unit of cultivable land, the latter, a tax on individuals. Taxes were levied on a proportional basis, the amount of the contribution being determined by the productivity and type of cultivation. As a rule, it was a sort of socioeconomic taxation based on the linkage between man and land in terms of either ownership or productivity. Assessments were made every five years; later, the system was consolidated into a cycle of 15 years called an *indictio*. This census of taxable adults gave rise to violent criticisms but had the theoretical advantage of replacing the arbitrary levies of the previous era. To be sure, the financial system was subject to excesses; but Diocletian's purpose was to obtain funds, and he did not even spare Italy, which had until then been free of land taxation.

Tax reform

This reform was accompanied by a monetary reform, including restoration of a sound gold and silver coinage of fixed design, creation of a new bronze coin, circulation of small coins to facilitate daily financial exchange, decentralization of minting, and an increase in the number of mints from eight to 15.

All of these measures tended to stave off financial crises. The famous edict *De Maximis Pretiis* was issued in AD 301, fixing wages and establishing maximum prices, so as to prevent inflation, abusive profits, and the exploitation of buyers. About 1,000 articles were enumerated, and violation was punishable by death; severe penalties were exacted of black marketeers. But even so, this regulation of prices and wages was not enforceable, and the edict was later revoked.

Freezing of wages and prices

Persecution of Christians. The end of the reign was darkened by the last major persecution of the Christians. The reasons for this persecution are uncertain, but various explanations have been advanced: the possible influence of Galerius, a fanatic follower of the traditional Roman religion; the desire to restore complete unity, without tolerance of a foreign cult that was seen as separatist and of men who were forming a kind of state within the state; the influence of anti-Christian philosophers such as Porphyry and governors such as Hierocles on the scholarly class and on the imperial court; the fear of an alienation of rebellious armies from emperor worship; or perhaps the disturbances provoked by the Christians themselves, who were agitated by doctrinal controversies. At any rate, some or all of these factors led Diocletian to publish the four edicts of 303–304, promising all the while that he would not spill blood. His vow went unheeded, however, and the persecutions spread through the empire with an extreme violence that did not succeed in annihilating Christianity but caused the faith of the martyrs to blaze forth instead.

Retirement. Diocletian had aged prematurely through illness. Perhaps he decided that, after 20 years of reign, his abdication was also "fateful." Of his own volition he decided to entrust the affairs of the empire to younger men and returned first to Nicomedia, then to the neighbourhood of Salona, on the edge of the Adriatic, where he had a magnificent palace built (the modern town of Split, in Yugoslavia, occupies the site of its ruins). He abdicated May 1, 305, and his death occurred almost unnoticed.

Diocletian had reorganized the empire without political romanticism. His reforms had not proceeded from a premeditated plan but had imposed themselves out of historical necessity. He may be accused of several things: of

Assessment

having been cruel, but his harshness was not the act of deep-seated brutality; of being miserly, but this miserliness was inspired by the desire to obtain resources for the state; of cutting a slightly muddle-headed, visionary figure, but these were the traits that led him to reflect on better methods of governing an immense territory; of having paved the way to bureaucracy and technocracy, but this was done with greater efficiency in view. Personally, Diocletian was a religious man. No doubt he did not manifest any unusual piety, but he always thought that the gods of the emperors governed the world. He exercised an absolute, "divine right" monarchy and he surrounded it with an almost Oriental majesty.

Partially he failed in his task, and one can rightly say that the state he created was not "the new house he intended to build, but rather an emergency shelter," which offered protection against storms with the help of the gods. The fact remains that he was, in his actions, his religion, and his time, *vir rei publicae necessarius*, "the man whom the State needed."

BIBLIOGRAPHY

Ancient Latin sources: Panegyrici latini; ARNOBIUS, *Adversus nationes;* LACTANTIUS, *De mortibus persecutorum;* EUSEBIUS, *Historia ecclesiastica;* AURELIUS VICTOR, *Caesares*, 39–40.

Modern works: W. ENSSLIN, "Valerius Diocletianus," in *Pauly-Wissowa Real-Encyclopädie*, 2nd series, vol. 14, col. 2419–2495 (1948), a general survey of Diocletian's career; W. SESTON, *Dioclétien et la tétrarchie* (1946); G. COSTA, "Diocletianus," in E. DE RUGGIERO, *Dizionario epigrafico*, vol. 2, pp. 1793–1908 (1922); E. FAURE, *Étude de la capitation de Dioclétien d'après le "Panégyrique VIII,"* vol. 4 of *Études de droit romain* (1961), and technical discussion; *The Cambridge Ancient History*, vol. 12 (1939), with topical bibliography.

(Je.C.)

Diplomatics

Diplomatics, broadly speaking, is the study of documents. The term is derived from the Greek word *diploma*, meaning "doubled" or "folded." Besides the documents of legal and administrative import with which it is properly concerned, diplomatics also includes the study of other records such as bills, reports, cartularies, registers, and rolls. Diplomatics is therefore a basic and not simply an auxiliary historical science. This article deals with its development and practice in the Roman Empire and in Europe. During Roman antiquity certain documents containing different sorts of authorizations were engraved on a bronze diptych and then folded and sealed, in order to keep the contents secret—hence the term diploma. Rarely found during the Middle Ages, the word was used by the Renaissance Humanists to denote formal documents of ancient rulers. The interest in and description of such documents came to be called *res diplomatica* after the famous 17th-century work *De Re Diplomatica Libri VI*, by Jean Mabillon, a member of the scholarly Benedictine congregation of Saint-Maur. Mabillon's work first made the study of old documents a reputable science. The major task of diplomatics is to distinguish between genuine and false documents, and this involves detailed examination of their external and internal features. Diplomatic studies have been applied mainly to Western documents, usually medieval ones, because it requires less specialist training to analyze more recent documents.

HISTORY OF THE STUDY OF DOCUMENTS

The forging of documents

Medieval and Renaissance work. The forging of documents took place on a vast scale during the earlier Middle Ages, partly because wars and disturbances so frequently upset possession and also because the increasing use of written records made it necessary for those whose title was, in fact, perfectly good in old unwritten "customary" law to give it written substantiation. Thus forgeries, partly intentionally honest, partly dishonest, occurred frequently, despite the fact that the Germanic tribes that settled in western Europe inherited, with other aspects of Roman law, the concept of forgery as a felony, which was soon also reinforced by the church's canon law. This legal concept of forgery was, however, mainly applied to cases concerning property or inheritance; and literary forgeries, such as the famous Donation of Constantine, which purported to be the gift by the Roman emperor Constantine I the Great (died 337) to Pope Sylvester I of spiritual primacy throughout the church and of temporal power in Italy, were not concerned. Serious critical efforts to detect forgery did not begin in the Middle Ages, although obvious forgeries might be challenged in the course of a dispute. As early as the 6th century, the Merovingian king Childebert II declared a charter recording the gift of land from himself to the Bishop of Reims a forgery on the simple ground that the royal official denied the signature on it to be his. Pope Innocent III (1198–1216) tried to establish infallible criteria for the detection of fraudulent papal documents, but knowledge of earlier documentary forms was totally inadequate. In the Renaissance the Humanists began to use philological and technical criteria; on these grounds Lorenzo Valla authoritatively pronounced the Donation of Constantine to be a forgery, though authenticity had already been questioned.

Post-Renaissance scholarship. Three events in the 17th century forced the development of more sophisticated standards of evaluation. The Thirty Years' War in Germany led to endless legal conflicts, and in France the nobility engaged in a concerted action known as the *bella diplomatica* ("diplomatic wars") to assert their ancient privileges against royal absolutism. The decisive impetus, however, came from a much more particularist dispute. Daniel van Papenbroeck, a member of the Jesuit commission known as the Bollandists (from another member, Jean Bolland), which was charged with the publication of the *Acta Sanctorum* ("Acts of the Saints,"), finding that some monastic documents he inspected were forgeries, assumed (1675) that this was true of almost all early-medieval documents. Since most of the monasteries with which the documents had been concerned were of the Benedictine Order, the Benedictines resented the suggestion, and Mabillon undertook to refute it. In his *De Re Diplomatica* (1681), Mabillon set out the fundamental principles of the science of verifying documents; Papenbroeck soon afterward acknowledged the correctness of his tenets. Nearly a century later, René-Prosper Tassin and Charles-François Toustain published their six-volume *Nouveau traité de diplomatique* (1750–65; "New Treatise on Diplomatic"), a work that surpassed Mabillon's only in its greater wealth of material. Another important event in the history of the science of diplomatics was the founding of the École des Chartes (an institute for the training of French archivists) in Paris in 1821. During the next decades important collections of early-medieval French documents were printed in the *Recueil des actes* by a variety of eminent editors. But the greatest advances were made by German and Austrian scholars, among whom Julius von Ficker investigated the differentiation between *actum* and *datum* (that is, between verbal legal procedure and its formal documentation), and Theodor von Sickel defined a basic technique of studying and comparing the script of charters and thus of identifying the individual notaries or scribes. The diplomas of the Carolingian and the German kings and emperors were edited in the series of the *Monumenta Germaniae Historica*, by members of the Institut für österreichische Geschichtsforschung (Institute of Austrian History Research), established by Sickel in 1854. Meanwhile, the *Regesta*, comprising short, synoptical condensations of the contents of papal documents down to 1198, published by Philipp Jaffé in 1851, gave a decisive momentum to the study of the papal chancery, while August Potthast covered the period from 1198 to 1304. Prominent scholars in the research of papal records in Germany at the beginning of the 20th century were Michael Tangl, Rudolf von Heckel, and, particularly, Paul Fridolin Kehr.

Lack of study in England

In comparison with the amount of work done in France and Germany, historical scholarship in England long paid relatively little attention to legal, as opposed to literary, records. Although John Mitchell Kemble published his collection of Anglo-Saxon documents, the *Codex Diplo-*

maticus Aevi Saxonici (1839–48), an extensive study of Anglo-Saxon and Norman legal and administrative documents was delayed until the 20th century. Since then notable contributions have been made by scholars such as Helen Cam, H.W.C. Davis, Vivian Hunter Galbraith, Frank M. Stenton, Dorothy Whitelock, David Charles Douglas, and many others. Christopher Robert Cheney has made important contributions to the research of papal documents. In Italy Luigi Schiaparelli made vital contributions to the study of Lombard documents. From the 19th century, some study of documents has formed part of the medieval-history curriculum in most European universities.

DIPLOMATIC METHOD

Types of documents. Documents that have been preserved are either originals, drafts, or copies. Originals, of which many survive, are formal documents drawn up on the order of the sender or donor, and they were designated to serve the recipient or beneficiary as evidence of the transaction recorded. Handwritten copies of documents, made either before or after the deed was actually executed (sealed), are not classified as originals. If made before an "original," they were in fact rough drafts of it; if made afterward, they were copies. The particularly Anglo-Saxon method of chirography, however, gave the possibility of producing several "originals." By this process two or more specimens of a document were written on the same page of the vellum sheet, and the free space between the texts was filled in with the word *chyrographum* ("handwriting") or other words and symbols. Then the sheet was cut irregularly right through these words or symbols; the originals thus separated could later be reassembled, an exact fit being complete proof of authenticity. But to provide documents having the force of "originals," copies of the original were usually made and formally certified as such, by public notaries, or by high ecclesiastical or secular dignitaries. Copies certified in this way were accorded the same legal value as the originals. In practice, lack of critical judgment on the part of the certifiers often led to the certification of forged records. In documents known as transumpts, which recited earlier documents or charters as part of their text, it often happened that the earlier document was forged, but, being included in the new, it received validation. The original documents and copies considered above were issued at the request of the recipient or beneficiary or of his legal heir. It also happened quite often that the sender or donor wished for various reasons to retain a record of the documents issued by him. The chanceries (record offices) of secular rulers or great ecclesiastics therefore kept copies of outgoing documents in registers, and often of incoming documents, too. The popes were among the first to adopt the old Roman practice of keeping registers; although nearly all the earlier ones have been lost, an almost uninterrupted series of papal registers is extant from the pontificate of Innocent III onward. An important group of registers are the rolls kept by the medieval kings of England; the earliest extant rolls date from the 12th century. The keeping of registers in the chanceries of the French kings began about the year 1200, in Aragon about 1215, in Sicily under the Hohenstaufen emperor Frederick II (died 1250), and in the German imperial chancery from the early 14th century. Another manner of studying documents is in the formula books of the various chanceries. Notaries drawing up the various forms of medieval documents did not usually compose each new text afresh but, rather, copied from books in which such text formulas had been collected, a practice that can be traced back to Roman procedure. These model texts frequently contained only the legally relevant passages, while the individually applicable parts, such as names, figures, and dates, were either abridged or totally omitted. During the time of the Frankish kings, important collections were made, such as the *Formulae Marculfi* (early 8th century) and the *Formulae imperiales* (828–832). Significant collections of formulas serving as models for papal documents have been preserved from the 13th century.

Use of formulas

Classification of documents. The documents of the Middle Ages are usually classified under two groups: public documents, which are those of emperors, kings, and popes, and private documents, which comprise all others. Another way of classifying documents is according to whether they are evidentiary or dispositive. The former merely record a valid legal act already executed orally, while the actual issuing of the latter forms in itself the legal act. This distinction, found among Roman documents from the 3rd century AD onward, gradually ceased to exist after the early Middle Ages. After the collapse of the Carolingian empire in the 9th century, private documents lost much of their function and were replaced by simple memorandums about legal acts and the witnesses to them. It was not until the late 11th and early 12th centuries that sealed charters of high secular or ecclesiastical dignitaries were again gradually considered as dispositive. Papal documents can be classified mainly as either letters or privileges, and royal documents can be classified as diplomas or mandates. Privileges and diplomas give evidence of legal transactions designed to be of long duration or even of permanent effect, while mandates and many papal letters contain commands.

Physical appearance of documents. *Material and ink.* Documents were written on a variety of material. In antiquity there were documents of stone, metal, wax, papyrus, and, occasionally, of parchment, but only papyrus and parchment (and, very occasionally, wax) were used during the Middle Ages. From the 12th to the 13th centuries, paper also was sometimes available. Papyrus, made from the stem of the papyrus plant, was produced mainly in Egypt; after the Arab conquest of Egypt in the 7th century, the import of papyrus to Europe became difficult. The Merovingian kings wrote their documents on papyrus until the second half of the 7th century, and the popes did so until far into the 11th century. North of the Alps papyrus had finally disappeared by the 8th century, when it was replaced by parchment. Parchment was made from animal hides and was thus easier to obtain. In southern Europe it was made mainly from sheep and goat hides; the insides of the skin were thoroughly smoothed and calcined, while the hairy sides were left rougher. In central and northern Europe, parchment was usually made from calves' skins, and both sides of the hides were thoroughly smoothed and calcined. Paper came originally from China. During the 8th century AD, it spread to the Arab world and from thence to Byzantium, where it was manufactured from linen and was used from the 11th to the 13th centuries for imperial documents. After that time ordinary paper was used in the Byzantine Empire. In the West the use of paper, most common at first in southern Italy and Spain, had begun to spread by the beginning of the 12th century. Germany and southern France began to import paper from Spain and Italy in the 13th century, and soon afterward it had reached England by way of Bordeaux. But paper did not altogether replace parchment, which long remained in use, especially for solemn documents. The medium for writing was ink, generally a mixture of oak gall and copper vitriol. Originally black, ink made north of the Alps sometimes shows a reddish-brown hue, while that made in Italy may contain tinges of brown and yellow. Over the centuries most of these colours have lightened as a result of atmospheric conditions. The Byzantine emperors used purple ink for their signatures. This custom was occasionally taken over by the Lombard rulers of Italy and, later, by the Norman kings of Sicily. Another custom of Byzantine origin is the use of gold lettering.

Introduction of parchment

Language, script, and abbreviations. Throughout the entire Roman Empire, the language used in documents was primarily Latin. Greek was also used, and, during the latter part of the 6th century AD, it slowly superseded Latin in the East. From then onward, Greek was the language of Byzantine documents until the end of the Byzantine Empire (1453). In the West, the collapse of the empire and the establishment of barbarian kingdoms led to a vulgarization of Latin, written as well as spoken.

Use of vernacular

Latin has always been used for papal documents and for most public and private charters, and it was used for international documents well into post-Renaissance times, until it was superseded by French as the language of diplomacy. In public and private documents, use of the vernacular alongside Latin gradually developed. Apart from its early and unique appearance in the documents of the Anglo-Saxons in England, no vernacular was used in charters before the 12th century. At the Norman Conquest (1066), use of Anglo-Saxon in English documents soon stopped, and no more vernacular was used there until some Norman French was introduced in the 13th century, and Middle English in the 15th century. There was an increasing use of the vernacular in Italian and French documents from the 12th century and in Germany from the 13th; but in medieval times Latin was never outstripped by the vernacular.

A correct assessment of the hand in which it was written is vital to ascertaining the provenance and authenticity of a document. Thus, the knowledge of paleography, different styles of ancient writing, is a skill essential to diplomatics. The broad basis of such knowledge begins with acquaintance with the general styles of writing current at particular times and places. This varied with the way the pen was held; whether the writing was cursive or had the letters formed separately; whether it was majuscule, all the letters being contained between a single pair of horizontal lines, or minuscule, with parts of the letters extending above and below the lines. There is a further distinction between what is called book hand and the business, or court, hand at one time used for documents (see CALLIGRAPHY).

In Europe the Roman capital letters, distinguished as rustic or square, uncial, and Roman majuscule and minuscule cursive, influenced all subsequent writing in the West. The Roman curial style (from the Curia, or papal court), used in the papal chancery until the 12th century, was a derivation of late Roman minuscule cursive. After the disintegration of the Western Empire, the Merovingian Franks used a Roman provincial script for their documents. Distinctive forms developed elsewhere, in Visigothic Spain and in Ireland. The Irish script, a half uncial (uncials are rounded letters) and a minuscule script, spread to Anglo-Saxon England and thence to the European continent. Under the Carolingian rulers, a particularly clear and attractive minuscule book hand was developed; modifications of this gradually became used in documents and eventually spread also to Italy, England, and Spain. A "Gothic," more pointed form of script developed since the 11th century in northern France and soon spread all over Europe, so that writing became more spidery in appearance. In the early years of the Renaissance, Italian scholars such as Poggio (Poggio Bracciolini) and Niccolò Niccoli developed a minuscule based on the Carolingian, and variants of this style were used by the Venetian Aldus Manutius and other pioneers of printing.

Abbreviations were used in both documents and books. Again, their particular characteristics would contribute to a correct assessment of the probable date and provenance of a document. Roughly two types were used: the suspension, involving the writing of only the first letter or syllable of a word; and the contraction, used first for Hebrew and Christian sacred names, the writing of only the first and last letter or letters of a word or syllable. The Carolingians sometimes used Tironian notes, a form of shorthand devised by Tiro, a freedman of the Roman orator Cicero.

Validation of documents. From Roman times the two most important methods of validating documents were by appending the signature or the seal of the sender or promulgator. The practice of using seals for this purpose (and not merely to close a document) was carried over from imperial usage and, by the 8th century, was current among the Lombards and other Germanic tribes in western Europe. Until about the 8th century, the signature of the Merovingian ruler or his delegate was also required for the validation of public documents, but thereafter the seal alone, together with the recognition by a high chancery official, was held sufficient, the king's signature dwindling into a monogram or mere stroke (the "stroke of execution"). This change was probably accelerated because many medieval kings could not write. Thus, in England, King John sealed, and did not sign, the Magna Carta.

Seals

Seals were made of wax or of metal; if the latter, they were called bulls (hence, the use of this term for a certain group of papal documents). The Byzantine emperors used gold seals for their documents; Byzantine officials and ecclesiastics used lead and silver for their bulls. Papal seals were of lead or gold. Wax seals were increasingly used from about the 11th and 12th centuries, and wax was also used for the impression when, later, less formal documents were validated by use of the signet or privy seal. Seals could be two-sided, suspended from the document, or impressed upon it.

Form and content of documents. Normally, each document was divided into three distinct parts: the introduction (protocol), the main text (context), and the concluding formulas or final protocol. There were various subdivisions, and not all the parts here mentioned are necessarily found in every document. The introduction comprises, first, the invocation (*invocatio*) of God, either by name or through a symbol such as the cross; second, the superscription (*intitulatio*), giving the name and title of the sender; and third, the address (*inscriptio*), naming those to whom the document is directed, usually followed by a formula of greeting (*salutatio*). The actual text of the document can be divided into a number of parts. The first, known as the *arenga*, expresses in general terms the motive for the issue of the document. The notification (*promulgatio*), briefly explaining the legal purpose of the document, is followed by the *narratio*, or exposition of the particular circumstances involved. In the *dispositio* the donor or promulgator firmly declares his purpose ("I hereby decree" or "I hereby give"); this clause is the vital core of the document, its legal decree of enactment. There usually followed the *sanctio*, a threat of punishment should the enactment be violated. The main text concluded with the *corroboratio*, a statement of the means to be used for validation of the document. The final protocol consisted of subscriptions or lists of names of all those, such as the scribe, who took part in the issue of the document and of witnesses to the enactment. The date and place of issue are given, and the final sentence, the *apprecatio*, is a short prayer for the realization of the contents of the charter. At the bottom of the document, the signs of validation (the recognition, monogram, seal) were then added.

Problems of chronology

The date given on a document might be either that of legal enactment (*actum*) or that of the issue of the document recording the (already performed) legal enactment (*datum*). The form in which dates are given in a document is of particular import in determining its provenance and authenticity. A wide variety of practices were followed at different places and times. For instance, days of the month could be given according to the old Roman system of calends, ides, and nones; by continuous counting throughout the month; or by reference to a saint's day. Years might be computed from the presumed time of the creation of the world; by the Roman indiction, a 15-year cycle; by the names of officiating Roman consuls; by regnal years of emperor, king, or pope; or from the birth of Christ. Moreover, there were also a variety of ways to determine when the year began.

DEVELOPMENT AND CHARACTERISTICS OF CHANCERIES

The Roman and Byzantine empire. Rulers, all of whom needed to issue directives and edicts, developed writing offices, or chanceries, in which formal documents were drawn up. The Roman imperial chancery, called the Office of Letters (*ab epistulis*), was subdivided into a Greek and a Latin department. In the 5th century four letter offices existed, all under the ultimate control of the *magister officiorum* ("master of offices"): the *scrinium epistolarum* ("letter office") handled mainly foreign, legal, and administrative affairs; the *scrinium libellorum*

("petitions office") handled petitions and investigations; the *scrinium memoriae* ("memorandum office") composed shorter imperial decrees; and the *scrinium dispositionum* dealt with administration. From the 4th century, a group (*schola*) of notaries had come into being, some of whom served the emperor as personal secretaries. Two centuries later a special confidential (*a secretis*) secretary existed. In the Byzantine Empire in the 8th to 9th centuries, the *scrinium epistolarum* and the *scrinium libellorum* merged to form a new department under the *koiaistor* (a high palace official), while the secretaries had all come under the office of the *protoasekretis* (head of the secretaries). An official called the *mystikos* handled the emperor's secret correspondence. In preparing edicts or other laws, the *koiaistor*, after consulting the emperor, made a first draft of the bill, had the official copy drawn up by notaries, and then verified its accuracy before it was validated. From the 9th century onward other high court officials participated in the validation of Byzantine charters.

The important governmental documents of the late Roman and early Byzantine empires include laws, edicts, decrees (imperial decisions concerning civil and penal law), and rescripts (the emperor's replies to inquiries from corporate and administrative bodies or private persons). In the Byzantine era documents concerning more day-to-day affairs can be grouped under the headings of foreign letters, privileges, and administration. Foreign letters include correspondence with other rulers, treaties (regarded not as an agreement between equals but as an act of grace or privilege granted by the emperor, and made out as such), and letters accrediting imperial ambassadors. The most solemn and splendid form of privilege was the *chrysobullos logos*, so named because the word *logos*, meaning the emperor's solemn word, appeared in it three times, picked out in red ink. Written in the carefully embellished chancery script reserved for the emperor's personal documents, the text consists of the usual parts—that is, the *invocatio, intitulatio, inscriptio, arenga, narratio, dispositio, sanctio,* date, and the *subscriptio*. It was sealed with the golden bull. From about the 12th to the mid-14th century, a simplified form, the *chrysobullon sigillion*, was used for privileges of lesser importance. It was not signed by the emperor himself but was held to be validated by the insertion, by the emperor, in red ink of the *menologema*, a statement of month and indiction. It, too, was sealed with a golden bull. The administrative documents of the Byzantine imperial chancery include the *prostagma*, or *horismos*, a plain and short document known since the beginning of the 13th century. If directed to a single person, the document starts out with a short address, but, in all other cases, it begins immediately with the *narratio*, followed by the *dispositio*. The emperor replaced his signature with the *menologema*. Unlike the privileges, this document was not rolled up but, instead, was folded, and then closed by means of a wax seal stamped with the imprint of the imperial signet ring.

Imperial privileges

In addition to those emanating from the imperial offices, there were other types of documents issued in the Byzantine Empire. These include those issued by despots and imperial officials and, in the ecclesiastical sphere, by patriarchs and bishops. There were also private documents. The documents issued by the despots carried a silver seal, showing their intermediate status between that of imperial documents sealed with the gold seal and that of documents drawn up by imperial officials and sealed with lead bulls. Documents issued by imperial officials were simpler. They lacked the protocol, and the personal signature of the issuing official was written in black ink. The detailed date comprises the *menologema*. The documents of the Byzantine patriarchs are in many respects analogous to the imperial documents and symbolize the high status of the patriarch of Constantinople. They were, however, sealed with lead bulls. Byzantine private documents are almost exclusively notarial instruments. They are immediately recognizable by the crosses marked at the top of the documents. Used in lieu of the signature, the cross was the mark of the sender and contained his name and official function in one of its angles. The document was either signed by witnesses, or at least the cross preceding their names is autograph. Following that is the signature of the notary. These documents were not usually sealed.

The papal chancery. Knowledge about early papal documents is scant because no originals survive from before the 9th century, and extant copies of earlier documents are often much abridged. But it is clear that the popes at first imitated the form of the letters of the Roman emperors. The papal protocol consisted only of the superscription and address and the final protocol of the pope's personal "signature"—not a mention of his name but merely a blessing. Toward the end of the 8th century, it became customary in certain documents to mention in the final clauses the name of the scribe responsible for the drawing up of the document; this was given with the date of issue, indicated by month and indiction, immediately following the subject matter of the document. There followed another clause, the great dating formula, *datum per manus* ("given by the hand of . . ."), naming a high chancery official and giving the date by reference to the regnal years of both emperor and pope. Both were used in documents containing decrees of permanent legal force, which came to be called privileges. Under Pope Leo IX (1049–54), the benediction written by the pope was changed into a monogram not written by him, but his signature was now introduced, placed in a round symbol, the *rota*. By the early 13th century, papal documents had evolved into two distinctive groups: solemn privileges and letters. Solemn privileges can be distinguished by their enlarged letters (*elongata*) of the first line, by the phrase *in perpetuum* ("in perpetuity") at the end of the address, by a threefold *amen* at the end of the text, by use of the *rota*, the pope's signature, the monogram, signatures of the cardinals, and by the *datum per manus*. Among letters, those whose bull was fastened on silken cords (*litterae cum serico*) brought some benefit to the recipient, while those with bulls fastened on a hempen cord (*litterae cum filo canapis*) contained either orders or the papal delegation in a dispute.

Great dating formula

The number of solemn privileges began to decline from the mid-13th century, and eventually they were completely discontinued, their function being partly taken over by the *litterae cum serico*, which became increasingly elaborate in form. A new type of document also developed, the papal bull, distinguishable primarily by its use of formulas such as *ad perpetuam rei memoriam* ("that the matter may be perpetually known") in the superscription. Yet another new papal document appeared at the end of the 14th century, the brief (*breve*), used for the popes' private or even secret correspondence. Written not in the chancery but, instead, by papal secretaries (an office dating from about 1338), the briefs were sealed on wax with the imprint of the papal signet ring.

The papal chancery of the 4th to the 8th centuries was similar to the late-Roman imperial chancery. Its notaries (*notarii, scriniarii*), organized in a guild (*schola*), were headed by the *primicerius notariorum* and the *secundicerius notariorum* (first and second of notaries) and included the especially important notaries of Rome's seven ecclesiastical regions. But, during the 9th century, the *bibliothecarius*, the papal librarian, became the most important chancery official; a little later, various important bishops and dignitaries seem to have acted occasionally as *datarius* (the official named in the *datum per manus* formula). During the mid-11th century, a phase of German influence led to the temporary employment of notaries from the court of the emperor Henry III, who drew up papal privileges according to imperial formulas. A more important and permanent outcome of German influence was the gradual replacement of the *bibliothecarius* by a chancellor as the highest chancery official. The chancellor was invariably a cardinal, and in his absence another cardinal acted in his place as vice chancellor. Lesser chancery personnel still included the seven regional notaries; increasing business involved the use of lesser paid scribes in addition to the established notaries. From the late 11th century, a papal chapel,

Chancery officials

modelled on those of contemporary emperors and kings, developed, and its staff was often employed in chancery tasks.

From the early 13th century, the vice chancellor became the permanent head of the chancery, the office of chancellor remaining vacant. During that century the vice chancellors were ordinary clerics, who renounced the office if elevated to the cardinalate; thus, the chancery became directly subordinate to the pope himself. Both the numbers and the official standing of the notaries in the chancery, which then functioned entirely separately from the chapel, gradually increased. Higher chancery officials were often distinguished canonists (legal experts), such as Sinibaldo Fieschi (later Pope Innocent IV), Godfrey of Trani, and Richard of Siena. From the beginning of the 14th century, bishops or cardinals filled the office of vice chancellor. During the Great Schism (1378–1417) there were two papal chanceries and two vice chancellors, one in Rome and one in Avignon.

Under Innocent III the procedure of the papal chancery had changed. Letters concerning matters of import to the papal Curia (*de Curia*) were drafted by the pope himself or else by a cardinal, the vice chancellor, or a notary. But the majority of the papal documents were elicited by their recipients, who had first to present to a notary the substance of their petition in a form the text of which largely anticipated the wording of the desired document. Professional proctors attached to the Curia assisted in the drafting and were also responsible for the documents during later stages of the procedure. Once a petition was approved, the notaries or the *abbreviatores* drafted a suitable document, drawing on a selection of formula books. After a final copy (engrossment) had been made and checked, it was read, if necessary, to the pope or in a special department of the chancery, the *Audientia litterarum contradictarum*. It was then passed to the Cistercian lay brothers who had charge of the papal bull, sealed, and given to the petitioner, who had had to pay a fee at almost every stage of the proceedings.

Insufficient research has so far been done on the papal chancery during the 14th and 15th centuries. Whereas formerly, when the vice chancellor was absent, one of the notaries had deputized for him, a new official, the *regens cancellariam*, was now created to fulfill this function. The number of notaries increased steadily, and, from the 13th century onward, an increasing number of public notaries worked in the papal administration. In order to distinguish between them and the papal notaries proper, the latter became unofficially known as protonotaries. The notaries were now in charge of the letters of justice, while the letters of grace were handled by the *abbreviatores*. The scribes remained in charge of the engrossments. A *computator*, aided by several assistants, was responsible for collecting fees.

The royal chanceries of medieval western Europe. Of the nations that held power in western Europe after the collapse of the Roman Empire there, the Ostrogoths, who occupied Italy from the late 5th to the mid-6th century, took over the ancient Roman imperial-chancery system in its entirety. Very little is known about the royal documents of the Lombards, their successors in Northern Italy, since not one of them has been preserved in its original form. But Lombard officials in charge of drawing up the documents were still trained in the Roman tradition. As well as *referendarii*, there were notaries who also acted as scribes. It is very likely that all of them were laymen.

Until the 12th century two main types of documents, diplomas and mandates, were produced north of the Alps, in the Merovingian, Carolingian, German, and French royal chanceries. Very little is known about the Merovingian royal chancery and its organization. The names of the scribes are never mentioned in the documents, but they were signed by high chancery officials, the *referendarii*.

The Carolingian chancery. When the Merovingian dynasty was supplanted by the Carolingians, chancery procedure changed drastically. In contrast to the Merovingian kings, the first Carolingian king, Pepin the Short, was unable either to read or write. He therefore entrusted the responsibility for the correctness of the royal documents to an official of the court. At about the same time, the task of drawing up documents was taken over by those clerics whose original duty had been to look after the most important relic of the royal court, the coat (*cappa*) of St. Martin of Tours. Collectively named the *capella* (chapel), these clerks were individually called *capellani*, chaplains. This close connection between the court chapel and the chancery existed under the later Carolingians and at the German and French and other royal courts, including that of England. Until well into the 12th century, European chanceries were not bureaucratic offices in the modern sense but, rather, in most cases an assemblage of chaplains suited for the task of issuing documents and usually working under a cleric who was not the head of the chapel. Not all chaplains wrote documents, however, and the chapel and chancery thus remained separate institutions. From the reign of the emperor Louis I the Pious (814–840), the heads of the chancery were not personally involved in writing the documents, a task performed by unnamed and unknown scribes. At first the scribes were indiscriminately designated as either *notarii* or *cancellarii* (higher, Roman provincial officials of the 5th and 6th centuries, who stood at the barriers, *cancelli*, of the council rooms), but, by the 9th century, the title of *cancellarius* was gaining ground and was increasingly applied to the head of the chancery. The 9th century was a period of transition, during which, for a while, the archchaplain, the head of the chapel, became also the head of the clerks who wrote the charters.

The *capella* of St. Martin

The German imperial chancery. Under the Ottonian dynasty, which came to power in the eastern division of the original Carolingian empire early in the 10th century, the German royal chancery developed the organization that was to characterize it throughout the remainder of the Middle Ages. The heads of the chancery were the archchancellors, but the office was entirely honorary and soon came to be automatically held, as far as Germany was concerned, by whoever was archbishop of Mainz. When the German kings or emperors established administrations in Italy, Italian bishops were at first made archchancellors for Italy, but in 1031 the office was attached to the archbishopric of Cologne. From the 11th century, Burgundian bishops were archchancellors for Burgundy, but, in the second half of the 13th century, the archbishop of Trier took over the office.

The actual heads of the chancery were the chancellors. At first there was a chancellor, as well as an archchancellor, for each separate part of the empire—Germany, Italy, and Burgundy—but from 1118 there was only one chancellor for all three kingdoms. But even the chancellors, all of whom were clerics, were rarely involved in the actual composition and engrossing of documents, being usually engaged, as important advisers to the king or emperor, in much weightier matters. They do seem to have been especially concerned, however, with decisions about the granting of charters, and they supervised the work of the scribes or notaries. From among the ranks of these notaries, a group of protonotaries gradually developed after the mid-12th century, as a result of influence from the chancery of the Norman rulers of Sicily. Often called upon to deputize for the chancellor, the protonotaries, from the late 13th century onward, frequently titled themselves vice chancellors.

Imperial chancellors

From the 12th century onward, the documents issued by the German royal chancery were divided into various classifications. The diploma, by then usually called a privilege, existed in two categories, the solemn and the simple privilege. A solemn privilege included the *invocatio*, the signum and recognition line, and a detailed dating or at least one of these three elements, which were entirely lacking in simple privileges. Gradually, simple privileges merged into documents called mandates; it is not always easy to distinguish between them, but, in general, privileges were concerned with rights in perpetuity, while the mandates dealt mainly with matters of only temporary importance. From the early 14th century, mandates

were superseded by the use of letters patent and letters close (open or closed letters). Privileges continued to be sealed with a hanging seal; the seal on letters patent was impressed on the document and was used to seal up letters close.

As the power of the German kings declined during the later Middle Ages, so that of the archchancellors increased, and in the 14th century they attempted to win control of the chancery. But, despite fluctuations in the power struggle, the king retained control of the chancellor, who, by the end of the 15th century, held the title of imperial vice chancellor.

The French chancery. Under the Carolingians and the first Capetians in France, various bishops and archbishops, especially the archbishops of Reims, held the office of royal chancellor. But at that time the office was merely titular, and, by the end of the 11th century, it disappeared entirely. From the 12th century onward, the title of chancellor became reserved to the head of the chancery. These new chancellors became so powerful that in 1185 King Philip II Augustus left the office vacant, and, during almost the whole of the 13th century, the chancery was administered by subordinate officials. Chancellors, often laymen, were appointed again in the 14th century, however, and the office remained important until 1789. As in other parts of Europe, the French chancellor merely directed the work of the notaries, and it was they who were responsible for drawing up the documents. From 1350 onward, the notaries were called secretaries, and both their numbers and their importance steadily increased. From the 15th century, the tremendous expansion of business occupying the Grande Chancellerie led to the establishment of several subsidiary *petites chancelleries*, all issuing royal documents sealed with the king's signet. Until the reign of Henry I (1031–60), the old Frankish type of diploma was issued almost exclusively. Then, gradually, charters in the simpler form of letters began to replace the diplomas, and, during the 13th century, the *lettres patentes* became the common type of document. These lacked the *invocatio*, the monogram, and the signature of the high dignitaries, and they gave the simple form of dating. From the 14th century, two forms of *lettres patentes* existed: the *charte*, which was sealed with a green wax seal hanging on red and green silk cords; and the *lettre patente*, used mainly for administrative mandates, which was sealed with a yellow wax seal on double and single cord. Besides, *lettres closes* were used from the 13th century onward.

The English chancery. The English royal documents of the Anglo-Saxon period (before 1066) can be divided into two large groups: the charters, mostly written in Latin; and the writs, written in Old English. The charters, for the most part concerning grants of land, began with either a verbal or a symbolic *invocatio* (a cross or the monogram XP for Christ). There was an *arenga* but no *intitulatio*. Charters were not sealed, the validation comprising the nonautographed signatures of the king and of ecclesiastic and secular dignitaries. Many of the extant charters of this era are forged or interpolated; the series of those apparently genuine starts shortly after the arrival in Canterbury (669) of the archbishop Theodore, and these show similarities with late-Roman private documents. It therefore seems probable that the Anglo-Saxon charters derive from Italian models brought to England by the Roman missionaries. Most of the charters were apparently written by the recipients. From the 11th century, the charters were gradually superseded by sealed writs, which became the most important type of document in medieval England. At first written in Anglo-Saxon, they were produced in Latin soon after the Norman Conquest. No continental document was anything like them. Written on a narrow strip of parchment, their entire text occupied only a few lines. A symbolic *invocatio* in the form of a cross was followed by the king's name and the address, which contained a salutation clause. There was no *arenga*, and the address was followed by a short description of the bequest, if the writ concerned a grant, or a directive, if the writ was a mandate. The Anglo-Saxon writs had no signatures of witnesses and no date, but after 1066 these elements were added. The dating consisted at first only of a mention of the place of issue, days and month being usually given only toward the end of the 12th century. From the very beginning the writs were supplied with a pendant seal as a means of validation. From the reign of Henry II (1154–89), it is possible to distinguish mandate-like writs of the old type and charter writs, which mainly concerned questions of feudal enfeoffments and confirmations of privileges and were usually more carefully executed than the mandate writs. By the early 13th century, the charter writs had developed into a new form of charter. It contained the *intitulatio*, including all the titles of the king; the general address; the dispositive text, which in most instances was introduced with the expression *sciatis* ("know that"); and the final clauses, consisting of the names of several witnesses and of the *datum per manus* clause mentioning the chancellor and giving the date and place of issue. During the final stage of its development, the great seal of green wax, pendant on silk cords, served as the means of validation. The ordinary writs further evolved into the common-law writs (containing orders to persons mentioned by name), the writs of summons, and, particularly, the letters patent (the latter eventually assuming the function of the charter writs, which disappeared at the end of the 13th century) and the letters close. The letters patent were furnished with a general address. Often introduced with the *sciatis*, the text concerns either limited grants or commissions to royal officials, introduced by the word *precipio* ("I order") or a similar expression. The dating begins with a series of witnesses and contains the place of issue, day, month, and year of the king's reign. Occasionally the dating is followed by the words "per N.," which are assumed to designate the household official transmitting to the scribe the royal order to issue the writ. The means of validation was the great seal in white wax, pendant on a single strip of parchment (*simple-queue*) or on a double strip of parchment or silk (*double-queue*). The letters close obtained their name from the fact that they were closed by means of the great seal. They contained either commands or information directed to a single individual or to several persons. Characteristic of these letters are the words *teste me ipso* ("witnessed myself") introducing the regular dating clause. These types of royal documents remained essentially unchanged throughout the Middle Ages and were imitated by both secular and ecclesiastic English magnates and dignitaries.

English writs

The English royal chancery grew out of the royal household. As on the Continent, it was at first merely a group of royal chaplains, and, until the Norman Conquest, there was no chancellor at their head. The chancellor was the keeper of the seal but usually took no part in the issuing of documents. During the 12th century the number of scribes was still low, between two and eight. At first this "chancery" travelled about with the king, only in the course of the 13th century establishing a permanent location at Westminster. At first English royal documents were not dispositive but merely evidentiary, confirming a previously arranged, orally discharged legal act, and the king's order for issuing the document was given orally. But, from the late 13th century, a royal secretariat came into existence, which was in charge of relaying to the chancellor, by warrant sealed with the privy seal, the royal order for issuing a writ under the great seal. In many ways the secretariat thus became a competitor of the chancery. In addition, during the first half of the 14th century, the Signet Office was established, so called after the small seal (signet). The king's secretary was also the head of this office. All these shifts made the issuing of royal documents increasingly complicated. From the end of the 14th century, the common procedure involved, first, the petitioner submitting a petition to the king. If the king approved of it, his secretary forwarded a warrant carrying the signet to the keeper of the privy seal with the request to send, in his turn, a warrant carrying the privy seal to the chancellor. The chancellor then ordered the issue of the document, which would bear the great seal. So far as the issuing of royal documents is con-

The Signet Office

cerned, the fact that the secretariat developed into the secretariat of state is of great significance. The king's secretary (later on, there were two of them) became the centre of the royal administration. It was in his office that the state papers originated, which already under Henry VIII (died 1547) far surpassed in importance the old chancery records. Among the state papers there are in-letters, out-letters, drafts, reports, and schedules. The decline of the rolls (document registers) during the 16th century gave rise to yet another new office, the State Paper Office, headed since 1578 by the clerk of the papers. The second holder of this office, Sir Thomas Wilson, established the division of the state papers into foreign and domestic. As departments of state proliferated during the 18th and 19th centuries, they developed their own archives. In 1838 all the public archives became subject to an official called the master of the rolls.

Other European chanceries. Smaller European nations usually modelled their documents on those of the papacy, the empire, France, or England. The influence of papal letters and privileges can be observed particularly in Aragon, Castile, and Portugal, while German royal diplomas served as models in Bohemia (which was part of the empire), Hungary, and Poland. Because of the close political ties between the two kingdoms, Anglo-Saxon influence can be traced in the seal of the royal Danish documents during the 11th century, but, in the course of the 12th century, royal and princely German documents became the models for Danish as well as Swedish royal documents. Norwegian royal documents were modelled on Anglo-Saxon writs, probably as a result of the influence of English missionaries working in Norway from the early 11th century. The Norwegian writs were drawn up in the vernacular. The chancery of the Norman rulers of southern Italy and Sicily was highly developed. Influenced by the form of papal documents, the Norman documents comprise mainly privileges, either formal (with *rota*, witnesses, and gold bull) or simple (with *rota*, leaden bull, or wax seal), and mandates. They all have a detailed dateline that includes the name of the chancellor or other high court officials, the number of years since the birth of Christ, the regnal year of the king, and the *apprecatio*. The mandates are more simply executed. They lack the *invocatio* and start out with a simple *intitulatio* and *inscriptio* that ends with a salutation clause. There is neither *arenga* nor *corroboratio*, but there is a command clause in the text. The dating consists only of the place of issue, the day of the month, and the *indictio*. There is no *rota* or signature. The seal is of red wax. The head of the royal chancery of the Norman kingdom of Sicily was the chancellor, a layman who was an influential court official. The notaries who drafted and wrote the documents were also laymen. Because the German Hohenstaufen emperors also ruled in Sicily from 1194 to 1250, Norman chancery practice influenced subsequent German documents.

The chancery of Norman rulers

BIBLIOGRAPHY. HARRY BRESSLAU, *Handbuch der Urkundenlehre für Deutschland und Italien*, 2nd ed., vol. 1 (1912), vol. 2, pt. 1 (1914), vol. 2, pt. 2, ed. by H.W. KLEWITZ (1931, all reprinted 1958, with separate index), although somewhat out of date, still the best handbook by far for Germany and Italy; OSWALD REDLICH, *Die Privaturkunden des Mittelalters* (1911, reprinted 1967), an excellent work on the private documents of the Middle Ages; LEO SANTIFALLER, *Beiträge zur Geschichte der Beschreibstoffe im Mittelalter*, pt. 1 (1953), the most up-to-date account of the history of writing materials during the Middle Ages; FRANZ DOLGER and JOHANNES KARAYANNOPULOS, *Byzantinische Urkundenlehre die Kaiserurkunden* (1968), the only study of the documents of the Byzantine emperors; C.R. CHENEY, *The Study of the Medieval Papal Chancery* (1966), an excellent general survey of modern research; R.L. POOLE, *Lectures on the History of the Papal Chancery down to the Time of Innocent III* (1915), the best book in English on this subject, though now out of date; PETER HERDE, *Beiträge zum päpstlichen Kanzlei- und Urkundenwesen im 13. Jahrhundert*, 2nd ed. (1967); and *Audientia litterarum contradictarum*, 2 vol. (1970), on papal letters of justice; WILHELM ERBEN, *Die Kaiser- und Königsurkunden des Mittelalters in Deutschland, Frankreich und Italien* (1907, reprinted 1967), an excellent supplement to Bresslau on the imperial and royal documents of the Middle Ages in Germany, France, and Italy; GEORGES TESSIER, *Diplomatique royale française* (1962), the best and most up-to-date handbook on the royal French diplomatic; ALAIN DE BOUARD, *Manuel de diplomatique française et pontificale* (1929), a handbook on French and papal diplomatic; F.M. STENTON, *The Latin Charters of the Anglo-Saxon Period* (1955), a good brief survey on research since the 18th century; P.H. SAWYER, *Anglo-Saxon Charters* (1968), the most up-to-date annotated list and bibliography; R.C. VAN CAENEGEM, *Royal Writs in England from the Conquest to Glanvill* (1959), the best book on writs; V.H. GALBRAITH, *An Introduction to the Use of the Public Records* (1934) and *Studies in the Public Records* (1948), two works especially useful for the later periods; C.R. CHENEY, *The Records of Medieval England* (1956) and *English Bishops' Chanceries 1100–1250* (1950); T.F. TOUT, *Chapters in the Administrative History of Medieval England*, 6 vol. (1920–33), the basic work on the subject, including the chancery; HARRY BRESSLAU, "Internationale Beziehungen im Urkundenwesen des Mittelalters," *Archiv für Urkundenforschung*, 6:19–76 (1918), the only survey on international relations in the documentary system of the Middle Ages; HORST ENZENSBERGER, *Beiträge zum Kanzlei- und Urkundenwesen der normannischen Herrscher Unteritaliens und Siziliens* (1971), the most up-to-date account of the chancery and documents of the Norman rulers in southern Italy and Sicily; H.O. MEISNER, *Archivalienkunde vom 16. Jahrhundert bis 1918* (1969), the best study of records from the 16th century through 1918 for central Europe, especially Germany; CHARLES CARTER, *The Western European Powers, 1500–1700* (1971), a very useful survey on archives and their records. All the cited works have comprehensive bibliographies, with some of them also containing bibliographies of editions and facsimiles of documents.

(Pe.He.)

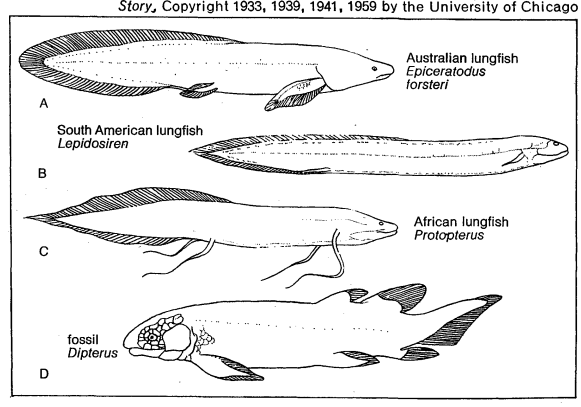
Dipnoi

The Dipnoi comprise an order of fishes that first appeared in the Lower Devonian Period (about 370,000,000–395,000,000 years ago) and that today is represented by only six species. Some authorities recognize only three extant species. Known as lungfishes, the extant species occur in Africa, South America, and Australia. They are especially interesting because of their characteristic body forms, their generally large size, their erratic distribution over the tropical regions of the earth, and their peculiar mode of life.

Economic importance

The economic importance of the lungfishes is slight. Only in certain parts of Africa, because of their abundance and size, are they of any value to man as food. They are obtained from the mud of dried river bottoms. The South American lungfish, which is obtained in the same manner, is eaten locally.

Drawing by J. Helmer; from (B,C) J.R. Norman, *A History of Fishes* (1947); Hill & Wang Publishers, (D) A.S. Romer, *The Vertebrate Story*, Copyright 1933, 1939, 1941, 1959 by the University of Chicago



Living and fossil forms of Dipnoi fishes.

Size range and distribution. Most species grow to substantial size. The Australian lungfish, *Neoceratodus forsteri*, attains weights of up to ten kilograms (about 22 pounds) and a length of 1.25 metres (about 50 inches). Of the African lungfishes, the yellow marbled Ethiopian species, *Protopterus aethiopicus*, is the largest, growing to a length of two metres (about 80 inches). The South American species, *Lepidosiren paradoxa*, reaches a length of 1.25 metres (about 50 inches).

The distribution of the Dipnoi strikingly parallels that of the unrelated osteoglossomorph fishes, another freshwater group. The Australian lungfish occurs in a very small region of Australia: in the marshes of Queensland, along Burnett River and St. Mary's River. Four species (*Protopterus*) occur in Africa, where they are chiefly concentrated in the equatorial belt, but occur as far north as Senegal and as far south as Mozambique. Within their areas of distribution, the African protopterids are abundant along the riverbanks, in submerged areas with plant cover, and in lakes. *Lepidosiren paradoxa*, the South American lungfish, is widely distributed in that continent. It is especially numerous and often associated with the eel *Synbranchus marmoratus* in the shallow and muddy watercourses of the Chaco River in Paraguay and in neighbouring areas.

Natural history. *Reproduction and life cycle.* The African lungfishes spawn in the last half of winter, coincident with the onset of the rainy season. *Protopterus* species build a nest in the form of a pit on the bottom of a watercourse. The egg is about 3.5 to four millimetres (about 0.14 inch) in diameter, and the tiny larvae emerge a week after the eggs are laid. The larvae have long, bright red tuftlike or fanlike external gills, which they use for breathing until the lungs are fully developed. The brood of young at first remain in the nest under the protection of the male.

The South American lungfishes dig a nest in the bottom in the form of a vertical passage, which frequently turns horizontally at the bottom. The male remains in the nest and guards the brood. During the spawning season, the pelvic fins of the male develop numerous tuft-shaped growths filled with small blood vessels (capillaries). These growths are believed to release oxygen from the blood, thereby oxygenating the water around the young.

The Australian lungfish lays gelatinous eggs among waterplants; the larvae, which have no external gills, breathe through internal gills.

Behaviour and ecology. Lungfishes are voracious, eating a variety of aquatic animals, including members of their own species. In captivity, African lungfishes eat earthworms, pieces of meat, tadpoles, small frogs, and small fish. The Ethiopian lungfish has at the front of the upper jaw two rather rounded teeth with a hard transverse (from side to side) bridge. The lower jaw has a number of crushing teeth. The prey is sucked in, crushed, and thoroughly chewed; such a manner of eating is rare among fishes.

Form and function. *General features.* The slim, eellike African protopterids and the even slimmer South American *Lepidosiren paradoxa* have long stringy, very mobile pectoral and pelvic fins that are in a constant state of agitation—touching and sensing surroundings. The tips of these fins have a highly developed sense of touch, which, together with the fish's well-developed sensitivity to pressure and turbulence and its good sense of smell and taste, largely make up for the weakness of the eyes. The fish are almost blind with respect to the perception of form and movement. Pressure and turbulence are sensed by means of sensory structures called lateral lines. At the anterior, or head, end, the lateral lines are modified into a pattern of intricately interlaced bright lines, which are a series of tiny bud-shaped terminal organs. The highly individual patterns are used in distinguishing species.

Fins as sense organs

The Australian lungfish has an entirely different appearance. It is more compactly built and has large, overlapping scales. The pectoral and pelvic fins are much broader. The African and South American lungfishes have paired lung sacs; in the Australian species the left lung sac atrophies.

Adaptations for breathing. There are a number of fishes that, in addition to or in place of gill breathing, have developed special organs through which they can breathe atmospheric air at the water surface. This phenomenon occurs almost exclusively in freshwater fishes. In lungfishes these organs for the intake of atmospheric air are, both in function and in structure, primitive lungs like those of amphibians. The name lungfish is thus well applied: these fishes have sac-shaped, pneumatic organs that lie along the alimentary tract. The inner surfaces of these air-breathing organs are covered with a great number of honeycomb-like cavities covered with fine blood vessels. As in terrestrial higher vertebrates, gas exchange takes place in the tiny air vesicles.

In order to breathe, the fish swims upward and positions its head so that the tip of the snout barely touches the water surface. The mouth is then opened wide, and the fish sucks in air from just above the water—a process often accompanied by a characteristic sound. The Australian lungfish reportedly breathes air through the nasal openings, the mouth remaining closed. In contrast to the higher bony fishes, lungfishes have a particular opening (choana) that connects the nasal cavity with the mouth.

In the Australian lungfish, gill breathing predominates at least some of the time—namely, in times of normal water level when the water is well oxygenated. At such times the fish rises less often to the surface to breathe atmospheric air. When the water level goes down, which usually occurs in August or September, the fish is often found in isolated waterholes in which the oxygen content is greatly reduced. Other fishes in such pools often die from lack of oxygen, but the lungfish survives, having changed over to the breathing of atmospheric air. During such a dry period the Australian lungfish surfaces about every 40 to 50 minutes for air. African lungfishes surface for air about every 30 minutes or, in some cases, at longer intervals.

Physiology and biochemistry. African lungfishes bore into the bottom of a riverbed or lake bed for their "dry sleep." After burying themselves they become encased in a sheath that gradually hardens. Here they spend the dry season, during which the waterline becomes lower and the riverbed or lake bed finally dries out. The African lungfish generally digs in and encysts in this manner, even if there is sufficient time to swim to deeper waters. African lungfishes also burrow into mud and ensheath themselves under experimental conditions. They have been kept alive in such an induced state for more than two years.

The "dry sleep"

The South American lungfish also bores into the mud in times of water shortage, but it forms no protective sheath. The Australian lungfish never buries itself in this manner.

Studies have shown that the "dry sleep" of the African lungfish is induced by a substance that inhibits the fish's normal metabolism (*i.e.*, together, all the internal physiological activities).

Extracts from the brains of such sleeping fish injected into rats have caused them to become lethargic; in addition, the body temperature of the rats falls 5° C, and the metabolic rate falls 33 percent. The day after receiving such injections, the rats stop eating. It is believed that the substance responsible for this effect is a protein-like substance.

Evolution. The oldest Dipnoi, from the Lower Devonian Period, had skull and dental features that are characteristically dipnoid but also had many similarities to the crossopterygians (*e.g.*, coelacanth). The Dipnoi was abundant until Triassic times (190,000,000–225,000,000 years ago), after which their numbers decreased.

Dipterus, one of the oldest lungfish, had leaflike pectoral and pelvic fins similar to those of the modern Australian lungfish, and it seems reasonable to assume that early forms also had functional lungs comparable with those of species living today. Hardened sections of clay, cylindrical in shape, have been found in Pennsylvanian (about 280,000,000–325,000,000 years old) and Permian (225,000,000–280,000,000 years old) deposits. Remains of the dipnoid *Gnathorhiza*, closely allied to the extant African and South American species, were imbedded in the clay, strongly suggesting that they passed unfavourable conditions buried in mud.

An evolutionary line can be traced from *Dipterus* to *Neoceratodus*, the extant Australian species. *Scaumenacia* and *Phaneropleuron*, common forms of the Upper Devonian (345,000,000–370,000,000 years ago), exhibited a much-reduced first dorsal fin (the first fin forward on the back); the second dorsal fin was enlarged and had

shifted further toward the tail. Lungfish of Permian times showed an apparent fusion of the fins along the back and the rest of the vertical midline into the so-called diphycercal tail (*i.e.*, tapering to a point) present in modern lungfishes. Various side branches also occurred in the evolution of the Dipnoi, none of which has survived to modern times.

Classification. The annotated classification given below primarily relates to living forms; for a classification including the extinct forms see the critical appraisal below.

Distinguishing taxonomic features. The separation of Dipnoi as a discrete group is based largely on the structure and arrangement of the skull bones and the teeth. The suborders, of which there are two, are mutually distinguishable mainly by the number of lungs (one or two).

Annotated classification.

SUBCLASS DIPNOI

Lower Devonian (370,000,000–395,000,000 years ago) to Recent. Cranium not divided into movable parts; teeth in upper jaw reduced and lost in later members; some teeth fused into plates for eating shellfish. A single order.

Order Sirenoidei

Suborder Monopneuma

One functional lung.

Family Ceratidae. Pectoral and pelvic fins reduced but not tentacle-like; scales large; grows to about 1.25 metres (about 50 inches); 1 living species: *Neoceratodus forsteri* (Australian lungfish).

Suborder Dipneuma

Two functional lungs.

Family Lepidosirenidae. Body eel-like in form; scales small, pectoral and pelvic fins modified into slender tentacle-like structures; passes dry periods in mud of dried river and lake bottoms; grows to about 2 metres (about 80 inches); 2 living genera: *Lepidosiren* of South America (1 species: *L. paradoxa*) and *Protopterus* of Africa (4 species).

Critical appraisal. Some writers assign Dipnoi to the ordinal level, subsuming several families, mostly extinct, within that order.

The following alternate classification is according to A.S. Romer (1966), an American vertebrate paleontologist (extinct families represented only by fossils are indicated by a dagger [†]):

ORDER DIPNOI

†Family Dipnorhynchidae

Lower to Middle Devonian; Europe, Australia, North America.

†Family Dipteridae

Devonian; Europe, Greenland, North America, northern Asia, Australia.

†Family Ctenodontidae

Carboniferous (280,000,000–345,000,000 years ago); Europe, North America, Australia.

†Family Phaneropleuridae

Upper Devonian; North America, Greenland, Europe.

†Family Sagenodontidae

Mississippian to Lower Permian (250,000,000–280,000,000 years ago); North America, Europe.

†Family Uronemidae

Mississippian; Europe, North America (?).

†Family Conchopomidae

Pennsylvanian (280,000,000–325,000,000 years ago) to Lower Permian (250,000,000–280,000,000 years ago); North America, Europe.

Family Ceratodontidae

Lower Triassic (210,000,000–225,000,000 years ago) to Recent; one surviving species, *Neoceratodus forsteri.*

Family Lepidosirenidae

Pennsylvanian to Recent; 2 living genera, *Lepidosiren* and *Protopterus.*

BIBLIOGRAPHY. H. SWAN, D. JENKINS, and K. KNOX, "Metabolic Torpor in *Protopterus aethiopicus:* An Anti-Metabolic Agent from the Brain," *Am. Nat.*, 103:247–258 (1969), an article on dry sleep in lungfishes; M. BLANC, F. D'AUBENTON, and Y. PLESSIS, "Mission M. Blanc-F. d'Aubenton (1954) IV. Étude de l'enkystement de *Protopterus annectens* (Owen 1839)," *Bull. Inst. Fr. Afr. Noire*, Series A, 18:843–854 (1956), a study of the encystment of *Protopterus annectens;* P. BRIEN, M. POLL, and J. BOUILLON, "Ethologie de la reproduction du *Protopterus dolloi* (Boulenger)," *15th Int. Congr. Zool.*, sect. 1 (1959); J.S. BUDGETT, "On the Breeding-Habits of Some West-African Fishes, with an Account of the External Features in the Development of *Protopterus annectens,* and a Description of the Larva of *Polypterus lapradei,*" *Trans. Zool. Soc. Lond.*, 16:115–136 (1901); K. CURRY-LINDAHL, "On the Ecology, Feeding Behaviour and Territoriality of the African Lungfish, *Protopterus aethiopicus,* Heckel," *Ark. Zool.*, Series 2, 9:479–497 (1956); A.G. JOHNELS and G.S.O. SVENSSON, "On the Biology of *Protopterus annectens* (Owen)," *ibid.*, 7:131–164 (1955), a detailed study of the habits of lungfishes in the flooding zones on both sides of the Gambia River; K.H. LÜLING, "Einige Notisen über afrikanische Lungenfische," *Dt. Aquar.-Terrar-Z.*, 12:12–14, 44–46 (1959), on the habits and distribution of the African lungfishes, together with a distribution map according to Poll; "Untersuchungen an Lungenfischen, insbesondere an afrikanischen Protopteriden," *Bonn. Zool. Beitr.*, 12:87–112 (1961), a detailed examination of the experimental encysting of the West African lungfish *Protopterus dolloi* in captivity; "Fische mit Lungen," *Neptun*, 6:80–83 (1966), a study of the morphology, anatomy, and the method of breathing of lunglike structures in the dipnoi; M. POLL, "Zoogéographie des protoptères et des polyptères," *Bull. Soc. Zool. Fr.*, 79:282–289 (1955), a discussion of the distribution of the four species of African lungfishes, with distribution map; H.W. SMITH, "Metabolism of the Lung-Fish, *Protopterus aethiopicus,*" *J. Biol. Chem.*, 88:97–130 (1930), the first modern physiological study of the encysting of Ethiopian lungfishes in captivity; "Observations on the African Lung-Fish, *Protopterus aethiopicus,* and on Evolution from Water to Land Environments," *Ecology*, 12:164–181 (1931); E.K. SUVOROV, *Allgemeine Fischkunde* (1959; German trans. from the 2nd Russian ed. of 1948), includes a chapter on the breathing organs of Dipnoi.

(K.H.L.)

Dipsacales

The Dipsacales, or teasel order, is a small order of four families of flowering plants including about 1,100 species, chiefly herbs or shrubs and, rarely, small trees or climbers. Though distributed throughout the world, they are centred mainly in the temperate parts of the Northern Hemisphere, where they occur in many types of habitats. The order is primarily important for the large number of hardy ornamentals it contains; for example, those of the genera *Lonicera* (honeysuckles), *Viburnum* (guelder rose, arrowwoods, and wayfaring trees), and *Scabiosa* (scabious). A few species are noxious weeds.

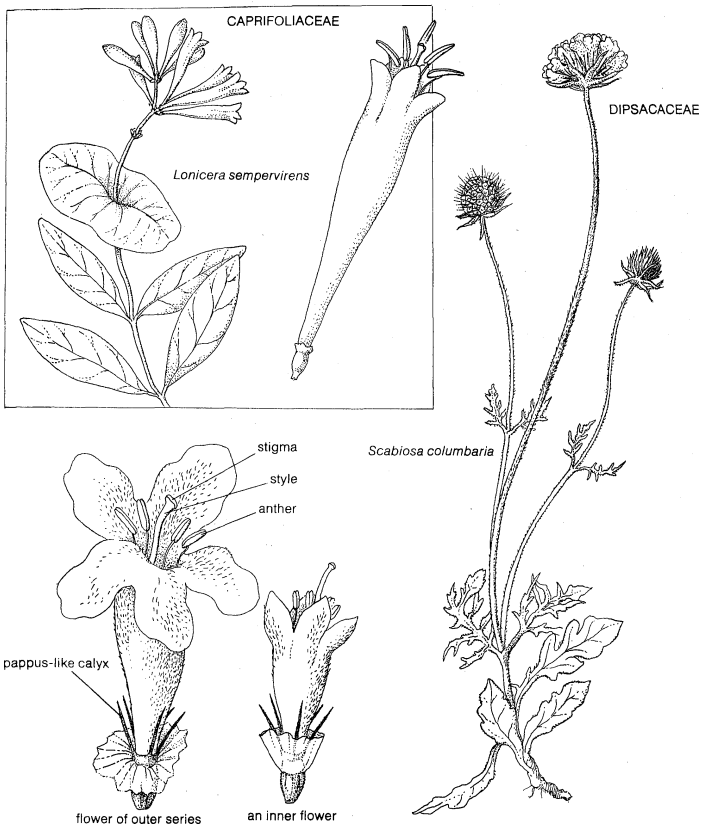
GENERAL FEATURES

Size range and diversity of structure. The order Dipsacales, although predominantly herbs or small shrubs, includes quite diverse life forms ranging from small desert annuals to trees. It occurs in habitats from the Arctic to the tropics. Notable in the order are several species of the genus *Lonicera* that are characterized by their climbing and sometimes strangling habit, and some species of *Valeriana* that have large tap roots and a cushion (low, densely branched, matlike) growth form. There are, however, no succulents or parasitic plants in the order, and none that are normally epiphytic, (*i.e.*, not rooted in soil but living among the branches of trees).

Growth forms

Utility and importance. The Caprifoliaceae family contains many valuable and beautiful ornamentals, including many species of *Viburnum, Lonicera, Weigela, Abelia, Symphoricarpos, Leycesteria,* and *Sambucus.* The hard wood of *Sambucus nigra* (elder) has been used in making musical instruments, and the large pith has been used for cleaning the pivots of delicate machinery. The fruits and flowers of this and other species of *Sambucus* are used for making jellies and wines. *Triosteum perfoliatum* (feverwort) roots are used as a mild cathartic. The dried and roasted fruits of this and other species have been used as a coffee substitute. *Lonicera japonica* (Japanese honeysuckle) introduced from Asia, is, in eastern North America, a pernicious and dangerous pest in woodlands, on roadbanks, and in hedgerows, overwhelming and strangling native flora. In China the leaves serve as a tea substitute, and in Japan and Vietnam extracts of the flowers and leaves are used as a diuretic.

In the family Valerianaceae there are no plants of great economic importance; a number are grown as orna-



Vegetative, floral, and fruiting structures of the order Dipsacales.
Drawing by M. Pahl

mentals as, for example, *Centranthus ruber* (red valerian), *Fedia eriocarpa* (African valerian), and *Valeriana officinalis* (common valerian or garden heliotrope). This latter species has been used as a source of the drug valerian sometimes used in the past in the treatment of functional nervous disorders and some cardiac ailments. Corn salad (*Valerianella locusta*), as its name implies, is used in salads. *Nardostachys jatamansi* (spikenard) is a native of high altitudes in Nepal, Bhutan, and other parts of the Himalayas extending into western Asia. The roots and spikelike woolly young stems, before the leaves unfold, are dried and yield an oil, which is used in perfumery. The spikenard referred to in the Bible is probably this plant. It was a popular perfume in Roman times and clearly costly, as the biblical references imply, due, no doubt, to the long distance it had to be imported from India. It is still used as a perfume in India, but the taste in perfume in western civilization has changed and spikenard might be considered disagreeable by some today.

The teasel

The teasel family, Dipsacaceae, has a number of genera of horticultural interest, and *Scabiosa* in particular is widely cultivated for its ornamental value, especially *S. caucasica*. *Cephalaria syriaca* can be a harmful weed in parts of the Middle East, the fruits contaminating grain and tainting flour. *Dipsacus*, the teasel, and the genus from which the order Dipsacales derives its name, consists of about 12 species in Europe, western Asia, and Africa. The name is from the Greek *dipsa*, "thirst," alluding to some species that have united, cup-shaped bases of the leaves that hold water. The genus contains the only member of economic importance, *Dipsacus sativus*, the fuller's teasel, so called because the ripened inflorescences (flowering heads) are used for fulling cloth. The heads are fixed to a cylinder that revolves slowly over the cloth while the plant's flexible receptacular bracts, which end in a stiff recurved spine, raise the nap or pile. The more appropriately named *D. fullonum* is very closely related, having erect receptacular bracts that are, however, too flexible for combing cloth. The fuller's teasel was widely cultivated in Europe and North America but the introduction of "metal teasels" have to some extent replaced it. Some are still cultivated in parts of Europe and in the state of Oregon for fulling high-quality cloth.

Historically, the properties of the teasel as an instrument for fulling cloth were known in Roman times and records of its cultivation in Great Britain since Richard I (1199) are known. The word "teasel" occurs in pre-English Anglo-Saxon and is directly related to the word "tease" in the sense of the act of disentangling fibres.

NATURAL HISTORY

Seed dispersal. The order shows a wide variety of methods of seed dispersal. Fleshy berries and both fleshy and dry drupes (single stony-seeded fruits) occur in the family Caprifoliaceae, representing adaptations to bird dispersal. In the Valerianaceae family the fruits of some genera are provided with a plumed structure or are winged. In the Dipsacaceae family similar winged structures that occur are mechanisms or adaptations for wind dispersal. Some species in the family Dipsacaceae have bristles or burrs that cling to the coats of passing animals. Dry capsules, which split open explosively, and indehiscent (nonopening) fruits also occur in this family. Slugs are said to help the dispersal of the seeds of *Adoxa*.

Pollination. The flowers of the majority of genera in most of the families are generally adapted to pollination by insects. A few are reported to be pollinated by birds. There are no records of wind pollination.

The flowers manifest adaptations to insect pollination in a variety of ways. Many species have a fragrant or characteristic odour, and nectar is secreted in most genera of the order. Brightly coloured flowers are frequent and in some species are aggregated into a head sometimes with large, sterile ray flowers around the margin of the cluster, causing the whole inflorescence to resemble superficially a lone, large flower.

Adaptations to insect pollination

The flowers in the order are most commonly protandrous—that is, the male stamens mature and release pollen before the female stigma of the same flower becomes receptive—an adaptation that favours cross-pollination. There is, however, a large diversity of mechanisms for both self- and cross-pollination throughout the order. There is considerable sexual polymorphism (*i.e.*, the different sexes of flowers have different shapes and sizes) with frequent occurrence of gynodioecism, in which separate plants are of one sex, but some bisexual flowers occur on plants that normally bear mostly female flowers. Dioecism, male and female flowers occurring on different plants, is known, too, in at least one species, *Valeriana dioica* (Valerianaceae). *Sambucus* (Caprifoliaceae) has widely diverging stamens and extrorse anthers (*i.e.*, the anthers open away from the central axis and female parts of the flower) that render self-pollination difficult. In *Adoxa* protogyny, the condition in which the female parts mature before the male, is reported with the anthers at first remote from the stigmas and later inclined inward to allow self-pollination. Similar mechanisms to effect self-pollination, in the event of cross-pollination failing, are found in the genus *Lonicera* (Caprifoliaceae). Species of *Lonicera* with long corolla tubes are adapted to pollination by long-tongued insects, and in the western United States some species of this genus are pollinated by hummingbirds.

Self-pollination may occur in species of *Lonicera* that have pendulous flowers, and in some species in the order, especially *Valerianella*, automatic self-pollination results when the style elongates through the anthers during growth.

FORM AND FUNCTION

Anatomy. There are two general anatomical groups in the order, woody and herbaceous. The order has few distinctive anatomical characters but the occurrence of glandular hairs is almost a common feature.

The Caprifoliaceae family is predominantly woody and has characteristic xylem vessels (water-conducting tissues) with scalariform perforation plates (*i.e.*, the vessel end walls have ladderlike slotted openings), and its wood

fibres have bordered pits. The genera *Sambucus* and *Viburnum*, however, differ from the rest of the family in having nectar-secreting organs located outside the flower. *Sambucus* is further distinct in having cellular crystals and thick-walled, long, narrow fibres and a large pith. It also has highly specialized secondary xylem. The genera *Alseuosmia*, *Carlemannia*, and *Silvianthus*, however, lack the characteristic anatomical features of the family.

The family Valerianaceae is characterized by the absence of parenchyma (a type of cell) and the presence of oil cells in the cortex and cork of the roots. Woody members have very small vessels.

The vascular system forms a closed ring early in development in the Dipsacaceae family, and a broad pith is a feature of the family.

Pollen morphology. The pollen in the order is variable. The families Valerianaceae, Dipsacaceae, and some members of Caprifoliaceae, however, show an overall pattern of similarity, most of their species having three-furrowed pollen, often with a pore in the furrow, and the pollen grain surfaces have spiny ornamentation.

Pollen shapes and ornamentation

The most common type of pollen in the Caprifoliaceae family is somewhat flattened with a spiny wall supported by densely packed rods, or bacula. The genera *Sambucus* and *Viburnum* have distinct pollen with a reticulate pattern of ornamentation on the wall. The genus *Abelia* is remarkable in having two distinct pollen types, typical Caprifoliaceae pollen and triangular pollen without ornamentation. *Adoxa* (family Adoxaceae) has pollen with reticulate sculpturing similar to *Sambucus*.

Two main types of pollen occur in the family Dipsacaceae, one that occurs in most genera of the family, in which the apertures have a surrounding halo and are usually covered with a lid and the surface has little spines or wartlike pegs; and a second type, found in the genus *Morina*, in which the apertures have no halo and a lid is usually absent and the surface has no spinules. The pollen morphology of the family Valerianaceae is not well known but generally resembles that of the family Dipsacaceae.

Biochemistry. The biochemistry of the order has not been studied extensively and only isolated facts are available. Volatile oils occur in the family Valerianaceae, especially in the roots and rhizome (a rootlike stem structure). The same oils are found in the genera *Sambucus* and *Viburnum*, members of the family Caprifoliaceae. The family Caprifoliaceae also contains several rather widespread cyanogenetic glycosides (toxic substances consisting of carbohydrates such as sugars combined with hydrocyanic acid), and rutin is reported. Alkaloids are rare and of little importance. Saponines are reported in the genus *Lonicera*. Some fruits of the family, especially in the genus *Sambucus*, are rich in pectin. The family Dipsacaceae is notable for the occurrence of pseudoalkaloids and pseudoindikases.

The serological correspondence in seed proteins—the measurement of the relative amounts of the major protein constituents of plant fluids—has been under investigation in the Caprifoliaceae, and the results are expected to prove important in the classification of the family.

Chromosome number. The chromosome number of the order is not well known although some genera have been investigated. The basic numbers in the order appear to be 7, 8, 9, and 10. Changes in chromosome number appear to have played only a small part in the differentiation of genera and species in the family Caprifoliaceae. Polyploidy—*i.e.*, the existence of more than the basic two full sets of chromosomes—has played a major role, however, in speciation in some of the genera in the families Valerianaceae and Dipsacaceae.

EVOLUTION

Fossil record. The earliest remains assignable to the order appear to be leaf impressions described as *Viburnum vetus* found in Cretaceous deposits (about 100,000,000 years old) in Portugal. Fossil leaves of *Viburnum* are fairly widely reported from Upper Cretaceous deposits of the United States and Europe. This genus was exceedingly widespread and characteristic of Tertiary times (between about 7,000,000 and 65,000,000 years ago), and both pollen and leaves occur in many American, European, and Asiatic deposits. Fossil pollen of *Sambucus* is reported from the Eocene (about 50,000,000 years ago) in Europe and North America, but it is by no means common. Fossil pollen of many members of the order has been reported, however, from Pliocene deposits (about 5,000,000 years old). Fossil leaves of *Sambucus*, *Lonicera*, and *Symphoricarpos* have been recognized in Miocene (about 20,000,000 years ago) and Pliocene deposits of Europe, Asia, and North America. Seeds of *Sambucus* have been identified from the Pliocene in Europe, and fruits apparently assignable to the genera *Valeriana* and *Valerianella* have also been reported.

Phylogeny. The adaptive significance of most of the characters that distinguish the Dipsacales order from other related orders is doubtful. Within the order there is a tendency for a progression from woody to herbaceous habit, from regular to irregular corolla, from several fertile locules (ovary chambers) to one, from numerous seeds to one, and from endospermous seeds to nonendospermous (endosperm is a starchy nutrient tissue for developing embryos)—a trend that has taken place in other orders of flowering plants. The family Caprifoliaceae seems clearly to be the most primitive family of the order, from which all the others have been derived. Several genera of the tribe Lonicereae are transitional toward the family Valerianaceae. The family Dipsacaceae, although specialized, is linked to Valerianaceae by *Triplostegia*, a genus first placed in the latter family but now accepted as a primitive member of the Dipsacaceae. The family Adoxaceae is a specialized herbaceous offshoot of the Caprifoliaceae.

Evolutionary trends

CLASSIFICATION

Distinguishing taxonomic features. Dipsacales is an order occurring predominantly in the temperate parts of the world. It consists primarily of herbs or shrubs, characterized by opposite leaves that usually lack stipules. The flowers are gamopetalous (*i.e.*, they have fused petals) and vary from being regular to irregular; the ovary is inferior.

Annotated classification.

ORDER DIPSACALES

Herbs or shrubs, seldom small trees. Leaves in basal rosettes or opposite, very rarely alternate, usually lacking stipules. Flowers regular or irregular, bisexual or unisexual, rarely some sterile. Calyx (sepals) absent or poorly developed. An epicalyx sometimes well developed in fruit. Corolla sympetalous, 3- to 5-lobed. Stamens 1 to 5, usually 3 to 5, alternate with the corolla lobes; filaments attached to the corolla tube. Pollen trinucleate. Ovary syncarpous, inferior, 2- to 5-locular, locules with 1 to many anatropous ovules. Fruit various, endosperm present or absent, embryo straight. Four families, about 40 genera and 1,100 species, with worldwide distribution but more abundant in Northern Hemisphere temperate regions.

Family Caprifoliaceae

Deciduous or evergreen shrubs, sometimes small trees or vines, rarely herbs, with opposite or alternate, entire or divided, exstipulate or stipulate leaves. Inflorescence a corymb, cyme, thyrse, or spike of whorls, or flowers solitary or in pairs. Flowers regular or irregular, epigynous, bisexual, rarely some sterile. Floral tube adnate to the ovary, with a distinct constriction at the level of the 3 to 5 usually small calyx lobes. Corolla of fused petals, 3- to 5-lobed, salverform, rotate, campanulate, funnelform, or tubular, often bilabiate, sometimes gibbous at or near the base. Stamens 5 (sometimes 4), epipetalous, alternate with the corolla lobes; anthers 2-locular, opening longitudinally, versatile, dorsifixed, introrse or rarely extrorse; pollen tricolpate. Gynoecium of 1 to 5 fused carpels; style 1 or wanting; stigmas as many as the carpels, distinct or united; ovary inferior, 2- to 5-locular; locules with solitary to numerous, pendulous, anatropous ovules; placentation axile or parietal. Fruit a berry or drupe, rarely a capsule or achene, with 1 to many seeds or as many seeds as locules. Endosperm copious; embryo straight, small. About 18 genera and 500 species, distributed primarily in the North Temperate Zone, especially in China and the Himalayas, but extending to the mountains of the tropics, South America, and Australasia.

Family Adoxaceae

A glabrous, perennial, rhizomatous herb. Rhizomes creeping

with fleshy white scales at the apex. Basal leaves 5-foliate or ternate, long petiolate, stem leaves ternate, exstipulate; leaflets divided into 2 or 3 stalked lobes. Inflorescence a capitulum-like cyme. Flowers regular, bisexual. Calyx 2- or 3-lobed. Corolla 4- or 5-lobed, imbricate, light green. Stamens 4 or 5, epipetalous, alternate with the corolla lobes, divided to the base and appearing as 8 or 10 each with 1 anther sac; anthers versatile, 2-locular, opening longitudinally; pollen tricolpate, reticulate. Gynoecium syncarpous, ovary semi-inferior, 3- to 5-locular with a solitary, pendulous, anatropous ovule; styles 3 to 5, stigmas capitate. Fruit a drupe with 3 to 5 pyrenes or nutlets. Seeds 3 to 5; endosperm copious; embryo small. One genus and one species (*Adoxa moschatellina*), occurring in moist, shady places throughout the North Temperate Zone.

Family Valerianaceae

Annual or perennial herbs, sometimes woody at base. Leaves in basal rosettes or opposite, pinnately divided or entire, exstipulate, the bases often sheathing. Inflorescence a monochasium, thyrse, or many-flowered compound dichasial cyme, sometimes condensed and capitate. Flowers irregular or almost regular, bisexual or unisexual. Calyx obsolete or developing late and becoming conspicuous only in fruit. Corolla tubular, 3-, 4-, or 5-lobed, imbricate, often basally spurred, or saccate or bilabiate. Stamens epipetalous and alternate with the corolla lobes, varying in number from 1 to 4; anthers versatile, 2- or 4-lobed, 4-locular, introrse, opening longitudinally; pollen tricolpate, echinate. Gynoecium syncarpous, ovary inferior, 3-locular, with 2 locules usually suppressed and 1 fertile, with a solitary, pendulous, anatropous ovule; style 1, stigma simple or lobed. Fruit dry, indehiscent, the calyx often developing into a winged, awned, or plumose pappus. Seed 1; endosperm absent; embryo large, straight. About 10 genera and 370 to 400 species distributed widely in the North Temperate Zone and well represented in the Andes Mountains of South America.

Family Dipsacaceae

Annual or perennial herbs, rarely small shrubs. Leaves opposite or rarely whorled, simple or compound, exstipulate. Inflorescence usually capitate or rarely a spike of false whorls or a panicle, bracts forming a calyx-like involucre. Flowers irregular or regular, bisexual with a calyx-like involucel of fused bracteoles borne in the axils of imbricate receptacular bracts. Calyx small, cuplike or tubular or divided into 5 to 10 pappus-like segments. Corolla gamopetalous, 4- to 5-lobed, imbricate. Stamens usually 4, sometimes 2, 3, or 5, epipetalous, alternate with the corolla lobes; anthers versatile, 2-locular, introrse, opening longitudinally; pollen with 2 to 6 apertures, smooth or echinate. Gynoecium syncarpous, ovary inferior, 1-locular with a solitary, pendulous, anatropous ovule; style 1, stigma simple or 2-lobed. Fruit dry, indehiscent, one-seeded, enclosed in the involucel, often crowned by the persistent calyx. Seed 1, endosperm present, embryo straight. With about 8 to 12 genera and 250 to 300 species distributed chiefly in the Mediterranean area and Near East, extending to North Europe, East Asia, and South Africa.

Critical appraisal. The system of families followed here accords closely with the most recent classifications. Some authors have included the small, anomalous, tropical American family Calyceraceae in the order, but this seems unsatisfactory from the evidence available. The classification system of one important authority shows large differences from that presented here because it places the woody family Caprifoliaceae in a division called "Lignosae" near the family Cornaceae, in an order Araliales, far apart from the herbaceous family Adoxaceae, which is placed in the order Saxifragales of that system. The families Dipsacaceae, Valerianaceae, and Calyceraceae fall in a division called "Herbaceae" in an order Valerianales in that classification scheme. Such treatment, however, seems to result in concealing affinities within the order.

Alternate classification systems

The inclusion of the family Rubiaceae with the families presented here as the order Dipsacales, as proposed by some earlier authors, appears to make a larger and more heterogenous unit. Some early authors have advocated the merging of the Caprifoliaceae family with the Rubiaceae family, but this would appear to be undesirable because various Caprifoliaceae genera would then be assigned to different rubiaceous tribes and the identity of the family would be lost.

A number of anomalous plants are generally included in the family Caprifoliaceae. For example, it is problematic whether the New Zealand genus *Alseuosmia* and the New Caledonian genera *Pachydiscus* and *Periomphale*, which have alternate leaves and valvate corolla lobes, belong strictly to the Caprifoliaceae family. These have been variously treated as a separate family, Alseuosmiaceae, either related to the Caprifoliaceae family or placed in the order Rosales near the families Pittosporaceae and Grossulariaceae. Similarly, the Asiatic genera *Carlemannia* and *Silvianthus*, which have been variously placed in the Rubiaceae family and in a separate family Carlemanniaceae, are at times considered as doubtful members of the Caprifoliaceae family.

The position of the genera *Sambucus* and *Viburnum* is somewhat problematic as they stand apart from the rest of the family, with their regular flowers and short style. *Sambucus* is further distinct with pinnately compound leaves and extrorse anthers, and it has been treated in some classifications as a segregate family Sambucaceae. Relationships between the genus *Viburnum* and the families Hydrangeaceae and Cornaceae have been suggested. The most accepted treatment seems to be to regard the genera *Sambucus*, *Viburnum*, and the herbaceous genus *Triosteum* as three distinct tribes. The remaining genera in the family form a more or less well-defined fourth tribe named "Lonicereae."

The genus *Adoxa* has been assigned to the families Saxifragaceae, Araliaceae, and Caprifoliaceae but is now generally regarded as constituting an independent family. The relationships with the Araliaceae and Saxifragaceae families depend on the herbaceous habit and the interpretation of the floral parts as involucre and calyx with the corolla suppressed. Morphological, serological, and palynological (pollen structure) evidence support a relationship with the family Caprifoliaceae, especially with the genus *Sambucus* of that family.

Members of the family Valerianaceae often have a characteristic, somewhat unpleasant, odour, which also occurs in *Sambucus* and *Viburnum* of the Caprifoliaceae family and may indicate some affinity to it. Variation in fruit characters has formed the basis for the classification of species in a number of genera, notably *Valerianella*, but the discovery of dimorphism in the fruits of species of *Plectritis* suggests a need for further investigation. The family Valerianaceae is a very clearly defined family and is closely related to Dipsacaceae.

Valerian family relationships

Dipsacaceae is mainly a natural family. The capitate inflorescence and involucre of bracts are characters that suggest affinities with the family Asteraceae (Compositae) but are considered also to be a result of parallel evolution. Differentiation has mainly taken place in the fruits of the family, and these together with variation in the receptacular bracts give good characters for taxonomic division into genera.

Two anomalous genera sometimes receive different treatment by various authorities. *Triplostegia*, with two species, is placed in the Valerianaceae by some workers while others treat it as a distinct family, Triplostegiaceae. *Morina*, a genus of about 15 species, is sometimes treated with the Dipsacaceae or as an independent family, Morinaceae.

Before the true affinities of the many anomalous genera in the various families of the order can be properly assessed, considerably more comparative work is needed to embrace detailed studies in such fields as palynology, cytology, anatomy, biochemistry, and phytoserology.

BIBLIOGRAPHY. A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968); F. EHRENDORFER, "Evolution and Karyotype Differentiation in a Family of Flowering Plants: Dipsacaceae," in S.J. GEERTS (ed.), *Genetics Today*, pp. 399–407 (1964), a general discussion of the morphology and taxonomy with special emphasis on cytology in relation to the evolution of the family; I.K. FERGUSON, "The Genera of the Valerianaceae and Dipsacaceae in the Southeastern United States," *J. Arnold Arbor.*, 46:218–231 (1965), and "The Genera of Caprifoliaceae in the Southeastern United States," *ibid.*, 47:33–59 (1966), a general account of these three families throughout their range, a discussion of their relationships, and a very full bibliography; K. FRITSCH, "Caprifoliaceae," and F. HOCK, "Valerianaceae" and "Dipsacaceae," in *Die Naturlichen Pflanzenfamilien IV*, 4:156–191 (1897), a full treatment in German of the order Dipsacales to the level of genus; T.A. SPRAGUE, "The Morphology and

Taxonomic Position of the Adoxaceae," *J. Linn. Soc. Botany*, 47:471–487 (1927), much biological information as well as a discussion of the relationships of the family; K. STURM, "Monographische Studien über Adoxa Moschatellina L.," *Vjschr. Naturf. Ges. Zürich*, 55:391–462 (1910), a full treatment in German including taxonomic morphology and ecological information; F. WEBERLING, "Morphologische Untersuchungen zur Systematik der Caprifoliaceen," *Abh. Math. Naturw. Kl. Akad. Wiss. Mainz*, 1–50 (1957), an authoritative account of the morphology in relation to the taxonomy of the family; "Caprifoliaceae" and "Adoxaceae," in G. HEGI, *Illustrierte Flora von Mittel-Europa*, 2nd ed., vol. 6, pt. 2, 1:3–96m (1966), a general discussion of many aspects of the two families although the detailed descriptions and keys are confined to European taxa.

(I.K.F.)

Diptera

Although many winged insects are commonly called flies, the name is strictly applicable only to members of the order Diptera, the two-winged, or "true," flies. Diptera, one of the largest insect orders, number more than 85,000 species that are relatively small, with soft bodies. Although the mouthparts of flies are of the sucking type, individuals show considerable variation in structure. Many flies are of great economic importance. Some bloodsuckers are serious pests of man and other animals. These insects, along with many scavenging flies, are important vectors of disease; others are pests of cultivated plants. Flies are beneficial, too—as scavengers, predators or parasites of certain insect pests, pollinators of plants, and destroyers of weeds noxious to man. Dipterous larvae, often called maggots or grubs, are found in many habitats (*e.g.*, in any kind of water, in plant tissue and soil, beneath bark or stones, in decaying plant and animal matter, even in pools of crude petroleum). Adults feed on plant or animal juices or other insects. Diptera fall into three large groups: Nematocera (*e.g.*, crane flies, midges, gnats, mosquitoes), Brachycera (*e.g.*, horse flies, robber flies, bee flies), and Cyclorrhapha (*e.g.*, flies that breed in vegetable or animal material, both living and dead).

GENERAL FEATURES

Flies range in size from midges of little more than one millimetre to robber flies more than seven centimetres long. In general the more primitive flies (*e.g.*, mosquitoes, midges, fungus gnats) are fragile insects with delicate wings. The more advanced flies (*e.g.*, blow flies, houseflies) are generally squat, sturdy and bristly; they fly in a more purposeful way than midges and gnats.

Distribution and abundance

Diptera are abundant throughout the world: in the tropics, the subarctic, at sea level, and high on mountains. They colonize beaches to low-tide level, but few go into deeper water, and only one or two midges are truly marine (*e.g.*, *Pontomyia natans* in the Pacific). On the other hand, migrating flies are found far out to sea.

IMPORTANCE

The abundance, worldwide distribution, and habits of flies combine to make them a nuisance to man. Swarms of midges are a common annoyance. Sweat flies and face flies gather around the eyes, nose, and mouth and also suck blood and pus from wounds and sores. Such flies move constantly from one victim to the next; in so doing they can transfer disease-causing organisms.

The housefly (*Musca domestica*) is dangerous because it feeds on almost any food and moves from person to food, drink, or feces. By transferring infective organisms from the skin and intestine, houseflies are agents in transmitting typhoid, dysentery, cholera, summer diarrhea in children, and other intestinal virus- and bacteria-caused diseases. Eye gnats are a nuisance in warm countries. Although the larvae are plant feeders, the small active adults feed on physiological secretions, particularly those around the eyes. Other flies pierce the skin of vertebrates and feed on their blood. Mosquitoes, black flies, sand flies, biting midges, and horse flies have evolved mandibles and maxillae that are bladelike, piercing stylets.

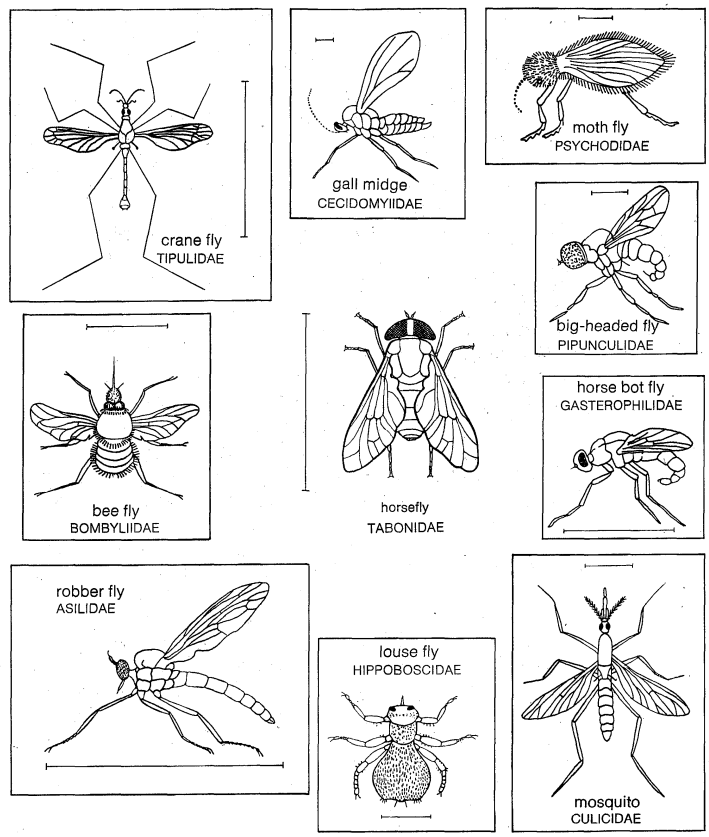


Figure 1: Diversity among Diptera. Line scales indicate the approximate size of each insect.
From *Invertebrate Identification Manual* by Richard A. Pimentel, © 1967 by Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company

These piercing organs are developed only in females, which use blood protein in egg production. Males do not feed on blood. Other groups of flies have evolved different mechanisms for obtaining blood. Tsetse flies, stable flies or biting houseflies (*Stomoxys*), and certain parasitic flies have developed a hard drill-like labium to replace the soft spongelike one. Both males and females have evolved this labium; therefore both feed on blood. A few flies related to the housefly have a spongy proboscis equipped with small teeth for rasping skin around wounds and sores to increase the flow of blood and lymph. Other groups (*e.g.*, robber flies) have developed a piercing proboscis used only against other insects.

Disease-causing organisms in the blood can be carried by a fly inserting its proboscis into successive victims. After piercing its victim's skin, a bloodsucking fly injects saliva into the wound; without the anticoagulant properties of this saliva, bloodsucking would be impossible since the tiny hole drilled by the proboscis would clog with clotted blood. If the mouthparts are contaminated with blood that contains micro-organisms, all are injected, along with the saliva, into the new victim. This is called direct, or mechanical, transmission of disease and occurs only if the fly, interrupted during a meal, finds a new victim before the micro-organisms die. One contagious disease that might be spread this way is tularemia, caused by a bacterium found in wild rodents. Trappers who cut themselves while skinning animals can contract the disease. In North America the bacterium is thought to be transmitted also by the deer fly (*Chrysops discalis*), common in wooded trapper country.

Surra, a disease of horses and camels in the Middle East and the Orient, is caused by *Trypanosoma evansi* and is transmitted by horse flies. Trypanosomes, transmitted by tsetse flies, cause sleeping sickness in man and nagana in animals throughout tropical Africa. These trypanosomes must spend part of their life cycle in the insect before they can infect a vertebrate; this is an example of cyclic disease transmission. The relationship between the parasitic dis-

ease organism and its two hosts, vertebrate and insect, is a result of evolutionary adaptation; however, it is not known whether the trypanosome was originally a fly parasite that spread to man and other vertebrates, or whether it was a human parasite that became adapted to living in a biting fly. An important cyclically transmitted disease is malaria. *Plasmodium*, the causative agent of human malaria, is an acellular protist nourished by red blood cells in the blood of man. Its reproductive cycles cause recurrent bouts of the disease. Occasionally sexual forms occur in the victim's blood; if this form finds its way into a suitable species of bloodsucking mosquito, another stage of the *Plasmodium* begins, preparing the organism to infect another human bitten by the mosquito host. Other diseases known to be cyclically transmitted include yellow fever, filariasis, and many viral diseases. It is likely that there are others not yet recognized.

Injuries to plants

Fly larvae are serious agricultural pests; they feed on young crop plants, retarding growth or killing them. Cultivated crops, because they provide pests with an almost unlimited food supply within a small area, can be devastated by uncontrolled population growth of a pest. On the other hand, wild food plants, because they are scattered and mixed with other varieties, do not usually provide so abundant a food supply and thus serve as a check on population growth. Frit flies can cause a 20 percent loss of an oat crop, and to the value of the lost oats must be added the cost of control measures necessary to save the remainder. Some crops, notably fruit trees and ornamental shrubs, are a financial loss if slightly disfigured by insect attack, though the life of the plant is not endangered. Fruit, although edible after attack by Mediterranean fruit flies, cannot be sold; a few infested fruits can result in loss of an entire consignment. Larvae of gall midges and leafminers lower the commercial value of ornamental plants.

NATURAL HISTORY

Life cycle. *General features.* The life cycle of a fly consists of four stages: egg, larva, pupa, and adult. Since larval forms, always morphologically distinct from adults, also occupy different habitats, flies in effect live two distinct lives and thus are able to adapt successfully to environmental changes. In some flies (*e.g.*, robber flies) neither the larval nor the adult stage predominates; the larva feeds actively in soil, and adult flies of both sexes catch other insects in flight. Among mosquitoes, black flies, and related bloodsucking flies, the larvae have characteristic structures and live active lives under water; the complex mating process of the adults is followed (in the case of females) by bloodsucking and egg laying.

Predominant stage in life cycle

There are many flies in which one stage is predominant. Swarms of adult midges (Chironomidae), for example, are conspicuous and troublesome; but the adult midge lives just long enough—usually less than a day—to mate and lay eggs. Thus, most of the life cycle is occupied under water by the larval stage. The larvae are wormlike in appearance. Some are adapted to oxygen-poor situations; the "bloodworm," for example, which lives in the mud of stagnant waters, uses hemoglobin as a respiratory pigment. Other midge larvae live in silken tubes, either filtering minute organisms from water for food or preying upon larger creatures. Some midge larvae have evolved an elaborate symbiosis, or mutualism, with other aquatic organisms; for example *Nostoc* (a genus of blue-green algae) and certain midge larvae utilize each other's excreta. Larval life as complex as this is not mere preparation for adult life; rather, the adult stage is a revitalizing and distributory stage for the larval one. The adult stage is of relatively little importance in a few other groups, too.

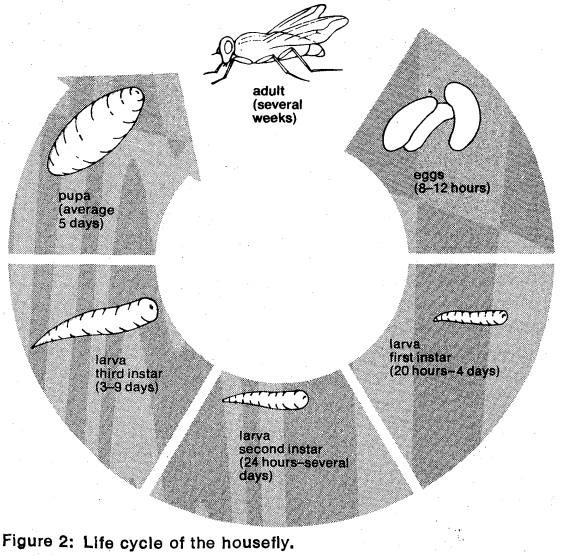


Figure 2: Life cycle of the housefly.

At the opposite extreme are tsetse flies (*Glossina*) and three families of pupipara parasites (*e.g.*, Hippoboscidae, which feed on the blood of mammals and birds; both Nycteribiidae and Streblidae feed only on bat blood). In these families a single egg is produced at one time and hatched internally. The larva, retained and nourished in a kind of womb, is expelled when it has matured and immediately forms a pupa. Thus, these flies have no free larval stage, no independent larval life. Since the pupa is immobile, the active life of the fly is passed as an adult. Most Hippoboscidae and Streblidae, and all tsetse flies, have wings and usually migrate to new hosts, but some species of these families, and all Nycteribiidae, cannot fly and often are wingless. Wingless flies can be identified as flies only after detailed morphological examination.

Eggs. The majority of flies lay eggs, which hatch into tiny larvae after a few hours or several days. The number of eggs laid by a female varies from 1 to about 250; however, a number of successive batches may be laid. The greenbottle fly (*Lucilia sericata*) has laid nearly 2,000 eggs in captivity; however, the total is probably fewer than 1,000 in the natural state when time and energy are lost looking for suitable places to lay. Egg-laying sites, chosen instinctively by the females, are related closely to larval habitats. Since many fly larvae feed in soft organic materials, many females have developed telescopic ovipositors, formed from the last three or four abdominal segments. The female uses the ovipositor to press the eggs into a mass of decaying material. Blow flies and houseflies push their eggs between the membranes of meat or into any convenient cavity in decaying organic material. The small fruit flies (*Drosophila*), which lay in rotting fruits and fermenting materials, also have this type of ovipositor; however, the large fruit flies (*e.g.*, Mediterranean fruit fly), which lay eggs in the rind of growing fruits, have a stiffer ovipositor. Elaborate ovipositors found in the robber flies are used to push eggs into the interstices of flower heads and the axils of grasses, sometimes even into plant tissues, to conceal them and protect them from drying. When hatched, the larvae drop to the ground and burrow under the soil.

Egg-laying sites

Larvae. Fly larvae have one common characteristic: all lack true, jointed, thoracic legs. Many fly larvae have "false legs" (prolegs or pseudopods) similar to those that support the fleshy abdomen of a caterpillar. Flies, much more versatile in this respect than caterpillars, can have prolegs around any body segment. Prolegs help the larvae crawl through narrow spaces or push through soil.

The evolutionary trend among fly larvae has been toward structural simplification; thus, generally, larvae of primitive flies are more structured than are larvae of more highly evolved flies, which show greater physiological versatility. Larvae of most members of the suborder Nematocera (see below *Annotated classification*) have a well-developed head, with antennae, palpi, and complex mouthparts similar to those of many adult insects. Often they are so structurally adapted to their special way of life that they are unable to adapt to any other. This is especially true among aquatic larvae (*e.g.*, mosquitoes) and perhaps reaches the extreme in mountain midge larvae, which live in rushing torrents and crawl on sub-

merged rocks. Their body segments are equipped with clinging processes and suckers.

Maggots

In contrast to highly specialized larvae, about half the fly species have larvae known as maggots. The maggot has lost the complicated head capsule of primitive flies; its pointed anterior end contains one or a pair of mouth hooks. The blunt posterior end has a pair of posterior spiracles (external airholes) that appear to the naked eye as black spots. Microscopically the spiracles are seen as a complex pattern of slits or pores that are useful in distinguishing species.

Although maggots show structural uniformity, they are diverse physiologically. Most maggots feed on decaying organic matter, but there are wide differences in the food preferences of different flies. Eight "waves" of maggots have been distinguished; each wave attacks dead animals in a strict sequence as decay progresses from the newly dead corpse through rigor and putrefaction to mummification. Although some maggots appear only during a clearly defined stage of animal decomposition, the large voracious maggots of many blow flies feed on any animal matter, including living tissues. The best-known blow flies are sheep blow flies, principally species of *Lucilia*. Maggots of *Lucilia sericata*, for example, feed in small dead animals, abattoirs, and garbage cans; they oviposit in soiled wool around the anus of sheep or in the pus exuding from scratches and wounds, where they are important agents of sheep strike disease. These maggots sometimes occur in soil near buildings in cities; their food source is not known.

Other maggots feed in living or decaying vegetable matter. Larvae of the frit fly of oats and the gout fly of barley are maggots of flies that belong to the plant-feeding family Chloropidae. The hessian fly of wheat is the destructive larva of *Mayetiola* (*Phytophaga*) *destructor* of the nematoceran family Cecidomyiidae (the gall midges). Although the external structure of most nematoceran larvae is complex, the structure of the gall midges, which live completely immersed in plant tissue, has evolved in the direction of simplification; gall midges are among the most difficult fly larvae to identify. Also known as gall gnats because feeding larvae cause formation of disfiguring galls on leaves or stems, gall midges harm many kinds of plants. Thus they have evolved simplified structure and physiological diversity regarding food plants as have maggots of more advanced flies.

Molts and larval stages. Among insects in general, the evolutionary tendency has been toward decreasing the number of molts during development, and flies are no exception. The number of larval stages, or instars, is six or seven in black flies (Simuliidae) and four in most other Nematocera. Along the second line of evolution of flies, Brachycera have from five to eight instars while the maggots of the most advanced flies (Cyclorrhapha) have only three. One or two species have no molts. Sometimes molts occur before the larva hatches from the egg. Muscidae, for example, are arranged in three groups according to whether they are trimorphic (*i.e.*, have three free larval instars), dimorphic (*i.e.*, pass the first instar in the egg, have two free larval instars), or monomorphic (*i.e.*, pass the first two instars in the egg, have one free larval instar). Monomorphic larvae are always predatory; the others feed first on decaying matter (are saprophagous), but they may or may not be predatory in their final instar.

Pupa. The external features of the adult fly (*i.e.*, eyes, antennae, wings, legs) are clearly visible in the pupa. The pupa, however, is not always exposed to view; it may be enclosed either in a cocoon of extraneous matter (*e.g.*, soil, or silk, or a mixture of the two) or in a puparium, which is a case formed by the hardening of the larval skin. A puparium is formed in flies of the family Stratiomyidae and others that have maggots as larvae (all Cyclorrhapha). Many families of flies form cocoons sporadically; the cocoon has evolved as an adaptive device that provides extra protection to the pupa. The pupae of mosquitoes, of black flies (Simuliidae), and of a few aquatic midges swim actively. Many pupae that lie in soil or in wood have developed thorns and spines to help them work their way to the surface just before emergence of the adults.

Cocoons

Adult. The adult fly emerges from the pupa soft and crumpled with a colourless skin (integument) and perfectly formed (though not fully pigmented) hairs and bristles. The newly emergent adult swallows air to expand its body and wings and to force blood through its body. In the more advanced flies of the group Schizophora (see below), the ptilinum, an inflatable membranous sac in the head, is used to aid this process. The ptilinum shrivels away after it has performed its function; however, it leaves behind the ptilinal suture, a horseshoe-shaped groove that runs over and beside the antennal sockets and is only found in Schizophora.

Ecology. It has been said that there is hardly any life-supporting medium in which dipterous larvae have not been observed. It is not possible to discuss all dipteran habitats, but the annotated classification below provides many examples. Maggots, however, are the most important larvae, because they play an essential role in breaking down and redistributing organic matter. The waste products excreted by the larvae provide nutrients for molds, fungi, and plants. In addition, the bodies of larvae, pupae, and many adult flies are an important food source for higher animals. Examples are aquatic larvae of midges and mosquitoes, which are staple food for fish. The terrestrial maggots of many flies also have a role in food chains. Since a blow fly can lay one to two thousand eggs, the blow fly population would increase calamitously if more than a few of them survived. Most of the larvae die of malnutrition, desiccation, or drowning, or are consumed by birds. The adult flies are snapped up by birds, small mammals, frogs, and toads. Swallows, swifts, and martins devour large numbers of flies that have been carried up into the air by convection currents. Thus, the population is maintained at a constant level.

The most fundamental importance of flies, therefore, lies not in the few familiar families that contain mosqui-

From H. Weber, *Grundriss der Insektenkunde*

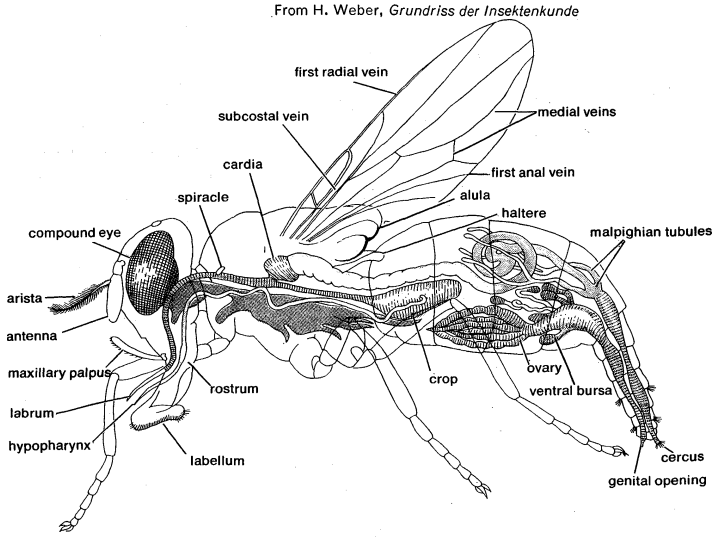


Figure 3: *Body plan of Diptera.*
(Top) Adult; (bottom) larva.

toes, tsetse flies, houseflies, and other nuisance insects, but rather in the large numbers of unfamiliar species that are an essential element in the food chains upon which all life depends.

FORM AND FUNCTION

External features of adult. *General appearance.* The thorax, abdomen, and legs of adult flies vary from long to short; the appearance of the fly is functional as well as decorative. Sometimes the bright colour and pattern of many flies is metallic (*e.g.*, blow flies), but most often the fly is covered with a fine coating called tomentum or dusting. Many flies, particularly those of more highly evolved families, are bristly; and the strongest bristles have a precise location, particularly on the thorax. The arrangement of bristles and the identification method based on them is called chaetotaxy.

Wings. Adult flies have only one pair of wings, on the mesothorax or second thoracic segment. The hind wings, modified into halteres, have a stalk and a knob, or club, that may be large and heavy relative to the size of the fly. The halteres vibrate up and down in time with the wings and act as gyroscopes in flight. If the fly yaws, rolls, or pitches during flight, the halteres, maintaining their original plane of movement, twist at their bases, where special nerve cells detect the twist and cause the fly to correct its flight attitude.

Wing venation

The wings of flies have a defined pattern of veins; each has a name and characteristic location, often of taxonomic value. Few true flies have a reticulation (*i.e.*, network of small veins) such as those in many other insects that are mistakenly called flies (*e.g.*, mayflies, dragonflies, dobsonflies). Primitive flies tend to have complex wing venation, while advanced ones have reduced and simplified venation. Some of the small midges (*e.g.*, Cecidomyiidae, Sciaridae, Mycetophilidae) have reduced wing venation also. Reduction or loss of wings occurs in many families, particularly those that inhabit windy places (*e.g.*, mountains, islands) or caves, or that are external parasites among fur and feathers.

Eyes. The eyes of flies often occupy most of the surface of the head, especially in males, where the eyes may meet in the middle line (holoptic). In female flies, with few exceptions, the eyes do not meet (dichoptic). In some families, notably robber flies and small acalyptrate flies, both sexes are dichoptic. Parasitic flies, or those that live in secluded places, may have very small eyes or none at all. Typically, however, the compound eyes of flies contain many facets; for example, the housefly has 4,000 facets in each eye, about average for insects.

Mouthparts. The mouthparts of flies are adapted for sucking. Most flies have maxillae; many also have mandibles, elongate blades that overlie a groove in the labium and form a tubular channel for sucking liquids. In some females (*e.g.*, bloodsucking flies, mosquitoes) the mandibles act as piercing stylets for drawing blood. Mandibles became functionless or were lost entirely relatively early in fly evolution and therefore bloodsucking families that evolved later had to develop other piercing methods. Tsetse flies and stable flies use the hardened labium; robber flies and dance flies use the hypopharynx; and Dolichopodidae (small, metallic green flies with very long legs) envelop prey in the spongy labella of the labium and crush it with specially evolved teeth. Most flies suck their food; the few exceptions have reduced mouthparts and possibly do not feed at all as adults. Thus the food of flies must be liquid or solids that can be liquefied by saliva and stomach juices. Flies also have a pair of labial palpi equipped with sensory cells that act as organs of touch, taste, and smell. The palpi and the antennae are essential for examining possible food sources and suitable sites for egg laying.

Evolutionary specializations

Antennae. All flies have antennae. Members of the suborder Nematocera (*e.g.*, crane flies, various midges, and gnats) have whiplike antennae with two basal segments (scape and pedicel) and a flagellum of many similar segments. All other flies, properly called Brachycera, or short horns because the flagellum is contracted into a compound third segment, have remnants of the terminal flagellar segments remaining as a pencillike style or a bristle-like arista. Considerable antennal structural differences exist among related genera and species.

Larval features. Larvae of flies have no wings, show no external traces of wingbuds (endopterygote insects), and do not have segmented thoracic legs. Larvae of primitive flies (most Nematocera and Brachycera) have a well-developed head, with chewing mouthparts. Evolution has favoured reduction of the head capsule and replacement of chewing mouthparts with a pair of mouth hooks that move in a vertical plane. Larvae with adaptive external structures (*e.g.*, prolegs) generally belong to the Nematocera or Brachycera. The maggots of the Cyclorrhapha have little external structure other than black mouth hooks and the posterior spiracles. Although a few of these larvae show secondary complexities (*e.g.*, some aquatic larvae of hover flies and shore flies), most cannot be identified beyond the family level.

Nutritional requirements. *Adults.* Nutrition involves balance between feeding habits of larval and adult flies. Primary feeding occurs during the larval stage; adult feeding serves to compensate the shortcomings of larval nourishment. At one extreme are nonbiting midges, with larvae that vigorously filter micro-organisms from water; the adults do not feed. Related to nonbiting midges are biting midges, mosquitoes, and black flies; adult females in these families must supplement an insufficient larval diet. Although one batch of eggs occasionally is laid without a meal of blood, blood is necessary to mature a second batch. Flies that lay one batch of eggs without blood are autogenous; those that cannot lay at all without blood are anautogenous. One species can have both types, possibly as a result of shifting populations or races arising from natural selection. For example, in the far north large populations of biting flies (*e.g.*, mosquitoes, biting midges, black flies, horse flies) occur during the short Arctic summer; obviously there are insufficient numbers of warm-blooded animals to provide food. If the flies find blood, they use it; if not, they still survive.

Blood, pollen, and nectar

Most adult flies visit flowers, which provide water, nectar, and pollen. Pollen, more difficult for a sucking insect to obtain than blood, is rich in protein and is an important source of this nutrient. Certain hover flies crush pollen grains between hardened portions of the labella before swallowing them; many flies actively probe into flowers, covering their heads and eyes with pollen grains. Nectar from flowers contains carbohydrates, and most adult flies use this syrupy liquid. Although their role in pollination is less well known than that of bees, flies are important pollinators of flowers. Some plants (*e.g.*, spurges) are often covered with small flies of different families. Small flies also feed on honeydew from aphids (see HOMOPTERA). Although the name *Drosophila* means "lover of dew," this insect sucks water and any other obtainable fluid. Flies feed on dung and liquid products of either animal or vegetable decay. They obtain nutrients from farmyard manure heaps and garbage dumps. These places also harbour many larvae that feed either directly on available organic food or are carnivorous on other larvae. A familiar example is the yellow dung fly; adults prey on other insects visiting the dung.

Larvae. The adaptability of flies is evident in the wide range of foods that larvae eat. Apart from parasites, the most specialized feeders are larvae that live in plant tissues (*e.g.*, leafmining Agromyzidae, many restricted to one plant species or group). Generally agricultural and horticultural pests (*e.g.*, cabbage root fly) are versatile species, feeding on a variety of wild hosts and modifying their diets when presented with concentrated plantings of commercial crops. Many carnivorous fly larvae (*e.g.*, asilids) probably live in soil and eat vegetable or animal matter, whichever is available. Since adult asilids (robber flies), however, feed on other insects, the larval nourishment is presumed to be inadequate. Some larvae, particularly maggots, that feed on vegetable matter during the first and second instars, become carnivorous during the third instar, when most of the growth takes place.

Respiration

Larval respiration is adapted to the medium in which the larvae live. Although a few parasitic larvae (*e.g.*, Pipunculidae, parasitic in froghoppers, and Drosophilidae, internal parasites of scale insects) get oxygen through the skin, most dipterous larvae need a tracheal system to distribute oxygen. Primitively, the tracheal system probably opened exteriorly by paired spiracles on each segment of the body. The soil dwellers, Bibionidae and Scatopsidae, retain this system, although most families have kept spiracles only on the thorax (one pair) and one at the tip of the abdomen. Even these are closed in some aquatic larvae (*e.g.*, luminous larvae of some fungus gnats and larvae of biting midges). However, mosquito larvae and those of most other water-living flies surface frequently to renew their oxygen supplies. Some larvae pierce the stems of underwater plants to obtain oxygen formed as a result of photosynthesis. Maggots of Cyclorrhapha rely heavily on complex posterior spiracles. Pupae respire through prothoracic spiracles that are sometimes equipped with long tubes extending outside the cocoon or puparium.

EVOLUTION AND PALEONTOLOGY

Diptera belong to the panorpoid complex, which includes Mecoptera (scorpionflies), Trichoptera (caddisflies), Lepidoptera (butterflies and moths), Siphonaptera (fleas), and Diptera (true flies). All are believed to have evolved from an ancestor that lived in moss; four-winged insects that resemble crane flies have been preserved as fossils in Permian deposits (see FOSSIL RECORD). Strata of the Lias Period (Lower Jurassic) contain many true midges; early Brachycera began to appear in Mesozoic; Cyclorrhapha appear in the Cretaceous. By the end of the Eocene Epoch most modern families of flies had evolved. Flies in Oligocene amber and copal are similar to living genera.

CLASSIFICATION

Distinguishing taxonomic features. The wings are the most distinctive feature of Diptera; they consist of a pair of functional forewings and reduced hind wings called halteres that serve as balancing organs. Except for male scale insects (see HOMOPTERA), only Diptera have hind wings modified into halteres. The thorax consists almost entirely of mesothorax filled with muscles that operate the forewings. This feature is useful in identifying wingless flies. The single pair of wings also distinguishes Diptera from other insects called flies (*e.g.*, caddisflies, dragonflies), while the posterior halteres separate the Diptera from other insects that have one pair of wings (*e.g.*, some mayflies and beetles).

Division into suborders is based on structure of antennae and wing venation. Another major feature is chaetotaxy, the arrangement of strong bristles, many in fixed positions and given individual or group names. Separation of Diptera into families is based on habitats and habits (*e.g.*, feeding) of larvae and adults. Genera and species are distinguished by details of head structure, shape and degree of separation of eyes, profile of head, and shape and proportions of leg segments. Abdominal shape often determines characteristic appearance of a genus, but it is difficult to define; the shape varies as the insect is starved, well fed, or pregnant (viviparous flies, such as tsetse).

Annotated classification.

ORDER DIPTERA

Size range one millimetre to 7.5 cm; wings, when present, number two; hind wings reduced to halteres; sucking mouthparts; 85,000 species; worldwide distribution; diverse habitats and diets in both larvae and adults.

Suborder Nematocera

Antennae consist of scape, pedicel, and flagellum with numerous similar segments; maxillary palpi with more than three segments, often pendulous; anal cell of wing open; larvae usually with well-defined head, mandibles horizontally opposed.

Family Tipulidae (crane flies). Elongated body, wings, legs; slow-flying; larvae in soil (leatherjackets), moss, rotten wood, mud, fresh water, littoral, marine.

Family Mycetophilidae (fungus gnats). Fragile, slender; flit about in damp, shady places, among decaying vegetation.

Family Sciaridae. Similar to fungus gnats but more compact, more often indoors.

Family Bibionidae (march flies in northern hemisphere). Compact, well-armoured flies; strong spurs on legs; often abundant on spring blossoms; larvae in soil, sometimes found in a tangled mass near roots of plants.

Family Scatopsidae. Similar to march flies, more often indoors.

Family Cecidomyiidae (gall midges). Tiny flies seldom seen as adults; shapeless larvae burrow into plant tissues, cause formation of plant galls, and deform leaves, stems, and roots; some horticultural and agricultural pests.

Family Psychodidae (moth flies). Tiny, with hairy wings; often seen singly in kitchens, on windows above sinks; some larvae numerous in sewage sedimentation tanks; larvae mostly aquatic.

Family Phlebotomidae (sand flies). Closely related to Psychodidae; adult females suck blood, carry dermal and intestinal leishmaniasis and sand fly fever.

Family Ceratopogonidae (biting midges). Tiny, often with spotted wings (*e.g.*, *Culicoides*); adult females with irritating bite suck blood, carry some parasitic worms; *Forcipomyia* suck blood of insects.

Family Chironomidae (nonbiting midges). Related to biting midges, but females do not suck blood; larvae aquatic; important fish food; adults swarm near water.

Family Simuliidae (black flies). Also buffalo gnats; small, humpbacked, with short antennae; females suck blood, carry parasitic worms that cause "river blindness"; forms nodules under skin; larvae aquatic, filter feeders, attached to stones, underwater vegetation, or fresh-water crustaceans.

Family Culicidae (mosquitoes). Small; elongated; proboscis prominent; palpi often long; best recognized by scaly wings; many females suck blood, carry human diseases (Anophelini carry malaria; Culicini carry yellow fever, filariasis, dengue, viral encephalitis); larvae and pupae aquatic.

Suborder Brachycera-Orthorrhapha

Name usually shortened to Brachycera; flagellum of antennae nearly always fused into a compound third segment, remaining diminutive segments form a stumpy "style" or bristle-like arista; anal cell of wing narrowed, nearly always closed on or before wing margin; palpi seldom with more than three segments, often two or one, held forward (porrect); larvae usually with well-defined head, mandibles move vertically or parallel, cannot be opposed; adult escapes from pupa by a rectangular slit ("Orthorrhapha").

Family Stratiomyidae (soldier flies). Colourful flies, found resting on vegetation with wings closed; males sometimes dance in air; larvae sometimes elongate, aquatic, active, carnivorous (*Stratiomys*); others in decaying vegetation (*Hermetia*).

Family Rhagionidae (snipe flies). Inconspicuous, usually rest on vegetation; some females (*e.g.*, *Symphoromyia*) suck blood; most larvae in soil or in water (some *Atherix* females form egg-laying swarms); some make pits in dust, like ant lions (*Vermileo*).

Family Pantophthalmidae. Large, archaic flies, now found only in tropical forests of South America; wood boring larval grubs sometimes damage commercial timber.

Family Tabanidae (horse flies, deer flies; march flies in Australia). Squat flies with big heads, brilliantly coloured eyes; some females (*Chrysops, Tabanus, Haematopota*) suck blood, are livestock pests; many primitive genera feed only from flowers; larvae in mud or wet soil, either vegetarian (*Chrysops*) or carnivorous (*Tabanus, Haematopota*).

Family Asilidae (robber flies). Adults catch other insects in flight, suck their blood; size varies from a few millimetres to eight centimetres (longest of all flies); characteristic "moustache" of bristles probably protects eyes from damage by fly's victim; larvae in soil or wood; eat any food.

Family Bombyliidae (bee flies). Hairy, scaly; superficially resemble bees, hover over flowers in similar way; often brightly patterned, pattern destroyed by rubbing scales; larvae scavenge in bee and wasp nests, or are parasitic (*e.g.*, locust egg pods, tsetse pupae).

Family Scenopinidae (window flies). Tiny black flies, on windows indoors; develop from larvae in carpets, feed on flea and clothes moth larvae; natural habitat, bird nests or similar dry debris.

Family Therevidae (stiletto flies). Adults resemble Asilidae, but not predatory; larvae like Scenopinidae, elongated, wormlike, carnivorous but sometimes attack plant roots.

Family Nemestrinidae. Rather like Bombyliidae; larvae parasitic in grasshoppers, locusts, perhaps beetles; remarkable for beautiful hovering.

Family Acroceridae (balloon flies). Grotesque; abdomen swollen, thorax small, head tiny; larvae parasitic in spiders.

Family Empididae (dance flies). Adults suck insect blood, also feed from flowers. *Hilara* darts over water, catches small insects; larvae in many habitats (*e.g.*, marine and freshwater mud, decaying vegetation, fungi, running sap from trees).

Family Dolichopodidae (long-legged flies). Tiny, metallic, bristly flies; large numbers sit on leaves in wet places; predatory on other insects; larvae like Empididae, elongated, with little external head structure, same habitats.

Suborder Brachycera-Cyclorrhapha

Usually shortened to Cyclorrhapha; characteristically form pupa inside last larval skin as a puparium; adult fly pushes off a circular cap, hence the name Cyclorrhapha; most families (Schizophora) with a ptilinum (membranous sac inside head), which emerges from a horseshoe-shaped ptilinal suture (identifies adult Schizophora) above antennae, is puffed in and out to help fly escape from puparium or soil or to inflate fly's body; ptilinum atrophies and only ptilinal suture remains; a small group of Aschiza, without ptilinal suture, are recognized chiefly by their wing venation.

Series Aschiza

Family Lonchopteridae. Little known; notable for parthenogenesis; few species; worldwide; sometimes abundant.

Family Phoridae (coffin flies). Tiny flies sometimes numerous indoors; larvae live in any organic debris rich in protein or nitrogenous decay products and scavenge in nests of wasps, bees, ants, termites; breed in carrion; many adults wingless or with short wings (brachypterous).

Family Pipunculidae. Tiny flies; head spherical, noted for precise hovering; larvae parasitic in Homoptera.

Family Platypezidae. Small flies; peculiar legs; rarely seen; appear to dance in smoke of wood fires; larvae live in fungi.

Family Syrphidae (hover flies). Vena spuria in wing runs between third and fourth veins; familiar everywhere; hover over flowers, settle on leaves; some larvae aquatic ("rat-tailed" maggots); larvae of many species feed on aphids on plant stems and leaves.

Family Conopidae (thick-headed flies). Wasplike flies; larvae parasitic in bees and wasps; may be a separate evolutionary line.

Series Schizophora

All flies with a ptilinal suture in head; larvae with no external head structure, mouth hooks visible through cuticle, one pair of prothoracic spiracles and one pair of posterior spiracles, each with either three slits or a mass of small pores; larvae with fore end pointed and hind end truncate are called maggots; larvae with both ends blunt and fleshy, with bulges and tracts of spines, are called grubs.

Section Acalyptrata

Thoracic squamae (*i.e.*, calypters that join base of wing to thorax) are small or evanescent; small soft-bodied flies; major families well established; placement of genera uncertain; families can be grouped according to food preferences of larvae.

Flies breeding in vegetable compost and dung.

Family Lauxaniidae. Larvae in decaying vegetable matter.

Family Helomyzidae. Like Lauxaniidae; most generalized of Acalyptrata.

Family Dryomyzidae. Like Lauxaniidae, but with wider range of food, including fungi; yellow flies often seen in winter.

Family Chyromyiidae. Yellow flies, 1 or 2 millimetres long; breed in debris of bird nests, mammal burrows, caves, cellars; seen singly on windows indoors.

Family Celyphidae (beetle flies). Scutellum enormously enlarged until it covers both abdomen and wings when at rest; tropical dung breeding.

Family Mormotomyiidae. Contains one wingless, African species; looks like a spider; known from only one locality in Kenya; breeds in bat dung.

Family Coelopidae (kelp flies, seaweed flies). Breed in wrack (*i.e.*, heaps of decaying seaweed stranded on beaches) chiefly in temperate countries; adults of some species attracted by trichloroethylene; sometimes pests.

Flies breeding in animal refuse, dung, carrion.

Family Sepsidae. Small, black, roundhead flies, sometimes with spots at wing tips; may breed to infestation level in sewage sludge.

Family Piophilidae (cheese skippers). Larvae in cheese, ham, cured meats, dried fruits, preserved skins and pelts; natural habitat in mummifying carrion; called "skippers" because larvae move both by crawling and "skipping" (*i.e.*, gripping the tip of the abdomen with mouth hooks and flipping the body through a relatively long distance).

Family Micropezidae. Large, long-legged flies; often with conspicuously patterned, blue-black wings; spectacular in tropics.

Family Sphaeroceridae. Tiny, black-brown flies; first tarsal segments of hind legs swollen; abundant throughout world in dunglike materials; some members live in seaweed on beaches; many short-winged or wingless species.

Family Sciomyzidae. Aquatic larvae eat both living and dead snails; may be valuable as controlling agents for injurious snails.

Family Milichiidae. Breed in dung; adults attach to predatory insects and spiders and feed on them; called "insect jackals"; *Madiza glabra* sometimes numerous indoors.

Family Carnidae. Scavenge in nests and burrows. Adults of *Carnus hemapterus* scavenge among bird feathers, break off wings.

Family Neottiophilidae. Nest-breeding; larvae suck blood of nestling birds.

Family Thyreophoridae. Among the rarest of flies; larvae in dead bodies of large animals.

Family Chamaemyiidae. Predatory larvae; known as controlling agents of aphids.

Family Braulidae (bee louse). *Braula caeca*, wingless fly, lives in beehives; larva feeds on wax and pollen stores; adult attaches to bee, may solicit nutritious saliva like other members of bee colony.

Flies with plant-feeding larvae.

Family Ephydridae (shore flies). Transitional; wide range of larval habitats; no substance unpalatable for larvae (*e.g.*, algae, sewage, excrement, carrion, urine, brine, hot springs, tar pools); carnivorous petroleum fly (*Psilopa petrolei*) lives in pools of crude petroleum seepage preying on trapped insects; many larvae feed in terrestrial and aquatic plants.

Family Diopsidae (stalkeyed flies). Some larvae live in decaying plant tissue, others mine in living plants.

Family Chloropidae (frit flies). Most important plant feeders; includes economic pests of cereal and other crops.

Families Opomyzidae, Geomyzidae, Psilidae. Small, usually yellow or grayish flies, plant feeders; *Psila rosae*, the carrot fly, an agricultural pest.

Family Agromyzidae (leafminers). Larvae feed in parenchymatous tissue of leaves, render epidermis transparent and produce either serpentine or "blotch" mines; rarely cause severe damage, but disfigure ornamental trees and shrubs.

Flies with fruit-feeding larvae.

Family Trypetidae (large fruit flies). Form galls in certain flowers particularly Compositae; many Trypetidae larvae feed in living fruits, and ruin them; now worldwide distribution; economic damage by several members (*e.g.*, the Mediterranean fruit fly *Ceratitis capitata*) has resulted in worldwide quarantine laws to regulate entry of fruit into countries.

Family Drosophilidae (small fruit flies). Larvae in decaying and fermenting fruit or any sweet substance; includes *Drosophila melanogaster*, used in genetic studies.

A number of smaller families have been formed to accommodate genera closely related to the two above. Otitidae (Ortalidae) and Lonchaeidae are the most clearly defined. Others such as Ulidiidae, Pallopteridae, Phytalmidae, Camillidae, and Diastatidae are debatable.

Section Calyptrata

Characterized by large squamae (calypters that join base of wing to thorax); Scatophagidae are transitional.

Family Scatophagidae (dung flies). Live around dung, other decaying materials; many also predacious as larvae and as adults.

Family Muscidae (housefly and allies). Many species include the housefly; some larvae carnivorous, especially in third instar; breed in decaying vegetable matter or dung; larvae of *Fannia*, the "lesser housefly" like materials soaked in urine; economically important muscid larvae feed on plant

stems and roots; subfamily (sometimes a separate family) Anthomyiinae contains dipteran plant pests; stable fly, *Stomoxys*, (biting proboscis in both sexes) may be placed in a separate family, Stomoxyidae; tsetse fly *Glossina*, confined to Africa, peculiar structurally and biologically, sometimes placed in the family Glossinidae, occurred in North America in the Miocene.

Family Calliphoridae (blow flies). Some bristly flies with carrion-feeding maggots; common blow flies, *Calliphora* (bluebottles), feed as larvae in dead meat; *Lucilia* (greenbottles) sometimes attack living flesh; screw-worms (*e.g.*, *Cochliomyia*, *Callitroga*) are dangerous feeders in living tissue.

Family Cuterebridae. Offshoot of Calliphoridae above; larvae are parasitic in rodents; one larva, *Dermatobia hominis* (human bot fly) also attacks man; eggs sometimes attached to mosquitoes and other biting flies and carried to their prospective prey.

Family Oestridae (bots and warbles). Larvae live under skin, in nose, and in other head cavities of large mammals; includes the sheep nostril fly (*Oestrus ovis*), warble flies of cattle (*Hypoderma bovis* and other species).

Family Gasterophilidae (horse bots). Larvae live in stomachs of horses, zebras, rhinos and elephants, attached to intestinal lining; relationship with other bot flies problematical; currently classified with other bot flies.

Family Sarcophagidae (flesh flies). Large, gray and black; common around refuse dumps; larval habits diverse, in living or dead animal matter; many viviparous species.

Family Tachinidae (tachinid flies). Ecologically important in balance of nature because larvae are parasites in other insects, spiders, woodlice, and centipedes; employed in biological control of pests.

Section Pupipara

Disputed group, families may merely be convergent in habit; lay living larvae, adults of both sexes feed exclusively on blood.

Family Hippoboscidae (louse flies). Feed as adults on blood of mammals and birds; many fly, some have wings reduced or lost (*e.g.*, sheep ked, *Melophagus ovinus*).

Family Streblidae (bat flies). Distinct, rounded head, wings often functional but fly little; cling closely to host.

Family Nycteribiidae (wingless bat flies). Always wingless; thorax weakened and de-sclerotized; live exclusively on bats; scarcely recognizable as flies.

Critical appraisal. Although there is general agreement concerning major groups of Diptera, disputes concerning relatively minor problems are not uncommon. After extensive study of relationships among families, probable lines of evolution within the order were traced in 1958. The order was surveyed according to the evidence of paleontology, and many fossil flies were illustrated in 1964; this resulted in subdividing the order into an unusually large number of families. Evolution of flesh-feeding maggots and classification and probable evolution of Oestridae has also been investigated.

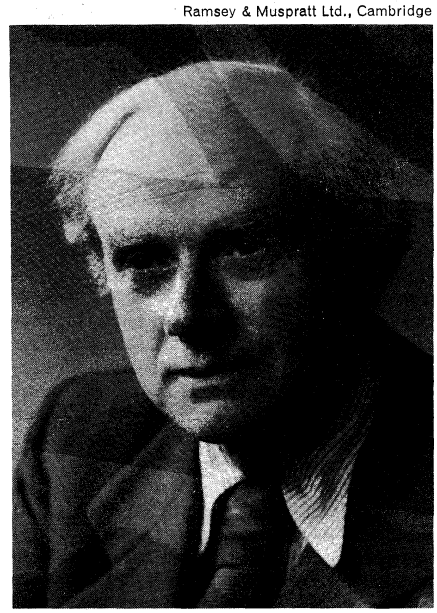
BIBLIOGRAPHY. J.R. BUSVINE, *Insects and Hygiene*, 2nd ed. (1966), comprehensive survey of insects, hygiene, insecticides; C.N. COLYER and C.O. HAMMOND, *Flies of the British Isles*, 2nd ed. (1968), limited to British flies but has general application; C.H. CURRAN, *Families and Genera of North American Diptera*, 2nd rev. ed. (1965), useful keys for primary identification; M. DEMEREC (ed.), *Biology of Drosophila* (1950, reprinted 1965), accounts "normal" development so that experimental results can be assessed; W. HENNIG, *Die Larvenformen der Dipteren*, 3 vol. (1948–52), the only comprehensive account of fly larvae, with references to published descriptions; M.T. JAMES, "The Flies that Cause Myiasis in Man," *U.S. Dept. Agric. Misc. Publ. 631* (1947), useful handbook on larvae that infest the human body; E. LINDNER, *Die Fliegen der palaearktischen Region* (1924, with periodic revisions), comprehensive account of flies with much information applicable throughout the world; I.M. MACKERRAS, "The Zoogeography of the Diptera," *Aust. J. Sci.*, 12:157–161 (1950); H. OLDROYD, *Diptera: Introduction and Key to Families*, in the "Handbooks for the Identification of British Insects Series" (1954); *Collecting, Preserving and Studying Insects*, rev. ed. (1970); *The Natural History of Flies* (1964), the only comprehensive account of flies in English; *Insects and Their World*, 2nd ed. (1966), simple account of the basic principles of structure and function in insects; and *Elements of Entomology: An Introduction to the Study of Insects* (1968), general reading about basic relationships of insects; L.S. WEST, *The Housefly: Its Natural History, Medical Importance and Control* (1951), comprehensive treatise on every aspect of housefly biology. More specialized references on classification may be found in W. HENNIG, "Die Familien der Diptera Schizophora und ihre phylogenetischen Verwandschaftsbeziehungen," *Beitr. Ent.*, 8:505–688 (1958); and in F. ZUMPT, "Some Remarks on the Classification of the Oestridae," *J. Ent. Soc. S. Africa*, 20:154–161 (1957).

(H.O.)

Dirac, P.A.M.

The English theoretical physicist P.A.M. Dirac played a prominent role in the mathematical restructuring of the science of physics early in the 20th century that advanced the understanding of atoms and made it possible to control their behaviour. An early practitioner of the mathematical methods that have come to characterize modern theoretical physics, Dirac applied these techniques with notable insight and originality. Physicists rank him as a theoretician with Albert Einstein and Niels Bohr.

Ramsey & Muspratt Ltd., Cambridge



Dirac, 1960.

Training in mathematics and engineering

Paul Dirac was born in Bristol, England, on August 8, 1902. His mathematical ability showed itself, as often happens, at an early age. At the school he attended in Bristol he was given rather advanced books on mathematics to study independently. His father, a Swiss by birth who was the French master at the same school, encouraged his son to develop his mathematical ability. He wished him also to become fluent in French, to the extent that, according to the son's report, the elder Dirac refused to speak to him unless he was addressed in the French language. This may have fostered Dirac's pronounced tendency to speak seldom and choose his words with utmost care. He avoided company, preferring to work alone. His chief pastime was solitary walks.

Toward the practical end of earning a living, Dirac studied engineering at the University of Bristol. The use of approximations that he learned in this study had a strong influence on his later work; it strengthened his confidence in the intuitive approach to problem solving. He came to believe that a theory expressing fundamental laws of nature could be constructed solely on the basis of approximations, guided by intuition rather than exact knowledge of the actualities. He declared that the actual phenomena were too complex ever to be pinned down in a precise way; a physicist must be satisfied to work only with approximate knowledge of reality.

Dirac's study of theoretical physics began only after he had received a degree in electrical engineering, had failed to find work in this field, and, aided by a grant, had entered St. John's College, Cambridge. From R.H. Fowler,

his faculty supervisor, who had collaborated with Niels Bohr in his pioneering work in atomic physics, Dirac learned the current state of that science.

In 1926, while still a graduate student, he made his first major contribution to physics by devising a form of quantum mechanics, the laws of motion that govern atomic particles. Other physicists (Max Born, Pascual Jordan) working in Germany anticipated Dirac in this achievement by only a few months. Dirac's version of quantum mechanics was distinguished, however, by its generality and logical simplicity.

Application of relativity to quantum mechanics

With the object of formulating atomic laws in the most elegant mathematical language, Dirac applied to quantum mechanics the ideas of Einstein's special theory of relativity. He had the revolutionary idea that the electron could be described by four wave functions, satisfying four simultaneous differential equations. It followed from these equations that the electron must rotate on its axis, an idea that had been developed by other physicists, and also that there must be states of negative energy. The latter conclusion did not seem to correspond to physical reality. In a later paper, however, Dirac suggested that a deficiency of an electron in one of these states would be equivalent to a short-lived positively charged particle. This idea was confirmed when Carl David Anderson obtained cloud chamber photographs showing the existence of positrons—*i.e.*, particles equal to the electron in mass but positively charged. In the experimental confirmation of this phenomenon, an apparent difficulty of Dirac's theory was turned into a triumph.

In his book *The Principles of Quantum Mechanics* (4th ed., 1958), Dirac developed the so-called transformation theory of quantum mechanics that furnished a machinery for calculating the statistical distribution of certain variables when others are specified. He also stated his philosophical position with respect to theoretical physics. The fundamental laws of nature, he wrote, "control a substratum of which we cannot form a mental picture without introducing irrelevancies." In his own work Dirac avoided using any pictorial model or mental picture of the phenomena described by his mathematical symbols.

In addition to refining mathematical descriptions of matter on the atomic scale, Dirac introduced a quantum theory of radiation. He was co-inventor of the Fermi–Dirac statistics. In 1933 he was awarded the Nobel Prize for Physics (shared with Erwin Schrödinger) and in 1939 the medal of the Royal Society. Dirac taught at Cambridge after receiving his doctorate there, and in 1932 was appointed Lucasian Professor of Mathematics, the chair once held by Isaac Newton.

BIBLIOGRAPHY. An introduction to Dirac's work may be found in NIELS H. DE V. HEATHCOTE, *Nobel Prize Winners in Physics 1901–1950* (1953). A historical account of the development of modern atomic physics that stresses the disparate contributions of individual theoretical physicists is BARBARA LOVETT CLINE, *The Questioners: Physicists and the Quantum Theory* (1965).

(B.L.C.)

Directing

The craft of controlling the evolution of a performance out of material composed or assembled by an author is called directing. The performance may be live, as in a theatre and in some broadcasts, or it may be recorded, as in motion pictures and the majority of broadcast material. The term is also used in film, television, and radio to describe the shaping of material that may not involve actors and may be no more than a collection of visual or aural images. Directing also may mean the instructing of anyone who takes part in any kind of radio or television program so as to make best use of what is said or done in front of the camera or microphone.

In the theatre a confusion has arisen between definitions as applied in England and in the United States. The director (as distinct from an old-time actor organizing rehearsals of a play in which he himself appears) emerged during the 19th century, and, like his actors, he worked for an employer who engaged both on contract. The employer, in England, came to be known as the manager, while the person directing the action was known as the producer; in the United States the producer is in fact the one who engages the actors and finances the production, while the artist who directs the actors and shapes the performance is known as the director. With the advent of films these terms were applied to the new industry, and the American usage eventually found its way into the London theatre. Nevertheless, many British provincial repertory companies retained the old terms, and the British Broadcasting Corporation still refers to producers of plays and features. The American usage, however, is spreading in the United Kingdom and the rest of the world, and eventually it will no doubt become general.

Role of the director

The role of the director varies a great deal, not only according to the medium in which he works but also according to whether or not he works with actors. There is always common ground between directors of drama, whatever the medium, because their success depends not only upon knowledge of the specialized form but also upon understanding of acting and human nature. But there need be nothing in common between a maker of fictional films and one who makes abstract films, even though both are film directors; the latter may be an excellent artist, yet may regard people as no more than colours on a palette.

The nature of the medium affects the responsibilities of the director. He who is in charge of a play is responsible, artistically speaking, solely to that play in the same way in which a symphony conductor is responsible to the score. But a film or television director may have a greater social responsibility, for his audience is larger. A documentary-film maker may have a greater one still, since in the editing (an area close to the director's and almost inseparable from it) the meaning of what is photographed or recorded can very easily be transformed into the opposite of what was originally intended.

In all directing there is a tension between content and form. Because in the newer media there are many opportunities for jugglery with technical tricks, directors can be tempted toward virtuosity at the expense of meaning. The justification for this in theatre directing is highly debatable; in any event, the value of immediate effect must be balanced against that of enduring significance, for the two are often mutually exclusive.

Outside the Western world, the director, as part of a general process of westernization, is today established, though only insofar as Western plays and Western-type plays are performed. Indigenous Oriental theatre, as typified by the classical theatre of China and the Nō and Kabuki theatres of Japan, is rooted in tradition. The aim is not discovery but the perfect presentation of what has been discovered. Chinese actors hand down their techniques from father to son. In Nō and Kabuki the positions of the performers on their acting platforms, the precise moments when they make their stylized gestures, and the way in which they use their voices have all been fixed for centuries. For these the director is superfluous.

DIRECTING FOR THE THEATRE

The director as a dominant force began to be recognized in the late decades of the 19th century. One of his functions, however (that of guiding the actors), was probably being regularly practiced as early as the time of the ancient Greek political orator Demosthenes (384–322 BC), who was said to have been given lessons in speech by an actor named Polus. It is a reasonable assumption that, from the beginning of the existence of an acting profession, it became customary for the most experienced performers to give advice and instruction to their less experienced colleagues: actors are seldom as confident as their performances can suggest, and they need repeated confirmation that their abilities approach their self-imposed standards. Such confirmation is likely to be sought from the most respected member of the company. There is a limit, however, to the value of the help given by a fellow actor; the perspective needed to see all the possibilities of performance is usually attainable only from a viewpoint outside the cast. The importance of this perspective is well illustrated in Hamlet's advice to the players, yet

it was well over 200 years after Shakespeare's death before acting companies officially ceased to direct themselves from within.

Nineteenth-century directing. The first professional to coordinate the acting, decor, sound effects, and lighting of a production without performing in it herself was probably Madame Vestris, who in 1830 controlled the Olympic Theatre in London. At her injunction, the company abandoned certain restrictive traditions of dress that had encouraged staginess and artificiality by inhibiting individuality of characterization. At the same time, she introduced varying degrees of realism into her productions, such as interior settings with real doors and windows (instead of painted ones) and sophisticated stage machinery.

Early directors in America

By the 1870s Augustin Daly, although a leading actor in America, was also achieving fame for the very personal direction he was giving his company in New York. Daly's even more famous successor David Belasco started his career as an actor and concentrated on directing and play doctoring, after the '80s. He had a flair for vivid staging and is probably the best known director in American theatre history, but he had little or no influence on acting. In common with a number of other early directors in the United States and England, he left behind him no tradition. It was left to an eccentric amateur in Germany to establish an acting company that was eventually to lead to profound changes in theatrical methods throughout the Western world. The Duke of Saxe-Meiningen's actors aimed to achieve a kind of psychological depth of characterization that was quite new for playgoers, and the company also paid a meticulous attention to detail in their settings and in the way they managed crowd scenes. In 1890 they toured Russia, where they greatly influenced Konstantin Stanislavsky, who was beginning to think along the same lines. A few years later, Stanislavsky and V.I. Nemirovich-Danchenko, who had established the Moscow Art Theatre, learned further from Anton Chekhov, a playwright so concerned with conveying the inner realities of human nature that his works could not be acted successfully except through entirely new directorial methods. Eventually, some two and a half decades later, similar methods seen through the independent mind of Theodore Komisarjevsky, who emigrated to England in 1919, were to affect the English theatre.

Meanwhile, in Scandinavia, France, England, and Germany there appeared realistic writers, very different from Chekhov, whose plays, however, also called for deeper and subtler acting; thus, a "theatre of significance" grew up alongside the "theatre of entertainment" in those countries, leading to the emergence of original and creative directors. Prominent among these were Harley Granville-Barker in England and André Antoine in France. By the turn of the century nearly all professional productions presented in Europe and America were directed by professional directors—though another two decades were to pass before directors were universally acknowledged in theatre programs.

Diversity and complexity of modern directing

Twentieth-century directing. The craft of the theatre director has become a matter of considerable diversity and complexity in the 20th century. His field includes the style of acting of his players, the interpretation of the play, guidance of the actors in the exploring of their parts, and sometimes, though controversially, a complete control over their performances. He also exercises overlordship in matters of decor, costuming, and lighting (sometimes he functions himself in these three areas, but union restrictions in certain countries, notably in the United States, prevent this, even though they allow him the final word). Incidental music, if any, and choreography in a musical play are also under his control, as are visual and sound effects.

These, the director's theatrical instruments, are more numerous and sophisticated than they were in the 19th century. Moreover, the actor is now much more aware than he used to be of the techniques of voice and speech, so that the director needs at least a theoretical knowledge of how a performer achieves the fully expressive tonal octave of which he should be capable. He must also be sensitive to the rhythms and dynamics of speech and how these are affected by emotion and situation; through the contrasts resulting from these, he can create in performance a pattern that can appeal both to the mind and to the senses. Even pauses can be transformed into theatre magic by the director's skill. A great deal depends, too, upon the composition of his stage pictures, which must never remain static for long, and on how and when his actors move. The function of a modern director might be summed up like this: he creates a succession of focal points that must irresistably attract the attention of the audience; then, through the quality of the acting, he tries to ensure that these focal points are as relevant as he can possibly make them.

Directorial styles. The backgrounds of individual directors—some have been actors, some stage managers, some have entered the theatre from other professions—have shaped their styles. Yet style in a director is difficult to gauge. It is much affected by his material, and he may be labelled by facile critics according to the kind of production with which he has been most obviously associated. Max Reinhardt (1873–1943) was famous in two continents as a "spectacular" director very largely because of *The Miracle* (premiered 1911), a play of no great distinction that owed much of its success to his spectacular treatment. His less publicized interpretation of Luigi Pirandello's *Six Characters in Search of an Author*, a psychological play requiring no scenery, was, however, at least as typical of this distinguished Viennese director. The name of Konstantin Stanislavsky is inextricably linked with that of Anton Chekhov, and he is commonly believed to have been the perfect interpreter of the great Russian playwright. The belief is, however, perhaps due less to a full understanding of Chekhov on the part of Stanislavsky than it is to the fact that Stanislavsky wrote repeatedly and at length about the kind of acting that Chekhov's plays needed. We know from the former's prompt script of *The Seagull* and from Chekhov's letters that the two men differed over some fundamental questions of artistic judgment. Chekhov's letters and Nemirovich-Danchenko's last production of *The Three Sisters* (1938), which held the stage for over 20 years, suggest that the Russian playwright was far better served by Stanislavsky's more reticent partner.

The success with which a director has carried out his task is not easily assessed by either playgoer or critic. Both can be deceived by exciting scenery or bold theatrical effects into overvaluing these tools of the trade and forgetting their purpose, which is to ensure an imaginative interpretation of what the author has written. Louis Jouvet, the distinguished 20th-century French director, once wrote:

> There are two kinds of director: the one who expects everything from the play, for whom the play itself is essential; and the one who expects nothing except from himself.

There is much truth in this statement, provided the extremes are not taken too seriously. Peter Brook of the Royal Shakespeare Company seemed to expect little from the author when he did an outstanding production of *Titus Andronicus* in 1955. If he had put greater faith in Shakespeare and had not eked out the script with a multitude of happy theatrical inventions, the public would have been the poorer. On the other hand, Brook's treatment of Shakespeare's great plays have not invariably been happy; these do not need eking out and rather seem to ask that director and actors should build a testament to Shakespeare's poetry and grandeur.

Comedic originality of Tyrone Guthrie

Possibly the best directors cannot be made to fit into categories. Tyrone Guthrie (1900–71), in 45 years of directing every kind of play, progressed from an almost perverse disregard of authors to an ungrudging respect for them. His work ranged from Shakespeare to Aeschylus and took in Gilbert and Sullivan on the way. In viewing the totality of Guthrie's work, what emerges most strongly is an irrepressible comedic originality: here he showed the maturity that overtook him after his early productions. The ability to discern comedy inevitably lurking behind the obvious sorrows of existence is evi-

dence of more than maturity; it also demonstrates the director's knowledge of dramatic ways and means, for the face behind the mask is in the best writers always discernible, and in the less good it is still there, to be forcibly exposed. The discovery of the comedy that is latent gives a highlight to good dialogue and the impression at least of an extra dimension to any writing that might otherwise seem featureless.

Jacques Charon of the Comédie-Française is noted for the positive yet subtle way in which he handles comedy and farce. His juxtaposition of the inevitable with the unexpected and his controlled yet apparently uninhibited comic invention put him in the topmost ranks of 20th-century directors. It is his strong technical discipline that makes him equally skillful and at home with French classical theatre as with modern drama. The same range of ability has been shown by the Englishman Sir Laurence Olivier, who is remarkable for his capacity to direct a play in which he is also acting the leading part. This is a rare talent. It is extremely difficult for the actor-director to maintain creative objectivity while becoming immersed at the same time in his own personal creation.

No inventory of outstanding modern directors would be complete without reference to the Swede Ingmar Bergman. His influence on films apart, that of his stage productions has been considerable. This work has spread over the years and is especially associated with the Royal Dramatic Theatre in Stockholm. He is a master of the art of drama in a multitude of forms and techniques, and probably the most profound director of modern times. His emendations to the classic plays of Europe and of his fellow Scandinavians are respectfully regarded even by those who do not agree with them. Part of the success of Bergman's productions is due to his handling of actors and the admiration that he inspires in them. Consequently, the best performers throughout the world have regarded an invitation to work for him as an accolade.

National conditions affect directorial vitality. The vast size of the United States and its hesitation to accept the cultural asset of professional drama, except in a very few cities, have polarized the American theatre in the post-World War II period. At one extreme is Broadway, highly professionalized but dependent upon the limited vision of speculative investors and demanding little of imaginative directors. Its intellectual sterility encourages the opposite extreme of "Off Broadway" and "Off Off Broadway," where there is often both experiment and imagination but also unfortunately much professional incompetence. Somewhere in between are a handful of resident theatres dotted about the country in and around the bigger centres of population. There the conditions imposed by union limitation of rehearsal hours do not encourage the emergence of directors of distinction.

Inhibitors to imaginative directing

The scarcity of good directors poses the question as to why modern dramatists should not direct their own plays. In the 17th century the great French dramatists Racine and Molière did so as a matter of course. Contemporary writers seem more introverted and tend to lack both the stagecraft and the desire to inspire a company of actors. The contemporary English playwright Harold Pinter is a notable exception. An actor himself, he is as good a director as he is a playwright. Other modern exceptions are the late Harley Granville-Baker and George Bernard Shaw, who saw to it that their plays were performed as they wished them to be. Moreover, they both had a marked effect on English acting and laid the foundations for its 20th-century pre-eminence in the English-speaking world. Shaw, in particular, inspired many a young performer with the understanding of stress, cadence, and rhythm, and even today his stage directions (with most dramatists these are no more than generalized indications and are useless for practical purposes) are models of instructive lucidity from which a director departs only at his peril.

The director's relation to the actor. A proper comprehension of and respect for the actor is indispensable to direction of the highest quality, since the acting in the theatre greatly outweighs such elements as settings, lighting effects, and visual ideas. On this point Louis Jouvet and George Bernard Shaw both have written aptly. The former said: "The profession of director suffers from the disease of immodesty." And the latter, hardly famous for underestimating his own abilities, advised in *The Art of Rehearsal:*

> Do not forget that though at the first rehearsal you will know more about the parts than the actors, at the last rehearsal they ought to know more about them than you, and therefore have something to teach you about them.

If a director has antagonized his actors and has not, on the contrary, stimulated their imagination so that they have become confidently creative, then failure for him is almost inevitable.

The task is difficult. The best way to communicate with any particular group of actors needs the most balanced judgment. Unlike the orchestral conductor, to whose aims the theatrical director's are closely analogous, he cannot control actual performance. Neither can he, except rarely, tell his actors precisely and in every detail exactly what he wants of them. The minutiae are solely their concern, just as in a concerto they are the concern of the soloist (all actors, basically, are soloists, and their creative powers generally are inhibited by drill sergeant methods). Actors need at least the illusion that their own imaginations have full freedom. To direct by guile is therefore most often the key to success, as was demonstrated by the British productions of Theodore Komisarjevsky, the illustrious Russian director whose subtle yet powerful influence in the '20s and '30s complemented that of playwright-director Granville-Barker and helped to complete the ripening of English acting that took place in that period.

Time limitations

The rehearsal process. The results of the director's efforts are naturally affected by the length of time given to rehearsals. These vary according to economic pressures, national customs, and union rules. In some countries, notably the United States, the actors' union has used its powers to escalate salaries and limit working hours in the 20th century. The American director is consequently hard put to find enough time to achieve the depth and the polish to which he aspires. His limit may stretch to four weeks on Broadway and to a mere one or two weeks outside New York City. In many parts of Europe, subsidized theatre has been long established, and, particularly in eastern Europe and Scandinavia, conditions are better. Rehearsals can last five or six weeks and may even extend into months. Despite unionization, the length of the rehearsal day for serious drama in Great Britain is left to the discretion of the artists themselves. The working day is long there, and four weeks is usually considered enough time.

Directorial capacities for patience and self-control usually are put to the test in the last few days of rehearsal. For weeks the company has worked together, and hopefully it has at last reached a result that, as far as the acting is concerned, seems very close to fruitful consummation. The actors may have rehearsed, as in many European countries, on the very stage on which they are to perform. Alternatively, if they are working in Great Britain or the U.S., they may have worked in nothing better than a room. In either case, a time must come when scenery, probably incomplete, appears and almost inevitably looks unlike what was expected. The actors have scarcely adjusted themselves to this when the lighting has to be set and cues established. Suddenly, the marvellous magic world that the artists have built together and come to inhabit as their own is shattered: the technicians have arrived. There is a sense of retrogression and despondency as door handles come off in the actors' hands and complicated speeches written to be spoken to the dying rays of a setting sun are repeated over and over again while the sun obstinately refuses to set. At this point the entire production must be born again and all its disparate components brought together in rhythm and harmony. Once again, the best served theatre artists are those in eastern Europe and Scandinavia, who are given as much as a week or more, before

the first dress rehearsal, in which they and the technicians can work together upon the stage. The British are fortunate if they have two days. The allotted time in the United States may be less than 24 hours.

Importance of casting

There is a crucial responsibility at the other end of the production schedule, before rehearsals even begin. It is the casting process, which is often regarded as an art in itself. An error in casting can be fatal, no matter how much imagination, hard work, and money have been invested in the production. The responsibility should always be the director's, but it is often usurped by the producer in box-office-oriented theatres, and it has become a frequent habit with the latter to seek either the "right type" for the part or a well-publicized "name," irrespective of the ability to act. The wisest casting often works by opposites, so that a hot-blooded character is best played by an actor whose own personality is cool and objective. Then, as a director of the Royal Academy of Dramatic Art in London observed:

> Two contrary elements fuse harmonically to create a character in depth [who] becomes "real" precisely because he is made up of opposite elements in the same way as people are in life.

Casting, the planning of schedules, and the coordination of differing streams of creativity become more complex in production outside the field of straight drama. The rehearsal period in musicals, for example, is longer, and the strain is greater not so much on the director's artistic resources (for he has a great many helpers) but on his powers of leadership. The presence of a choreographer, with a conductor and an orchestra, the cost of which in England and America prohibits the use of musicians before the dress rehearsal, in addition to a heightened emphasis on lighting and visual effects, all contribute to creating a vast potential for discord. The director must resolve these and even turn them to advantage.

Ice shows introduce another complication: dancing, singing, and acting capacity has to be injected into skaters by specialists, each of whom may compete with each other for precious rehearsal time and may vie with each other for precedence. All large-scale entertainments reach the ears of the audience by means of electronic amplification. This introduces the alien element of the sound engineer, who is likely to be more technocrat than artist and may require tactful handling.

Clearly, the director of live entertainment needs to be a man with many qualities, some of which are in conflict with each other. Of all the necessary characteristics, patience perhaps is one of the most important. Unlike directors in other media, he is both artist and maintenance man: in a long run it is his duty to watch the show at least every two or three weeks, thus ensuring that all that happens on the stage continues to be true to his original intentions. Such recurring watchfulness calls for critical balance and powers of endurance not easily found.

DIRECTING FOR THE MOTION PICTURE

In contrast to the theatre, the motion picture's roots are only partly in acting, with the complex heredity of that art in myth and ritual. The film director's ancestry is mundane: the original interest of film makers was in actuality reproduced as accurately as the technique of primitive photography would allow. The artifices that have always been inherent in drama and that gave it a potential for development into a sophisticated art developed only when the novelty of photographing movement began to fade. A new world was then discovered by the cinema. It was a childlike world of make-believe, whose inhabitants could achieve the impossible, such as multiplying themselves 50-fold in the flash of a second or vanishing like ghosts into nothingness.

Out of such conjuring tricks grew the more compelling magic sought by the technocrats of the films. These, the film editors (or cutters) and the film directors, developed the power to vary the duration of a film shot and the movement of the film camera. It was they, far more than the actors (who at first had little or no control of what they did in front of the lens), who controlled timing, suspense, and emphasis. When the motion picture finally became accepted as an art, the directors and editors emerged with a measure of artistic control incomparably greater than a theatre director has ever possessed: they had power over the shape of a manufactured product after the actors had completed their own contributions to it.

The era of the silent film. Rapid movement and sudden change for their own sakes were the staple ingredients of the early films, and so, when the Lumière brothers caused a sensation in 1895 by photographing the arrival of a train, they needed neither actors nor director. And when another Frenchman, Georges Méliès, made the film *Cinderella*, in 1899, though he used actors, his success was due less to them than to his own cinematic inspiration. Méliès invented multiple exposures, slow and rapid motion, and (by an accident) the fade-in and -out, as well as many other devices that seemed mysterious and fascinating to those seeing them for the first time. Such novelties could give substance to fairy tale and seemed to give the lie to "reality" itself. Méliès made altogether about 4,000 films, from which others learned: he can be fairly called the first film director.

The first film director

Since the camera can see farther than the human eye and since what is most easily seen is movement, leading logically to action, incident, and conflict, the natural material of the motion picture became melodrama, together with melodrama's reverse image, farce. This did not require sophisticated actors. The early directors springing from among fairground showmen, photographic handymen, and inventive opportunists were not dramatically informed. They were concerned with inventive photogenic action, composing their frames as good photographers, and the better use of their cameras.

Directors in the United States. In 1903, the impact of a new innovator was felt in the United States. Edwin Porter, a photographer in Thomas Edison's laboratory, had carefully examined the achievements of Méliès. Méliès had invented out-of-sequence filming. He did this for economic reasons, since he could greatly reduce costs by keeping a set standing and shooting on it until it was completely exploited, regardless of chronological order. Porter, however, thought of deliberately interrupting the sequential flow of action so that it no longer appeared on the screen as one long scene, in the manner of the theatre, but had at certain moments other frames from different scenes cut into it. By this he heightened dramatic tension and pioneered the art of editing. Out-of-sequence filming subsequently has become standard practice, since the economic advantages outweigh its defects. The problems it raises for directors and actors, however, are very real. There have been occasional instances where the appearance of the actor and even the manner of the acting have betrayed the order in which the scenes were photographed. The "continuity girl" is employed for the sole purpose of taking accurate note of every detail of every shot so that mistakes may be avoided. Porter also invented the close-up in *The Great Train Robbery* (1903). Here he showed a huge head and shoulders of the villain pointing a revolver at the audience. The film made history, establishing the basic convention of the western and its prime climactic ingredient, the chase.

Innovations of Edwin Porter

Although the French led the world in filming until World War I, and many pictures were made before that in Scandinavia, Italy, and Germany, the growing industry had not by then elevated the film director to prestigious and publicity-earning status. The first to reach that status was D.W. Griffith of the U.S.

Griffith began as an actor, and he brought an improved standard of acting into films. He looked on motion pictures, crude though they were, as something that had possibilities of form and content and so could be shaped into a kind of art. He was not too proud to learn techniques from Porter and Méliès, but he had an original cast of mind and carried innovation a great deal further. Before Griffith the screen was still thought of as a sort of proscenium or frame through which actors, viewed from a fixed position, moved and grouped them-

selves exactly as they had always done upon the living stage. Griffith realized that a film audience was not bound to watch drama from a single point. By changing the placement of his camera—which he correctly saw as the audience's eye—they could be made to imagine that they were actually moving among the characters of the story. He thus discovered a new and tremendous power for contrast and emphasis. His camera moved during shots, it panned from side to side, it tilted up and down, it tracked into close-up and out again. This new flexibility of vision and particularly the new intimacy it gave to dramatic action elevated an unsophisticated mass audience to a position of privilege: now they could almost see the heartthrobs of their heroes and heroines.

Innovations of D.W. Griffith

Griffith invented the "language" of the motion picture. By varying the length of the scenes he shot, he made filmic rhythm possible. His ideas in general greatly influenced the growing motion-picture industry, and many of them have prevailed throughout the 20th century.

By the end of World War I, which temporarily halted developments in Europe, America had a virtual monopoly in film making. The "star system," which Griffith had created by finding unknown performers and building them up with publicity, had become established. The convention of epic romance, of which *The Birth of a Nation* (1915) was an example, became a stereotype of the "big picture." There were other stereotypes, too; the western invented by Porter and the comedy (which was really film slapstick) made famous by the film producers Mack Sennett and Hal Roach. These stereotypes, which guaranteed success at the box office and were perfect instruments for building an industry based on unthinking mass audiences, did not (except for comedy) encourage creative directing. But they did make for slick technicians and continued technical development. The techniques of editing and directing became more and more intertwined. It came to be realized that clever editing could entirely transform what had been performed on the studio floor. A disaster of acting or directing could even on occasion be turned into success in the cutting room. In Hollywood the stars and the producers decided what should be made, but their ideas were seldom new ideas. The conditions were not healthy for growth of original directorial minds.

European directors. In Europe after the war, however, and particularly in Germany in a resurgent industry, the creative director was able to develop. There the market and the audience were smaller, but the tradition of stage drama was very much alive. Films inexpensively made with good actors on modest salaries could create a cinema that was individualistic, free from the obligation to appeal to the lowest common denominator of comprehension, and really a kind of extension of the legitimate theatre that the Germans had always appreciated so much.

Suddenly, the motion picture became respectable fare for bourgeois and intellectual society. The art of the film began to be the subject of critical analysis in the better class newspapers. Such German directors as Robert Wiene (*The Cabinet of Doctor Caligari;* 1919), F.W. Murnau (*The Last Laugh;* 1924), and American-born Arthur Robison (*Warming Shadows*) began to be talked about in circles where such discussions had formerly been limited to the fine arts. The directors of the new Soviet cinema contributed greatly to the general ferment over the new medium. Vsevolod Pudovkin appeared to transform peasants into actors; Sergey Eisenstein developed the cutter's art of montage so that documentary films and dramatic films merged into a single statement. In Sweden Mauritz Stiller discovered Greta Gustafsson and changed her in the United States into Greta Garbo.

Innovations of Soviet directors

By the middle of the 1920's the masters of the silent film had transfigured the primitive trickery of Méliès into a true magic art, backed by the most sophisticated technical achievement. Perhaps its most remarkable monument was *Warming Shadows,* in which Robison succeeded in reducing the subtitles to no more than four, of which only two actually broke into the story. At last the silent film almost literally spoke for itself.

The sound film before World War II. In 1928, recorded sound suddenly intoxicated the tycoons of the industry, and for a while all creative direction seemed to be abandoned. Camera technique retrogressed. Cameras were immobilized in soundproof booths to protect the conquering but undirectional microphones from the sound of their whirring motors. Shooting scripts were peppered with directions for inflicting the public with inappropriate noises, from the thunder of well-oiled locks to the patter of little canine feet.

In the United States the coming of sound killed the Hollywood film comedy, which posterity was to rank as a unique contribution to motion-picture art. Hal Roach and Mack Sennett had each taken hold of the American clichés of sentimentality and melodrama and had turned them upside down. The results were a long succession of remarkable inventions, described by the famous U.S. film critic James Agee as

> such a majestic trajectory of pure anarchic motion that bathing girls, cops, comics, dogs, cats, babies, automobiles, locomotives, innocent bystanders, sometimes what seemed like a whole city, an entire civilization, were hauled along head over heels in the wake of that energy like dry leaves following an express train.

This was near to genius, even if Sennett had not given their first chance to the world-renowned comics Buster Keaton and Charlie Chaplin and others who developed the form to a high level of inspiration.

The tradition created by Sennett and Roach never really died, though they themselves went out of business when dialogue displaced the visual gag. René Clair, the first important director to use sound in an original way, owed much to Sennett. By the 1930s, cameras had become mobile again, and the satirically minded Frenchman seized his chance with both hands. He found new and unknown actors and formed them to his own style. He wrote his own dialogue. He used composers of reputable quality to write background music that made a witty commentary on the action and characterization at a time when Hollywood was still, on the whole, happy with Tin Pan Alley music. Clair was the first screen satirist of sound films. His *À Nous la Liberté* (1931; "Liberty for Us") was a technically innovative comic–poetic elegy for industrial civilization that immortalized the eccentric aspects of the Gallic personality.

René Clair's original use of sound

France seemed to lead the world in cinema until World War II. Jean Renoir, the son of the great Impressionist painter and René Clair's contemporary, earned an enduring reputation with beautifully composed pictures, each frame of which plainly showed its ancestry. His films were of the very earth of France, deep and memorable. His camera examined men's natures rather than recorded their actions.

Hollywood directors, with their huge financial resources, maintained a technical lead, and there were exceptions to a general impression that they did no more than turn out a mass of glossy pictures without enduring substance. Ernst Lubitsch imported his Viennese sophistication and wit and kept up a consistent standard in his comedies. Frank Capra reconciled the qualities needed for creating light entertainment with an honest yet affectionate observation of American character. Preston Sturges, a sort of American René Clair, became a brilliant farcical satirist of middle America.

At this time and for some years after, a kind of stability had become established in the industry, and it seemed that the film makers' art and methods had permanently formed their conventions. Choice of story remained almost everywhere with the producers; and these were largely guided by the box office and what would tempt stars to work for them. The handling of the story was much influenced by the producer, although it was the domain of the director and the cutter, who were complementary to each other and were sometimes even the same person. The director's ability to obtain a good performance did not depend on a detailed knowledge of acting (scenes only needed to be sustained for brief periods in a succession of short shots and so did not call for the endurance and technical resource of stage

Directing conventions of the 1930s

actors). What the director most needed was an instinct for making the performer seem real in front of the camera—by giving him the briefest instruction in the simplest possible terms. The classic example remains the injunction given to Garbo for her tragic closing shot in *Queen Christina* (1933): she was told to "think of nothing at all." If the dialogue was of unusual quality and made particular demands upon those who had to speak it, then a dialogue director was especially engaged and imported from the theatre or from the faculty of a drama school.

Photography had improved vastly over the years, but camera technique had not changed a great deal since the early days. The director still took his "master shot," covering everybody taking part in the scene. He then photographed portions of the same scene several times at closer range and from different angles. Finally, equipped with long shots, a number of three shots or two shots, and a couple of close-ups, he was able to give the editor ample footage from which to choose the best assemblage of frames. The conventions of dissolves and fade-outs to emphasize lapses of time or changes of locality remained the same as they had been for a long time. The biggest change, in the middle '30s, was the achievement of deep focus, which permitted for the first time a sharp image in the background of a close shot. It was used to good effect by the U.S. director Gregory La Cava in *My Man Godfrey* (1936).

The sound film after World War II. Then, in midcentury, films were rocked by television and social change. The mass public, which had loyally supported the movies even through the Depression years, drifted away to the new medium. Insecurity invaded the industry. Perpetuation of old clichés no longer guaranteed success. Producers lost some of the arrogance born of overconfidence and listened more to the views of directors. A new generation of directors was willing and able to take charge of their own destinies.

Even before television had taken hold, the revolution had begun. The U.S. actor-director Orson Welles struck the first blow with *Citizen Kane* (1941). This was something new for Hollywood: realism had sprung before out of its magazine-story romanticism, but it had been a realism that reflected the melodrama of the tabloid headlines or sustained, as in John Ford's *Grapes of Wrath* (1940), the national patriotic myths. Realism inside the home was something else.

Postwar European directing. The realities of World War II made nonsense of the romantic dream. Hitherto, Great Britain had counted for little in quality film making, excepting for contributions of Alfred Hitchcock and the documentaries of Robert Flaherty and Humphrey Jennings. Now British directors began to grow into prominence. Carol Reed came to maturity with *The Third Man* (1949). This film, with a story by the English novelist Graham Greene, revitalized the thriller tradition. Reed's roots were in the theatre, and most of his cast were theatre actors: the depth and subtlety of the acting of this film gave great support to a plot whose merit was that its presentation of the postwar world of disillusionment and the iron curtain might actually be true. Other British directors arrived on the scene and made reputations at Ealing Studios. They were not only prolific but also surprisingly eclectic, as exemplified by Robert Hamer's *Kind Hearts and Coronets* (1949)—the only filmed example of an artificial comedy in the Oscar Wilde tradition, especially written for the medium. The British film benefitted, too, from an influx of American directors, like Joseph Losey, displaced from Hollywood by the anti-Communist political pressures of the '50s.

Emergence of English directors to prominence

In France there was also resurgence. At first the influences were literary. Only France would have given a writer such as Jean Cocteau the opportunity to direct a film. His *Belle et la Bête* (1946; *Beauty and the Beast*, released in the U.S. in 1947) established a tradition of highly articulate storytelling that has continued with René Clément, François Truffaut, and Claude Chabrol. France, too, gave repeated opportunities to the Spanish director Luis Buñuel, a passionate recorder of human suffering whose films are not easy to forget because of the contrast between the beauty of their backgrounds and the harshness of the events depicted in front of them. In the same immediate postwar decade there suddenly appeared Jacques Tati, Chaplin's French counterpart. He, like Chaplin, was a brilliant director and comedian combined; both men spend years of meticulous thought between the films they write, direct, and dominate.

Directorial art of Ingmar Bergman

Sweden evolved a monolithic figure in Ingmar Bergman. He is imitated by many in a country where motion-picture standards have always been high. He is generally conceded as not having been surpassed in the postwar period, though, like all directors, he has had his failures. One mark of his genius is that in two remarkably fruitful decades he has hardly ever failed to produce the unexpected. Another is that it is almost impossible to judge where, in the many aspects of film making, his pre-eminence lies: he excels in them all—in the way he handles actors (with whom he spends weeks of preparation before he begins shooting), in his cutting and camera work, in always getting the right actor for the right part, and in the intellectual richness of his ideas.

Bergman continued his unpredictable course in the '60s and '70s, while the world of films changed around him. The mass audience had deserted to television, though it could still at intervals be lured back to the movie house, if the "big picture" was good enough and the star popular enough. On the other hand, the contradictions and upheavals that had affected every kind of art in the middle of the century—and which some thought mere perversity and others the road to an artistic utopia—created a climate of great opportunity for enterprising producers. They could risk adventures when the stakes were infinitesimally smaller than in the old days of the huge capitalizations of pictures intended for world consumption. Directors benefitted: they attained a freedom of which their predecessors had never dreamed, seeming to explore a multiplicity of paths in all directions at once.

The Italians were in the forefront not only of change but also of the new catholicity. Luchino Visconti, high stylist of direction in the opera house, made "operatic" pictures, in which elaborate and beautiful sets supported complicated emotional stories acted in depth by experienced and star-studded casts. His films were the *haute couture* of the industry—tasteful, traditional in form (if not always in technique), and magnificent. At the other extreme, Federico Fellini, Vittorio De Sica, and Pier Paolo Pasolini asserted their own directorial importance by deliberately diminishing that of the actor. They claimed that amateurs could more easily be "real people," and their pictures often sought to bridge the gap between fiction and the documentary film. Fellini and Michelangelo Antonioni in Italy and Alain Resnais in France, with younger directors François Truffaut and Jean-Luc Godard, ignored the old conventions of editing that had guided the public for decades and started something of a new fashion. Their avant-garde films darted hither and thither, through time and space, replacing the fade-out by the sudden jump cut, indicating contrasting psychological planes by changing from colour to black and white and back again, and emphasizing strong situations by the frozen frame, in which life seemed for a moment to stand stiff and still as in an old daguerreotype. Truffaut and Louis Malle (together with Fellini) were associated with the so-called "auteur theory," whereby a director assumed *total* responsibility for his picture, often doing without a script, and creating, as he directed, by improvisation. The success of such directors and their followers did a valuable service; it demonstrated that conventions and aesthetic principles are not the same, and that even when they are for the most part convenient and practical they may sometimes be flouted to advantage.

Conventions of avant-garde directors

Outside Europe and the United States the film industry expanded, particularly in Japan and India, during the '50s and '60s. Kurosawa Akiwa evolved a great many styles, assimilating Western idioms and translating these into his own. The Indian Satyajit Ray's leisurely long films achieved the persuasiveness of good novel writing

and have achieved considerable success outside his native land.

Directing in the '60s and '70s. In the '60s and '70s there was a tendency for films to grow more subjective, many directors asserting a right for their work to express their personal attitudes. Continuing social evils throughout the world and the futilities of power politics led to films of social consciousness and protest. In the U.S. the virtual collapse of the great wealth and power of the big studios encouraged a striking break with the old Hollywood traditions of make-believe, although films of traditional sentimentality continued to be made and were highly successful. The documentary approach increased in popularity, stimulated by social violence in many countries. In France the phrase *cinéma vérité* was coined to express an allegiance to the unglamorous truth. Such films were shot with the news photographer's "minicamera," often held with a deliberately shaky hand to express the utmost dissociation with the motion picture's glossy past. In 1969 the American director Haskell Wexler solved formidable matching problems by successfully combining colour news film of the Chicago 1968 National Democratic Convention with fictional footage made with professional actors. The result was *Medium Cool*, a thinly disguised protest film.

From the quasi-documentary to the frankly propaganda film was the smallest of steps: from that to the underground film was smaller still. Ever since the 1940s, when the New York director Sidney Peterson made *The Cage*, there had been many purposeful individuals, some with genuine talent and others with none at all, eager to exploit the new and cheaper featherweight cameras in order to demonstrate their personal views. Needing minimal equipment, avoiding professional actors, and combining the function of director, cameraman, and editor, they could make films for small audiences at astonishingly low cost and in perfect freedom. Their subject matter ranged from Freudian obsessions of every kind to an anarchic contradiction of the very nature of the motion picture, the latter being well illustrated by the fixed camera shooting for several hours on a sleeping man done by the U.S. artist-director Andy Warhol in *Sleep* (1963). The ultimate in abstract pattern making brought the underground-film director almost to the position of the fabricator of the animated cartoon. Both were absolute dictators, and neither had any use for the actor as intermediary to help them communicate with their public.

Directors of animated films

In the United States the animator remains rooted in the Walt Disney tradition of the 1930s, somewhat wanting in taste and with an efficient but unadventurous technique. In Europe, however, animation has evolved, perhaps because there is it taken seriously. Cartoonists Jiří Trnka, Zdeněk Miler, and Jiří Brdečka in Czechoslovakia, with Gyula Macskássy and Gábor Kovásznai, are all artists of originality, while in Italy Guglielmo Giannini and Emanuele Luzzati together showed in "The Thieving Magpie" that frivolity is not necessarily inconsistent with subtlety of colouring and a witty conception.

So very catholic are today's habitués of the movies that films of infinitely varied kinds hold the screen. Some, such as those of Alfred Hitchcock, seem to be directed almost exactly as they used to be three decades ago. Others, such as the U.S. director Stanley Kubrick's *2001: A Space Odyssey*, range wide in philosophic and imaginative splendour or, as in John Schlesinger's *Midnight Cowboy*, dig deep into the psyche of ordinary people. In the main, directors have given good value in return for the freedom that the crumbling of the old tyrannical film empires has given them. Many of the most creative are now their own producers, and they have taken charge of their own destinies.

DIRECTING FOR TELEVISION

Although the output of theatre in audiovisual programs is greatly exceeded by other televised material, drama, particularly in the forms evolved for the motion picture, has put its stamp on direction for television. The television director, like his theatre colleague, should lead the eye of the audience to a succession of focal points in a picture of meaningful activity. Like the film director, he should give composition to his images, achieving a variation of intimacy in the shots he chooses, as well as the rhythm that cutting can give to his sequences. It is arguable whether any director ever reaches the goal of his ideal; in the case of the television director, there is no doubt: he cannot do so. The reason lies both in the cost factor and in the very nature of the medium, which combine to deprive him of what every director needs—time in which to have enough second thoughts. Although he is allotted a period for pre-preparation, a most important and necessary part of his work is done at the very same time as his production is being recorded for subsequent screening—and sometimes, though rarely, as it is being screened "live."

Unique role of the television director

Primarily because of high production costs, the choice of shots he uses—the kind of decision the film director-editor is able to make from a number of alternatives at leisure—has either to be incorporated beforehand in the camera script and adhered to within narrow limits or else it is improvised in the control room during recording, in a succession of decisions that must be immediate. Always, the actual moment of cutting away from one picture to another (to which the film cutter or director can give both time and thought in choosing the exact frame in the cutting room) is a choice boldly made during the action being photographed. The decision in a live show is irrevocable: in a recorded television program it is not entirely so. The director may stop the running of the videotape and re-record a section of it. But at best he can do no more than to repeat, with some improvements, the pictures he has already taken. But opportunity for alternative pictures is not available to him, since they do not exist. Occasions when he may stop the tape are limited by the total of hours agreed with the unions for each recording session. This allowance is neither parsimonious nor generous, and a director may well wish to exceed it. The cost of doing so, however, can be high.

Requisites for drama directing

In the case of a drama director, two kinds of experience are called for. He needs theatre knowledge, since he will want actors who can sustain character and technical power throughout the entire length of their parts. This means stage actors, not film actors, whose time in front of the camera is measured in seconds or minutes. Since, however, television essentially consists of pictures in motion, all directors must also have knowledge of film technique and the techniques of the cutting room. Early television directors were drawn from radio, films, and the theatre. Each discipline had its contribution to make, and radio experience proved particularly useful in the vast field of news, discussion programs, "participation" programs, competitions, music, interviews, and the whole realm of sports coverage.

The pioneers learned as they worked, making mistakes that were forgiven by a minority public easily fascinated by novelty. As demand increased, it began to be realized that the new medium must have a discipline of its own. Courses of training for direction were established. The lecturers often insisted that anyone trained to a sufficiently high standard for drama would be fully equal to dealing with television of other kinds. On the whole, this insistence was justified: a good director of a football telecast has to fulfill the drama director's function of "leading the eye to the focal points." He cannot guarantee their relevance, however, for he never knows beforehand where the most exciting television pictures need to be looked for. Such a director hopes, through experience and intuition, that his cameras will succeed more often than they fail in shooting in the right direction at the right time. Because of this unpredictable element, the sports director himself presses the buttons that determine which camera is on the air or is being recorded.

In drama and other scripted programs the movements of the participants have been rehearsed, and the shots have been planned, numbered, and laid down in the camera script. The unpredictability lies only in the minor yet important changes of camera work that the director may be able to make while the production is actually in progress or being taped. In a typical production the drama

director sits in the control room with a vision mixer (British) or a technical director (American), who is connected by earphones with each of four to six cameras on the set and who "calls the shots" according to the script, unless ordered by the director to do otherwise. Others in the control room are the production assistant and the sound mixer, who is in touch with the various microphone boom operators on the set and who, fortunately for the director, ensures that voices and sounds are always of appropriate volume and quality. The director keeps his eyes on the half dozen or so monitor screens in front of him. One of these shows the shot that is being recorded or transmitted at any given moment: another, called the preview monitor, indicates the next shot to come. Others inform the director exactly what each of the cameras is "looking at." If camera one is "on transmission" it will indicate the same picture as the transmission monitor, and, if camera three has the next shot according to the camera script, it will duplicate the preview monitor: cameras two and five will show preparations for what is further ahead in the script. The director concentrates his attention on the picture that is being transmitted and also on the preview monitor, so that at the right moment he may cut from the first to the second and call the next shot on the cue as indicated in the camera script. This is where unpredictability enters, for if the director feels impelled to improve on the camera script he may alter the precise moment of the cut. He may even become aware that one of the camera monitors is showing a better picture for his next shot than the one that was planned and so change it. He is thus like a juggler keeping several balls in the air at once. An additional screen can be his concern from time to time. This is the monitor upon which film may be running at certain times so that it can be cut in and out at the appropriate cue.

Use of film in television directing

Films can be valuable aids to a television production. A script may call for a sequence that cannot be realized in the studio—some scenes, for example, might be taking place on a fishing boat entering a harbour. Or an actor could be required to make a complicated change of make-up and clothes between scenes separated by "script time" but connected in screen continuity, which could waste many precious minutes of recording time. In such cases, filmed footage shot prior to the recording of the production could fulfill the requirements. The use of film, however, also presents its problems: conditions for photography in two such contrasted media as film and television are very different. There are virtually no opportunities in television for altering lighting and thus extracting the best possible photographic advantage from each shot. Film shots, on the other hand, are separately lit, as there is always plenty of time between each in which to change the lighting setups, which differ according to whether they are close-ups or long shots. Thus, the quality of the picture varies from one media to the other.

Unsurprisingly, many television dramatic productions turn out to be substandard. A high degree of adaptability, strong nerves, unfailing picture sense, and an ability to make quick decisions are all called for in the television director—in addition to all the attributes of good theatrical direction.

In the current-affairs program, with its journalistic affinities, ephemerality is appropriate. But here there are responsibilities for the director that are both public and onerous. Because he is a director, he must be something of an artist, but he must also maintain at all times the objectivity of good reportage. Television is a public service, and the director must not bring his personal opinions to the screen. If his program depicts, for instance, a gathering of politicians whom he dislikes, or if he photographs two people in weighty argument, he must favour both sides equally with his angles and the closeness of his shots.

The obstacles in the television director's path mean perpetual and exciting challenge, for the medium is still evolving. The field, therefore, attracts young recruits, although it often fails to keep them. It is a useful stepping stone to the film industry, where there can be more time for aesthetic and creative considerations.

DIRECTING FOR RADIO

Although radio has in some ways become television's poor relation, directing for this medium can show qualities that provoke admiration. The range of programs calling for a director is far from narrow. Though the bulk of material that employs them consists of drama, there are also discussion programs, sound documentaries, interviews, and talk shows in which the director's function is vital.

The fact that sound is the sole means of conveying meaning has important consequences. The director, like a blind man whose other senses are heightened by affliction, acquires a highly sharpened sense of what is essential and what is not. The mistakes made in the early sound films were not repeated by the early radio directors, who knew how easily too much of anything spoils the effect. They also discovered a remarkable refinement in evoking reality through well-chosen music and by inventing strange and original sounds.

Qualities of a good radio director

In good radio directors two qualities particularly stand out. They are, in a sense, pedants in the way they concentrate on minutiae in nuance of inflection and stress. Many a public speaker, thinking himself an expert in tonal variety, is humbled when he first rehearses a radio talk. In half an hour of professional instruction, his typed script has become pockmarked with a multitude of dactyls and spondees or their equivalent—all indicating how little that speaker had really explored the possibilities of expressive speech. The other interesting and paradoxical quality of the good radio director is the vividness of his visual imagination. In order to be confident that the microphone will convey everything possible in a play that will be heard but never seen, he himself must "see" the totality of that play in his mind's eye: he must be aware of all that the audience will gather only by implication. Thus, radio directing demands an unusually perceptive and retentive mind, as well as all the expected qualifications of a theatre director. Possibly this explains why, despite the decline in status of this vocation since the advent of television in the 1950s, the radio director is often a particularly dedicated individual who obstinately insists on remaining in the field, since there is nearly always enough rehearsal time to do his work well.

BIBLIOGRAPHY. H. GRANVILLE-BARKER, *Prefaces to Shakespeare*, 4 vol. (1963), an analysis of the theatrical, as well as the obvious poetic values, to be discovered by anyone staging Shakespeare's plays; TOBY COLE and HELEN CHINOY (eds.), *Directors on Directing* (1963), a collection of essays useful for the light it sheds on the ways of a score of directors; KONSTANTIN STANISLAVSKY, *My Life in Art* (Eng. trans. from the Russian, 1924; reprinted 1956), the odyssey of the man who more than anyone else codified the actor's inventory of tools and methods; V.I. NEMIROVITCH DANCHENKO, *My Life in the Russian Theatre* (Eng. trans. from the Russian, 1937), a valuable complement to *My Life in Art;* JOHN FERNALD, *Sense of Direction* (1968), a detailed examination of the way in which a director works with actors, including analyses of many scenes from classic plays, showing how to approach them as a director; ELSIE FOGERTY, *Speechcraft: A Manual of Practice in English Speech* (1930), an authoritative guide as to what actors should be able to do and the mistakes they can make; TYRONE GUTHRIE, *A Life in the Theatre* (1959), personal revelations about directing and the theatre; G. WILSON KNIGHT, *Shakespearean Production*, 3rd ed. (1964), a first-class complement to Granville-Barker's *Prefaces;* RICHARD L. STERNE, *John Gielgud Directs Richard Burton in Hamlet* (1967), showing in detail exactly how a great Shakespearean directed this play over a period of four weeks' rehearsal; RAYMOND SPOTTISWOODE, *Film and Its Techniques* (1951), an excellent technical book on film making, though now somewhat out of date; HARRY M. GEDULD (ed.), *Film Makers on Film Making* (1967), the film complement of *Directors on Directing*, presenting the personal views of directors from Méliès to Buñuel; RUDOLF BRETZ, *Techniques of Television Production*, 2nd ed. (1962), and GERALD MILLERSON, *The Technique of Television Production*, rev. ed. (1968), two technical television books with full explanations of the methodology of the television studio—the first on U.S. practice and the second on practice in Great Britain; LANCE SIEVEKING, *The Stuff of Radio*, 3 vol. (1934), a work on radio direction and writing by a great radio pioneer.

(J.B.F.)

Disciples of Christ

The three major bodies of the Disciples of Christ stem from a common origin.

The Churches of Christ (1970 membership 2,400,000) emphasize rigorous adherence to the New Testament as the model for Christian faith, practice, and fellowship. They reject general religious institutions other than the congregation, practice a dynamic evangelism based on a literal view of the Bible, and remain aloof from interdenominational activities.

The Christian Church (Disciples of Christ), which reported 1,592,609 members in the U.S. and Canada in 1970, has developed organizations for missionary work and ecumenical involvement and accepts its status as a denomination (uniting a number of local organizations).

The congregations loosely related in the Undenominational Fellowship of Christian Churches and Churches of Christ (estimated membership in 1968 1,000,000) were formerly identified with the Disciples but refused to follow them into the Christian Church. They earlier had refused to follow the Churches of Christ in rejecting musical instruments in worship and missionary organizations as a matter of biblical principle; they later repudiated the openness of their fellow Disciples toward biblical criticism, theological liberalism, ecumenical involvement through "official" channels, and development of denominational institutions.

In a larger sense, Disciples of Christ includes "sister churches" in Britain, Australia, and New Zealand, known locally as Churches of Christ, with origins largely independent of America. It also denotes churches in other lands resulting from the missionary efforts of all these bodies.

NATURE AND CHARACTERISTICS

The original ethos of the Disciples blended the independence and pragmatism of the American frontier with an uncomplicated biblical faith in the restoration of the "ancient order" in the church. They repudiated "human creeds" and traditions as requirements for Christian fellowship, understood Baptism as the immersion of believers only, and recognized no churchly authority beyond the congregation. This simple formula's typical "sectarianism" was combined with a strong catholic impulse: a plea for the union of all Christians, the regular celebration of the Lord's Supper in weekly worship, and the use of inclusive biblical names such as disciples, Christians, saints, and brethren, that would not distinguish them from other believers.

The original equilibrium among these varying emphases did not last for long. Without rigid structures of official dogma or government, Disciples of all persuasions have responded to historical influences by changing, wittingly or unwittingly, the original formulation.

HISTORY

Origins. The movement emerged on the American frontier through various efforts to cut through the complexities of sectarian dogma and find a basis for Christian fellowship. Out of the Great Western Revival (1801) in Kentucky arose the short-lived Springfield Presbytery, which dissolved in 1804 so that its members might "go free" simply as Christians. Their leader, Barton W. Stone, championed revivalism, a simple biblical and noncreedal faith, and Christian union. In the upper Ohio Valley, Presbyterian Thomas Campbell, a contemporary of Stone, organized the Christian Association of Washington (Pa.) in 1809 to plead for the "unity, peace, and purity" of the church. Soon its members formed the Brush Run Church and ordained his son Alexander, under whose leadership they accepted immersion of believers as the only scriptural form of baptism and entered the Redstone Baptist Association. Alexander Campbell rapidly gained influence as a reformer, winning fame as preacher, debater, editor (*Christian Baptist*), and champion of the new popular democracy. His colleague Walter Scott developed a reasonable, scriptural "plan of salvation" based on "positive," or objective, steps into the church: faith, repentance, Baptism, remission of sins, gift of the Holy Spirit. The formula attracted thousands who longed for religious security but, perhaps because of temperament, did not "experience" the emotional crisis and subjective "assurance" that characterized the prevailing revivalism.

Thomas and Alexander Campbell

By 1830 the regular Baptists and the reformers parted company, the latter terming themselves Disciples. Two years later Stone and many of his followers joined with them, though continuing to use the name Christians.

Alexander Campbell from 1830 on turned to constructive churchmanship. He established Bethany College, then in Virginia (1840), for the education of leaders and agitated unsuccessfully for a general church organization based on congregational representation. The first general convention met at Cincinnati, Ohio, in 1849 and launched the American Christian Missionary Society as a "society of individuals" and not an ecclesiastical body. Similar cooperative organizations emerged in various states to support evangelists and to establish new churches. The Christian Woman's Board of Missions (1874) and the Foreign Christian Missionary Society (1875) initiated successful programs overseas, and other boards were soon founded to promote building loans for new churches, care for aged ministers, homes for orphans and the aged, temperance, and other "causes." The Centennial Convention at Pittsburgh in 1909 claimed an attendance of 30,000; they had come to celebrate a century of triumph for the New Reformation, or Restoration Movement.

Controversy and separation. Meanwhile, schism had begun to sunder the ranks, yet without shaking the confidence of the Disciples in their plea for union. They had held together during the controversy over slavery and through the Civil War, when major American denominations had divided. In the succeeding era of bitterness, however, the Disciples also suffered schism. New developments in response to growing urbanization and sophistication brought two sharply divergent responses. The conservatives regarded such developments as unauthorized "innovations," while the progressives (pejoratively termed digressives) looked on them as permissible "expedients."

Plea for union

Discord first arose over the "society principle" involving general missionary work. Alexander Campbell's biblical view of the church had kept pushing him toward a general church organization, but he could never find a convincing biblical text to support his proposals. Frontier independence and pragmatic popular biblicism prevailed. The "society principle" seemed to its advocates a legitimate solution: entertaining no ecclesiastical pretensions as a secular corporation, the missionary society provided a means by which individual Disciples could work in voluntary cooperation. But the opponents saw in it a repudiation of the Bible as the determining rule of practice.

The introduction of musical instruments (reed organs) into Christian worship concerned congregations much more directly and led to many local disputes. Other innovations added occasion for controversy—the infringement of the "one-man pastoral system" on the local ministry of elders, introduction of selected choirs, use of the title Reverend, and lesser issues.

Dispute over music

In 1889 several rural churches in Illinois issued the Sand Creek Declaration, withdrawing fellowship from those practicing "innovations and corruptions." In 1904 a separate "preacher list" issued unofficially by some conservative leaders certified their preachers for discounts on railway tickets. The Federal Religious Census of 1906 acknowledged the separation between Churches of Christ and Disciples of Christ (who commonly took the name Christian Churches) even though many congregations did not decide which they were for some years.

The crucial issue centred in the manner of understanding biblical authority. Both conservatives and progressives accepted the New Testament as the only rule for the church. The conservatives, heavily concentrated in the South, applied a strict construction to Scripture; this

required a specific New Testament precept to authorize any practice. The progressives tended toward a broader construction, accepting as expedient such measures as they found harmonious with Scripture or not in conflict with it.

Disciples in the 20th century. Disciples had experienced their most rapid growth in rural America. Their leaders undertook to respond to the passing of the frontier, the growth of cities, and the emergence of urban expectations. Whereas the Churches of Christ had opted for the practices established in the rural past, regarding them as biblical, the Disciples of Christ (progressives) were able to find some flexibility in the biblical rule. Nevertheless, rural and small-town Christian Churches predominated in numbers and membership even past midcentury, and the newer social and cultural influences did not affect all of them simultaneously.

Development of a professional ministry

Urban churches demanded full-time leadership, and Disciples gradually developed a professional ministry. In the first half of the century they worked hard to establish collegiate education as standard for ministers. As late as 1930, only 11 percent had graduate education, and the rapid growth of theological seminaries did not come till after World War II. The expanding corps of educated leadership reworked the inherited formulas, introducing both ideas and practices that troubled the more traditional.

The cooperative organizations underwent notable changes. In 1917 the old general convention, a week-long series of annual meetings of the various societies, gave way to the International Convention (U.S. and Canada), to which all cooperative agencies were expected to submit reports for review and advice. It consisted of two houses: a large Committee on Recommendations, constituted of delegates from state conventions, and a mass assembly; in 1967 it made the assembly a body of voting representatives from congregations.

Meanwhile, a number of the agencies had combined in 1920 to form the United Christian Missionary Society. Ten years later most state and national agencies entered Unified Promotion, a cooperative program of fund raising, with voluntarily accepted restraints on independent campaigns, and with distribution on the basis of agreed allocations. Thus they gradually evolved, in effect, one general budget. From the start the United Society drew intense criticism for ecclesiastical giantism and theological liberalism. There was especially intense opposition to reports of "Open Membership" in the China mission. (Open membership, increasingly practiced in the U.S., meant reception of unimmersed Christians from other denominations.)

In 1927 traditional forces established the North American Christian Convention. Many churches gave their support to "independent" missionaries in large numbers, as well as to "independent" Bible colleges, youth camps, district meetings, Bible school curriculum, various publications, and a yearbook—all of them explicitly denying official status—more or less parallel to the "cooperative" agencies. The power struggle focussed on the placement of ministers and resulted, on the cooperative side, in enhancing the leadership of the state secretaries and creating the pressure for delegate conventions in the states.

Ecumenical participation

The cooperative conventions (state and international) also became instruments of ecumenical participation, electing representatives to the old Federal Council of Churches (and to the succeeding National Council and the World Council of Churches), as well as to the state councils, and to the Consultation on Church Union. Thus for the sake of their original catholic commitment, the "cooperatives" accepted status as a denomination, a compromise that the independents rejected.

A growing sense of moral obligation toward the common cause led in 1950 to the formation of the Council of Agencies, which included all organizations reporting to the International Convention. Legally independent, they sought by consultation to avoid overlapping and to develop a common mind. From the Council came a proposal for a Commission on Restructure, appointed by the convention in 1960. In 1967 the convention approved the commission's Provisional Design for the Christian Church (Disciples of Christ), ratified in the ensuing year by all 40 area conventions and 15 national agencies. In 1968 the International Convention of Christian Churches reconstituted itself under the Design as the General Assembly of the Christian Church (Disciples of Christ). All general agencies became provisional administrative units of the Christian Church in the United States and Canada, and the state organizations became provisional regions. Congregations retained full legal independence, but the system provided for corporate unity through decisions by representatives from congregations and regions.

Fear of infringement on congregational freedom and theological opposition to the doctrine of the church underlying the restructure proposals led to active opposition. Many independent congregations formally requested withdrawal of their names from the Yearbook of Christian Churches (Disciples of Christ), and a campaign led many cooperative churches to follow suit. From 1967 to 1969 the number of congregations listed dropped from 8,046 to 5,278.

Schism

Meanwhile, a self-appointed Chaplaincy Endorsement Commission for the Undenominational Fellowship of Christian Churches and Churches of Christ asked recognition by the U.S. government to represent those congregations that had elected "to continue as free, independent, and completely autonomous local churches" apart from the restructured Christian Church.

The World Convention of Churches of Christ since 1930 has sponsored mass meetings for fellowship and inspiration at five-year intervals. It attracts both cooperative and independent Disciples from America and from many nations, but very few from the American Churches of Christ. As a world confessional body it appointed official observers to the second Vatican Council, but it exercises no administrative or programmatic functions.

Churches of Christ in the 20th century. The Churches of Christ continued to avoid any general organization and until midcentury at least had shown little interest in their own history; they centred attention on the New Testament and on the Restoration Movement (involving the restoration of the doctrines and practices of primitive Christianity) of the early 19th century. Records of their 20th-century history still remain as raw material in their journals.

In 1906 the vast preponderance of the members and leaders of the Churches of Christ lay in the South, with heaviest concentrations in Tennessee and Texas. The reported membership of 159,658 apparently did not include all who held their general position. In the ensuing half-century they have grown into the largest of the three Disciple groups. The migration from the rural South to urban centres has brought impressive membership gains in the North and the West—aided by a vigorous evangelism that has made intensive use of radio. Missionaries have established churches in Asia, Africa, Latin America, and Europe, winning converts especially from Roman Catholicism. Many churches now forward their missionary funds to an agent for disbursal, all the while making certain that the actual appointment of missionaries remains the prerogative of congregational elders.

Missions

The churches' doctrine permits individual initiative in certain types of religious (not ecclesiastical) enterprises. A vigorous journalism has flourished for more than a century, the most influential papers being the *Gospel Advocate* (Nashville, Tennessee) and *Firm Foundation* (Austin, Texas). Benevolent homes provide care for children and the aged. A number of churches conduct Christian day schools, while private colleges offer Christian higher education and receive support from churches. A graduate school of religion at Harding College in Memphis, Tennessee, offers a three-year Master of Theology degree.

Variations of conviction about specific practices (whether a single, "common" cup or many cups are to

be used in communion) and doctrines (especially millennial ones about the perfect age of Christ's reign on earth) have produced sharp controversies and withdrawal of fellowship, but the continuing absolute stand against any type of ecclesiastical structure makes it difficult to trace the effects of such discord.

After two generations of refusal to have fellowship with their divergent brethren, some leaders in the Churches of Christ took initiative in the 1960s to set up informal forums or conferences on unity with members of the Christian Churches, both cooperative and independent. With no official status, these meetings provided opportunity for a limited kind of ecumenical dialogue. Their doctrinal stance, in repudiation of ecclesiastical organization, prevents members of both the Churches of Christ and the Undenominational Fellowship of Christian Churches and Churches of Christ from official participation in ecumenical gatherings.

BELIEFS, WORSHIP, ORGANIZATION

Influences and relationships. Alexander Campbell summarized his theology in *The Christian System* (1835), the most influential book in shaping Disciple thought. In it he outlined a commonsense biblical doctrine against the complex theories of the schools and the sects. The influence of Renaissance humanism on Campbell is evident in his insistence on going to the sources and his rules of biblical interpretation. He shared the presuppositions of the Enlightenment in its romantic primitivism and in its rationalism as tempered by John Locke, "the Christian philosopher," and by the Scottish Common Sense school, which opposed the paradoxes of subtle philosophies. From Locke he took over the political theory considered self-evident in the young American nation and readily transferred it to his doctrine of the church. He also followed Locke's theory of knowledge based on sense experience, which established the grounds for Christian faith in historical events and objective evidence (recorded in Scripture) rather than in mysticism or subjective religious "experience." He therefore repudiated the Calvinist (and revivalist) concept of miraculous conversion and the similar concept of miraculous call to the ministry. Debates on these issues, as well as on the damnation of unbaptized infants, which Disciples denied, led them to think of themselves as anti-Calvinist.

Legacy of John Locke

The general framework of their thought nevertheless followed Reformed (Calvinist) outlines, modified by the influence of British Independents (the originally Scottish Glasites—or Sandemanians—in practice a strictly New Testament sect, and the Congregationalists). Disciples shared the orthodox Protestant emphasis on the authority of Scripture. Their classic biblical position differs from that of other Protestants in being a product of the early 19th rather than of the 16th or 17th century.

Characteristic doctrines. Early Disciples understood their uniqueness to lie in the rigour, precision, and simplicity with which they set forth the biblical basis for the unity of all Christians. Campbell distinguished sharply between Old and New Covenants (Testaments), limiting to the latter any authority for "the original faith and order" of the church. Only explicit apostolic teaching or precedent belonged in the realm of faith, of the essential; all else, however logical or helpful, fell in the area of opinion and consequently of Christian liberty. Thus they rejected creeds as tests of fellowship; they believed such tests usurped the sole authority of the New Testament and set forth demands not found there. The popular Disciple bias against theology as a divisive preoccupation with human opinions—as well as Alexander Campbell's early protest against ecclesiastical institutions as unwarranted by Scripture and threatening to freedom—also was inferred from the New Testament.

Standards of apostolic teaching

Campbell saw the biblically authorized ministry as that of elders and deacons, ordained by the congregations, and of evangelists, who served the church at large. The latter provided the general leadership in his day; he succeeded in establishing no extracongregational procedure for their authorization and discipline.

He regarded immersion and "the breaking of bread" (*i.e.*, Baptism and communion) as ordinances of the Lord (Christ). While the insistence on believer's Baptism alone separated Disciples from the "paedobaptists" (those advocating Baptism of children), weekly communion served as a universal element in their worship and tempered their rationalist bent. Despite their memorialist doctrine (that communion is a commemoration of Christ's Last Supper involving no miracle of transubstantiation), they understood the service as present communion with their Lord.

Internal differences. The divisions in the movement expressed varying attitudes toward Scripture, as the norm of faith and practice: Churches of Christ construing it strictly, Disciples more loosely. Many who introduced organs in worship held the same view of biblical authority as those who refused to do so; their interpretation simply led to a different conclusion about the use of musical instruments in apostolic times. They provided the constituency for the "independent" Christian Churches, whereas Disciples tended to find more and more flexibility in the principle of expediency.

Controversies in biblical interpretation

In the world of scholarship since the early 19th century, a revolution has occurred in the understanding of the biblical documents and the nature of their authority. In general, the Churches of Christ hold steadfastly to older views of the Scripture. Disciple leaders tend to accept the results of critical scholarship. While the "biblical theology" that flourished in the second quarter of the 20th century enabled some Disciples to work out a contemporary formulation of their tradition within the ecumenical context, the conservative biblical position of their forefathers has largely been eroded. Of the three groups, Disciples have been most influenced by the theological movements of the 20th century; *i.e.*, biblical criticism, liberalism, existentialism, ecumenism, neoliberalism. The independents have been much less influenced, the Churches of Christ least of all.

Recent theological trends. At the beginning of the 20th century, the most influential Disciple scholar was J.W. McGarvey, a champion of the traditional doctrines and view of the Bible and an opponent of the musical instrument in worship. Early in the century Herbert L. Willett, E.S. Ames, and C.C. Morrison led in a liberal reformulation of the plea, emphasizing a pragmatic and reasonable approach to faith, the repudiation of creeds, an openness to the scientific world view, and a commitment to Christian unity. The neo-orthodoxy that dominated Christian thought at midcentury had less appeal to most Disciples, but a leading English theologian, William Robinson, gained attention for neo-orthodoxy's emphasis on biblical doctrine. A Panel of Scholars appointed by two of the national agencies published three volumes of papers in 1963 showing the impact of the new mood. With the rapid growth of their theological seminaries and college faculties of religion and with their deep ecumenical involvement, Disciples went through a theological renaissance in the 1950s.

Early 20th-century liberal leaders

The institutional developments leading to restructuring were accompanied by a reformulation of the old doctrine of the church. The founders of the early movement had spoken of the Church of Christ as a local congregation; they recognized no other organization as a church. The new generation of Disciples could no longer deny the churchly character of the institutions that had been developed. They began to speak of three manifestations of the Christian Church—congregational, regional, general (U.S. and Canada). The name that they adopted in restructuring—the Christian Church (Disciples of Christ)—they found to have been dictated by their history. They saw that church manifesting itself organizationally "within the universal body of Christ" and as committed to "responsible ecumenical relationships."

Worship. After fruitless attempts to derive a stated order of worship from the New Testament, Disciples settled into an informal but relatively stable pattern composed of hymns, extemporaneous prayers, Scripture, sermon, and breaking of bread. Except for its omission of the Decalogue, the public confession of sin, and the

creed, it resembled classic Reformed (or Presbyterian) worship, especially in its austerity of spirit. In the second half of the 19th century it took over more of the mood of popular revivalism, which still prevails among Churches of Christ and the independent Christian Churches.

Because many churches in the 19th century had the services of a preacher only occasionally but regularly observed the Lord's Supper (communion) after the Bible School (Sunday School) hour, the breaking of bread came to precede the sermon, which was simply added on when a preacher was present. At the Lord's Table, two local elders presided, one offering a prayer of thanksgiving for the bread and the other for the cup. The minister now commonly presides, but the elders ordinarily offer these prayers.

Christian Worship: A Service Book (1953), edited by G. Edwin Osborn, a semi-official manual for voluntary use, exerted wide influence in restoring and stabilizing the typical pattern, with an emphasis on much use of scriptural sentences. More recently Disciples have responded to the liturgical movement, with greater use of responsive readings (minister and people reciting alternate verses), affirmations of faith, and accommodation to the historic pattern of the Christian liturgy.

Social issues. On social questions Disciples have held positions characteristic of the American denominations of English background. Alexander Campbell applied his principles literally in the matter of slavery, admitting that Scripture and the law of the land permitted the practice, though as a matter of opinion he favoured emancipation. By this line he prevented schism. During the Civil War a number of leading Disciples, especially in the Border States, espoused pacifism on biblical grounds.

Liberal Disciples took up the Social Gospel (a theological movement that stressed the social this-worldly aspects of Jesus' teachings) early in the 20th century, whereas most Disciples, along with members of the Churches of Christ and the independent Christian Churches, held to the emphasis on personal salvation. Disciple representatives to the National Council of Churches and the World Council of Churches have supported those organizations' general stand on social issues but resisted the tendency to accept the new radicalism of the later 1960s.

BIBLIOGRAPHY

History: WINFRED ERNEST GARRISON and ALFRED T. DEGROOT, *The Disciples of Christ: A History*, rev. ed. (1958), a comprehensive scholarly study; WILLIAM GARRETT WEST, *Barton Warren Stone: Early American Advocate of Christian Unity* (1954); LESTER G. MCALLISTER, *Thomas Campbell: Man of the Book* (1954); EVA JEAN WRATHER, *Creative Freedom in Action: Alexander Campbell on the Structure of the Church* (1968).

Beliefs, worship, and organization: ALEXANDER CAMPBELL, *The Christian System . . .* (1835, later reprints), the classic summary of Campbell's theology; ROYAL HUMBERT (ed.), *A Compend of Alexander Campbell's Theology* (1961); ROBERT O. FIFE, DAVID EDWIN HARRELL, JR., and RONALD E. OSBORN, *Disciples and the Church Universal* (1967), statements by representatives of the three major groups within the tradition; WILLIAM ROBINSON, *The Biblical Doctrine of the Church* (1948); WINFRED ERNEST GARRISON, *Heritage and Destiny: An American Religious Movement Looks Ahead* (1961); KEITH WATKINS, *The Breaking of Bread: An Approach to Worship for the Christian Churches (Disciples of Christ)* (1966), a historical and theological analysis; DAVID EDWIN HARRELL, JR., *Quest for a Christian America: The Disciples and American Society to 1866* (1966).

(R.E.O.)

Disease

Disease commonly is considered to be a departure from the normal physiological state of a living organism sufficient to produce overt signs, or symptoms. The concept applies to the mental, as well as the physical, state of the organism; thus, it is customary to speak of "mental" disease when referring to deranged thought processes.

The normal state of an organism represents a condition of delicate physiological balance, or homeostasis (*q.v.*), in terms of chemical, physical, and functional processes, maintained by a complex of mechanisms that are not fully understood in their entirety. In a fundamental sense, therefore, disease represents the consequences of a breakdown of the homeostatic control mechanisms. In some instances the affected mechanisms are clearly indicated, but, in most cases, a complex of mechanisms is disturbed, initially or sequentially, and precise definition of the pathogenesis of the ensuing disease is elusive. Death in man, and other mammals, for example, often results directly from heart or lung failure, but the preceding sequence of events may be highly complex, involving disturbances of other organ systems and derangement of other control mechanisms.

Disruption of physiological balance

The initial cause of the diseased state may lie within the individual organism itself, and the disease is then said to be idiopathic, innate, primary, or "essential." It may result from a course of medical treatment, either as an unavoidable side effect or because the treatment itself was ill-advised; in either case the disease is classed as iatrogenic. Finally, the disease may be caused by some agent external to the organism. This may be an inert but toxic agent, in which case the disease is noncommunicable; that is, it affects only the individual organism contacting it. The external agent may be itself a living organism capable of multiplying within the host and subsequently infecting other organisms; in this case the disease is said to be communicable.

NONCOMMUNICABLE DISEASE

Metabolic defects. Noncommunicable diseases arise from metabolic faults present at birth that leave the organism ill-equipped to deal with the natural materials it encounters in its daily life. In man, for example, the lack of a certain enzyme necessary for the metabolism of the common amino acid phenylalanine leads to the disease phenylketonuria (or PKU), which is often associated with mental retardation. Other metabolic defects may make their appearance only relatively late in life. Examples of this situation are the diseases gout and diabetes in man. Gout results from an accumulation within the tissues of uric acid, an end product of nucleic acid metabolism; and diabetes results from an impairment of the synthesis of insulin by the pancreas and a consequent inability to metabolize sugars and fats properly. Alternatively, the metabolic fault may be associated with aging and the concomitant deterioration of control mechanisms, as in the loss of calcium from bone in the condition known as osteoporosis. That these late-developing metabolic diseases have a genetic basis—that is, that there is an inherited tendency for the development of the metabolic faults involved—seems to be definitely indicated in some instances but remains uncertain in others.

Environmental hazards. Metabolic derangements also may result from the effects of external environmental factors, and this relationship may be suggested by the apparent confinement of a disease to sharply delimited geographic areas. Notable examples are goitre and mottled enamel of the teeth in man. The development of goitre is attributable to iodine deficiency in the diet, which leads to compensatory growth of the thyroid gland in a vain effort to overcome the deficiency. In the absence of deliberate inclusion of iodine in the diet in such standard items as table salt, the disease tends to occur in inland areas where seafood consumption is minimal. Mottled enamel of teeth is observed to result from consumption of excessive amounts of fluoride, usually in water supplies, but conversely, dental caries is found to occur to a greater extent in areas with water supplies deficient in fluoride. Analogous conditions in herbivorous domestic animals result from deficiencies in the so-called trace elements, such as zinc, in the soil of pastures and, therefore, also in plants making up the diet. Similarly, plant growth suffers from deficiencies in the soil of essential elements, particularly nitrogen, potassium, and phosphorus. The correction of these conditions by supplying salts to domestic animals, such as cattle, and the application of fertilizers to soil, is well known.

Deficiency diseases

These, generally, are diseases of deficiency; there also are diseases resulting from toxic substances added to the environment in sufficient amount to produce symptoms of greater or lesser severity. Although best known in man, untoward effects of such contamination of the environment occur also in plants and animals. The problems caused by environmental toxic agents are largely, if not entirely, man-made. Best known of the environmental diseases, perhaps, are the occupational diseases, especially those of the respiratory tract, including asbestosis, silicosis, and byssinosis (caused, respectively, by inhalation of asbestos, silica, and cotton dust). Also important in this regard are metal poisoning, as with mercury, lead, and arsenic, poisoning with solvents used in industrial processes, and exposure to ionizing radiation. Of more importance to the population at large are the diseases that result from exposure to insecticides and atmospheric pollutants. Such diseases usually, though not invariably, are of a chronic nature, requiring prolonged exposure to the noxious agent and developing slowly. Environmental diseases of all kinds, however, also may predispose the individual to other diseases, notably tuberculosis in the case of respiratory diseases such as silicosis, and perhaps, in the longer view, to cancer.

COMMUNICABLE DISEASE

Host–parasite relationships. Communicable, or contagious, diseases are diseases transmitted from one organism to another; infectious diseases are diseases caused in the host by infection with living, and therefore replicating, micro-organisms such as animal parasites, bacteria, fungi, or viruses. Practically, these two classes of disease are the same, for infectious diseases generally are communicable, or transmissible, from one host to another, and the causative agent, therefore, is disseminated, directly or indirectly, through the host population. Such spread is an ecological phenomenon, the host serving as the environment in which the parasite lives; it becomes complex when the parasite occurs in more than one host species. The host–parasite relationship, therefore, must be considered not only with respect to the individual host–parasite interaction but also in terms of the interrelationship between the host and parasite populations, as well as those of any other host species involved.

Disease as an ecological problem

The degrees of dependence of the parasite on the host form a continuous spectrum, ranging from the almost complete independence of free-living forms that survive on decaying matter (are saprophytic) and only incidentally to their own life history may produce disease in a host, to that of the highly adapted obligate intracellular parasites, which are capable of undergoing reproduction only within the cells of the host species. The causative agent of botulism in birds, and in mammals (including man), is a saprophytic inhabitant of the soil that is unable to invade host tissue. When it grows in foods, it produces a potent toxin, and disease in the host is a consequence of the ingestion of preformed toxin. The tetanus and gaseous gangrene bacilli also are essentially saprophytes, but they may live as "passengers," or commensals, in the large bowel of animals. There they are capable of setting up circumscribed foci of infection in previously injured host tissue, and disease results from the absorption of toxins produced during the growth of the micro-organisms.

The majority of the pathogenic bacteria are obligate parasites; that is, they are found only in association with their hosts. Some, such as staphylococci and streptococci, can proliferate outside the body of the host in nutritive materials infected from host sources. Within the tissues of the host, these organisms set up local infections that tend to spread throughout the body. Still other bacteria, such as the glanders bacillus, and the gonococci, meningococci, and pneumococci, are more closely adapted parasites, multiplying outside the body of the host only under the artificial conditions of the laboratory. All of these micro-organisms have complete cell structures and metabolic capabilities.

Pathogenic bacteria

A greater degree of dependence upon the host is shown by the micro-organisms known as rickettsiae and viruses. Rickettsiae have the cell structure of bacteria, and they exhibit a small degree of metabolic activity outside of cells, but they cannot grow in the absence of host tissue. The ultimate in parasitism, however, is that of the viruses, which have no conventional cell structure at all and consist only of nuclear material wrapped in a protective protein coat. Viruses are obligatory intracellular parasites, capable of multiplying only within the cells of the host, and they have no independent metabolic activity of their own. The viral nuclear material, containing directive information for the synthesis of virus materials and certain enzymes, enters the host cell, parasitizes its chemical processes, and directs them toward the synthesis of virus substances.

The occurrence of these various degrees of parasitism suggests that the host–parasite relationship is subject to continuing evolutionary change. The adaptation of the micro-organism to its parasitic existence, in this view, is accompanied by progressive loss in metabolic capability, with eventual complete physiological dependence of the parasite upon the host, as is the case with viruses.

Parasite specificity. The condition of obligate parasitism is associated with a degree of specificity of the parasite with regard to the host—that is, the parasite generally is more closely adapted to one species of host than to all others. This may represent the development of dependence of the parasite upon the chemistry of the host species, and it also may represent an adaptation of the host species to infection by the parasite. Micro-organisms adapted to plant hosts, with only rare exception, are unable to infect animal hosts, and conversely micro-organism parasites of animals rarely occur in plants. A number of host species may be susceptible to infection with a given parasite, and the pattern of host susceptibility need not correspond with taxonomic relationships, including, for example, hosts varying as widely as vertebrates and invertebrates.

The ability to produce consistently highly fatal disease in a host is clearly of negative survival value to the parasite, because it is quite likely to eliminate quickly all available hosts. Consistent with this, there is a tendency for disease resulting from infection to be less severe when adaptation of the parasite to the host has become close. A change in severity of a disease, presumably resulting from adaptation, has been observed in the case of the spirochete that causes syphilis, the disease in man being less severe today than it was in the 16th century.

Adaptation of parasite to host

Disease produced in related host species may be either milder or more severe than in the definitive host. In certain cases, adaptation is so close that the parasite is unable to infect any other hosts under natural conditions; this is true of many micro-organisms producing disease in man. On the other hand, natural infection of secondary hosts may occur, leading to severe or fatal disease. Rabies, for example, is a highly fatal disease in almost all animal hosts, but the virus persists in bats, presumably the natural host, as an asymptomatic infection. Similarly, spotted fever rickettsiae, which cause severe and sometimes fatal disease in man, appear to be so closely adapted to their tick host as to be congenitally transmitted. In this host they constitute a permanent, and apparently harmless, infection.

Host resistance. The specificity of pathogenic micro-organisms with regard to their hosts is an expression not only of differences in microbial character but of differing host resistance. The ability of a micro-organism to produce disease can be evaluated only in terms of the host reaction, and conversely the resistance, or immunity, of the host can be judged only with regard to its effect on the micro-organism. In short, the two are but different facets of the same phenomenon, and either may be evaluated by holding the other constant and varying it. Commonly, for example, virulence of an infective agent is determined experimentally by inoculating groups of hosts with graded doses of the agent and determining, by interpolation, the dose that produces a typical reaction in 50 percent of the host individuals inoculated. This dose is termed the 50-percent-effective dose, or ED_{50}; it is related in inverse fashion to virulence and in a direct way to re-

sistance. In other words, in a given host, the higher the 50-percent-effective dose, the less virulent the infective organism; or with a micro-organism of known virulence, the higher the ED_{50} with the host it is tested against, the greater the resistance of that particular host. Customarily, in different host species, resistance is expressed as an *n*-fold increase or decrease (with *n* equal to a whole number) in the ED_{50} over that of the normal host species.

This kind of assay is possible because both virulence and resistance tend to occur in approximately normal, or bell-shaped, frequency distributions; that is, most members of the host and micro-organism populations occupy a central position with regard to these properties, exceptional individuals appearing at both extremes. With reference to host resistance, this explains the varied incidence of disease in a host population exposed to a statistically constant dose of the infectious agent. In most practical considerations, of course, the dose is only statistically constant, for it varies greatly from one host to another depending upon circumstances relating to transfer of the infectious agent. Individual variation in host resistance to infection, however, is due to more than mere numbers of infectious agents encountered; it is also clearly due to innate factors in the individual host organism. At any rate, the result of variation in host resistance is that not all individuals making up a population essentially universally susceptible to infection with a newly appearing infectious agent (such as the Asian strain of the influenza virus in 1957) will contract the disease on primary exposure.

Distribution of resistance and virulence

Apparent and inapparent infection. Because infection is not an all-or-none affair, individual variation in resistance to disease also results in different degrees of reaction to the infectious agent; *i.e.*, the outcome of the interaction of host and parasite is variable in each individual instance. Some individual hosts show symptoms typical of the disease; they are readily recognized. Others, having greater resistance, exhibit symptoms of the disease in only a mild or atypical form, and these individuals are not clearly recognizable. Still other host organisms become infected with the invading parasite but show no symptoms of the disease. Distinction, therefore, must be made between infection and disease, the former occurring on occasion without any sign of the latter. There may be, of course, no such thing as totally symptomless infections. What are taken to be such may be, in fact, only those infections with symptoms occurring beneath the level of observation. Nonetheless, such inapparent infections, or "carrier" states, clearly exist and serve to transmit the infection to susceptible hosts.

The overt consequence of infection of a host population of relatively high resistance is the sporadic occurrence of cases of disease, and a high carrier–case ratio. The infection, in other words, is widely prevalent in the host population in asymptomatic form, and the relatively rare, observed cases of disease represent the highly susceptible few in the host population making up the extremes of the bell-shaped frequency distribution curve. Examples of human diseases of this kind are poliomyelitis, meningococcal meningitis, and cholera.

This type of irregularity in the occurrence of cases of disease tends to occur in host populations of high, but not too high, resistance to the infectious agent. If host resistance is too high, or too low, the disease will die out: in the former case, because the infective agent is unable to maintain itself and, in the latter, because it eliminates the host. One of the best known illustrations of the importance of relative host resistance to survival of the parasite is that of the plague bacillus. Plague is primarily a disease of rodents and persists as foci of infection in these hosts. The black rat and the less susceptible gray sewer rat are commonly associated with this disease but are too susceptible to allow its persistence; *i.e.*, the host is destroyed. The infection persists, however, in relatively resistant wild rodents. In Kurdistan and adjacent areas four species of gerbils are involved, two resistant and two highly susceptible species, and the parasite population is maintained in permanent foci through the infection and interaction of these rodent populations. In Mesopotamia and India, foci of infection persist for periods of up to several years but do not become permanent because the wild rodent population, also species of gerbils, in which the disease occurs is not sufficiently resistant. Plague continues to occur in these areas through reintroduction of the parasite from permanent foci of infection elsewhere, usually by means of the highly susceptible rat, which transmits the infection to the indigenous gerbil population.

Foci of infection

Inheritance of resistance. That there exists genetic control of resistance is suggested by the mere fact of host specificity, and such control has been demonstrated amply by experimental studies on both plant and animal hosts. The former, for example, had wide practical application in the development, by selective breeding, of strains and races of plants of economic importance, especially grains, that are resistant to a wide variety of plant diseases. Although similar studies on animal hosts have not had practical application, intensive investigation of lower animals has shown that resistance, or susceptibility, unquestionably is determined genetically.

In general, resistance developed by selective breeding is only partially specific; that is, the observed resistance to infection with pathogenic micro-organisms, and to the toxins of such organisms, is manifested toward groups of related micro-organisms producing similar diseases, and not to single organisms alone. Although resistance to disease has been found in a few instances to be a function of a single gene, in the majority of cases, several genes appear to be involved—as might be expected from the complex nature of resistance.

For many years there has been considerable interest in the possibility of differences in resistance to disease associated with the different races of man. While marked differences in morbidity and mortality occur between whites and nonwhites in the United States, for example, it is often difficult to rule out differences in exposure to infection, socio-economic factors, and differential application of preventive and therapeutic measures in accounting for them. Nevertheless, there are fragmentary indications that there may be sufficient genetic segregation among races of man to give differences in resistance to certain diseases. The case fatality rate in tuberculosis appears to be lower in Jews than in others, for example, and gonorrhea seems to be a less serious disease in the Negro than in the Caucasian.

Racial differences in man

Epidemiology. The interaction of host and parasite populations constitutes the subject matter of epidemiology (the term being more inclusive than suggested by its relation to the word epidemic). In most instances the epidemiology of infectious disease is characteristic of that disease and is an outgrowth of biological properties of the parasite and the host, including host specificity and the behaviour of the host populations as populations.

Aside from the saprophytic micro-organisms that occasionally produce disease, most pathogenic micro-organisms are adapted sufficiently closely to their hosts that they cannot compete successfully in the physical, chemical, and biological environment outside the host tissues. Exceptions to this generalization occur in the cases of those micro-organisms whose life history includes a resistant spore stage. This occurs with various fungi responsible for plant disease, as well as certain parasites of animals. Among the latter are species of the fungus *Coccidioides*, which infect both rodents and man (producing desert fever in man), and the anthrax bacillus, which causes disease in cattle, sheep, and other domestic animals, and occasionally infects man as well.

Survival of the parasite ordinarily requires transmission from an infected host organism to a susceptible one more or less directly. The precise route of infection often assumes primary importance, some micro-organisms requiring direct access to internal tissue, others being able to initiate infection on mucous membranes of the nose and throat, and still others able to establish primary infections in the intestinal tract. These particular modes of infection then generally demand transmission by way of, respectively, biting insects, coughing and sneezing, and contamination of food and water.

The occurrence of a given parasite in more than one host species may markedly affect its epidemiological character; it may persist, for example, in one or another of its hosts as a reservoir of infection, sallying forth to encounter the alternate host only on occasion. It is fairly common to find transmission of a parasite from one vertebrate or plant host to another by means of an insect carrier, or vector. Often animal parasites have intermediate hosts in which one or more phases of their life cycles occur; this results in an obligatory sequence of hosts in the life history of the parasite. With the disease schistosomiasis in man, for example, the blood flukes responsible for the disease spend one phase of their larval life in snails. Under such circumstances, and they are not uncommon, the dissemination of the parasite in a host population is dependent not only upon the interaction of the parasite population with that of the host population but also upon the interaction of the intermediate or vector host population with both parasite and host populations.

The role of vector hosts

Such interrelationships may be the basis of geographical and seasonal differences in the incidence of disease. An insect-borne disease transmitted from one host species to another requires the simultaneous presence of all three populations in sufficient numbers for its dissemination. Such a circumstance may be sharply limited by location and season.

Behavioral patterns of host populations often nave a very great effect on the transmission of infectious agents. Crowding, for example, facilitates the spread of infection. Bovine tuberculosis is largely a disease of domesticated cattle in barns, and the age incidence of the human diseases of childhood is lower in urban than in rural populations, suggesting that in the more crowded urban environment children are exposed to disease at an earlier age. That transmission of human childhood diseases is favoured by aggregation also is indicated by the prevalence of these diseases during the times when children are in school.

When a disease is prevalent in an area over long periods of time, it is considered to be endemic in that area. On the other hand, when the prevalence of disease is subject to wide fluctuations in time, it is considered to be epidemic during periods of high prevalence. Epidemics prevailing over wide geographical areas are called pandemics.

Epidemic prevalence of disease occurs in a wave, the number of cases rising to a peak and then declining. The period of rise occurs when each case gives rise to more than one additional case; *i.e.*, when the parasite population is growing more rapidly than the host population; and the decline occurs when each case gives rise to less than one new case; *i.e.*, when the parasite population begins to die off because it encounters only immune individuals among the host population. The rise and fall in epidemic prevalence of a disease is a probability phenomenon—the probability being that of transfer of an effective dose of the infectious agent from the infected individual to a susceptible one. After an epidemic wave has subsided, the affected host population contains a sufficiently small proportion of susceptible individuals that reintroduction of the infection will not result in a new epidemic. Since the parasite population cannot reproduce itself in such a host population, that host population as a whole is immune to the epidemic disease, a phenomenon termed "herd" immunity.

Herd immunity

Following such an epidemic, however, the host population immediately tends to revert to a condition of susceptibility because of (1) the deterioration of individual immunity, (2) the removal of immune individuals by death, and (3) the influx of susceptible individuals by birth. In time the population as a whole again reaches the point where it is susceptible to epidemic disease. This pattern of rising and falling herd immunity explains why epidemic diseases tend to occur in successive waves; *i.e.*, exhibit a periodicity in prevalence. The time elapsing between successive epidemic peaks is variable and differs from one disease to another. In human measles uncontrolled by immunization, the interval between periods of epidemic prevalence is about two years and in meningococcal meningitis it is seven to eight years.

Immunity. An animal host reacts to the presence of parasites within its tissues by the formation of antibodies to the parasite and its products. Effective immunity to disease is a function of antibodies directed toward some, though not all, of the chemical substances (antigens) from the parasite. In general, such immunity is dual in nature, being made up of components active against the organism itself and against the toxins it produces. The relative importance of the two types differs from one kind of micro-organism to another. In those diseases that result largely from the action of microbial toxins, such as tetanus and diphtheria, the antitoxic element is of primary importance and the antimicrobial immune response is more or less irrelevant. When toxins are not formed by the micro-organism, or are of a low order of potency, antibodies to the cell substance of the micro-organism are predominant in effecting immunity to disease. The antimicrobial antibodies function in part by coating the microbial cells in such a way as to facilitate their destruction by the white blood cells of the host.

Although the immune mechanism is primarily defensive in nature, it may contribute in some cases to the toxic course of the disease. In rheumatic fever, for example, sensitivity to the causative streptococcus organism is associated with symptoms of the disease. The immune response to various environmental substances, including materials such as plant pollens and chemotherapeutic drugs, also is responsible for the diseases grouped under the general head of allergies. The immune response itself may become abnormal in such human diseases as multiple myeloma, macroglobulinemia, Hodgkin's disease, chronic lymphocytic leukemia, and infectious mononucleosis. Such abnormal immune responses, however, seem to be of minor significance to the course of these diseases.

One category of disease is associated with an immune response to antigenic components of the host itself (autoantigens). The occurrence of such diseases, which are called auto-immune diseases, is firmly established on a laboratory basis. Diseases produced by the inoculation of experimental animals with extracts of their own organs and tissues include encephalitis, thyroiditis, hemolytic anemia, polyarthritis, myocarditis, and dermatitis. These experimental diseases mimic, to a greater or lesser degree, certain human diseases, the autoimmune nature of which is strongly implied, though unequivocal proof of this relationship is not available. Diseases of this kind include rheumatoid arthritis, systemic lupus erythematosus, polymyositis, thyroiditis, encephalopathy, encephalomyelitis, and multiple sclerosis. Certain of these diseases tend to merge into one another, and in all of them there is some evidence of the participation of immunological reactions in their pathogenesis. Generally, therapy with corticosteroids, which are known to suppress the immune response, alleviates their symptoms.

Auto-immune disease

CONTROL OF DISEASE

Prevention. Most diseases are preventable to a greater or lesser degree, the chief exceptions being the idiopathic diseases, such as metabolic defects. In the case of those diseases resulting from environmental factors, prevention is a matter of eliminating, or sharply reducing, the responsible material in the environment. Because these materials originate largely from human activities, prevention ought to be a simple matter of the application of well-established principles of industrial hygiene. In practice, however, this may be difficult to achieve.

The infectious diseases may be prevented in one or the other of two general ways: (1) by preventing contact, and therefore transmission of infection, between the susceptible host and the source of infection; and (2) by rendering the host unsusceptible, either by selective breeding, or by induction of an effective artificial immunity. The nature of the specific preventive measures, and their efficacy, varies from one disease to another.

Quarantine, a theoretically effective method of preventing transmission of disease in principle, has had only

Quarantine

limited success in practice. It has served to control rabies in Britain and to exclude foot-and-mouth disease from the United States. Rigorous application of quarantine procedures also effectively prevented the occurrence of cholera in Japan in one of the great pandemics of this disease. Other than these instances, however, quarantine has not achieved prevention of the spread of disease across international borders, and quarantine of individual cases of disease in man has long been abandoned as ineffective.

It has not been possible to prevent effectively the dissemination of airborne disease, notably airborne fungus diseases of plants and human diseases of the upper respiratory tract. Nor is disease ordinarily controllable by elimination of reservoirs of infection, such as those that occur in wild animals. There are certain exceptions in which the reservoir of infection can be greatly reduced, however; for example, chemotherapy of human tuberculosis may render individual cases noninfectious, and slaughtering of infected cattle may reduce the incidence of bovine tuberculosis.

When infection is spread less directly, through the agency of living vectors or inanimate vehicles, it is often possible to break one or more of the links connecting the susceptible host with the source of infection. Malaria, for instance, can be controlled effectively by elimination of the mosquito vector, and louse-borne typhus in man can be regulated by disinfestation methods. Similarly, diseases spread in epidemic form through the agency of water or milk are controlled by measures such as the chlorination of public water supplies and the pasteurization of milk.

Artificial immunization

Artificial immunization against certain diseases provides a solid immunity and may be utilized in these instances, particularly when other methods of control are impractical or ineffective. Mass immunization in childhood has been highly effective in the control of diphtheria, smallpox, and poliomyelitis, and, to a somewhat lesser extent, whooping cough. Under special circumstances, as in certain military populations, it has been possible to control with prophylactic medicinal agents the spread of disease for which effective vaccines have not been developed.

Treatment. Treatment of disease in the affected individual is twofold in nature, being directed (1) toward restoration of a normal physiological state and (2) toward removal of the causative agent. The diseased organism itself plays an active part in both respects, having the capacity for tissue proliferation to replace damaged tissue and to surround and wall off the noxious agent, as well as defense and detoxification mechanisms that remove the causative agent and its products or render them harmless. Therapy of disease supplements and reinforces these natural defense mechanisms.

Metabolic faults also may sometimes be corrected—for example, by the use of insulin in the treatment and control of diabetes—but more often specific therapeutic measures for idiopathic diseases are lacking. When disease is produced by environmental factors, there is commonly no specific treatment; removal of the affected individual from exposure to the agent generally allows normal detoxification responses to take over. Again, there are notable exceptions, as in the treatment of lead poisoning with ethylenediaminetetraacetic acid, an agent that forms complexes with lead that are excreted by the kidney. Similarly, mercury poisoning, and less effectively arsenic poisoning, may be treated with the substance dimercaprol, which forms inactive complexes with the metal atoms.

Treatment of infectious diseases is more effective in general; it assumes several different forms. Treatment of diphtheria with antitoxin, for example, neutralizes the toxin formed by the micro-organisms, and host defense mechanisms then rid the body of the causative micro-organisms. In other diseases, treatment is symptomatic in the sense of restoring normal body function. An outstanding example of this is in cholera, in which disease symptoms result from a massive loss of fluid and salts and from a metabolic acidosis; the highly effective treatment consists of restoring water and salts, the latter including bicarbonates or lactates to combat acidosis. More often, however, therapy is directed against the infecting micro-organism by administration of drugs such as sulfonamides or antibiotics. These substances do not kill micro-organisms in the host but inhibit their proliferation and give host defenses an opportunity to function effectively. For other infectious diseases there is no specific therapy. There are, for example, very few antiviral chemotherapeutic agents; treatment of virus diseases is solely symptomatic in that it is directed toward relief of discomfort and pain, and recovery, if it ensues, is largely a matter of an effective defense against the invading virus by the host.

BIBLIOGRAPHY. H. SMITH and J. TAYLOR (eds.), *Microbial Behavior in Vivo and in Vitro* (1964), an exhaustive and informative account of the present status of knowledge of parasites in the host–parasite relationship; M. SAMTER and H.L. ALEXANDER (eds.), *Immunological Diseases* (1965), the first volume to cover all aspects of the poorly known collagen diseases; *Cecil-Loeb Textbook of Medicine*, 13th ed., by P.B. BEESON and W. MCDERMOTT (1971), the classic textbook on human diseases of all kinds, including etiology, pathogenesis, prevention, and therapy; *Defense Reactions in Invertebrates: A Symposium*, vol. 26 of the Proceedings of American Societies for Experimental Biology (1967), a symposium by various experts giving up-to-date information on all aspects of the response of the invertebrate host in the host–parasite relationship; F.M. BURNET, *The Natural History of Infectious Disease*, 3rd ed. (1962), a unique consideration of infectious disease as an ecological and evolutionary phenomenon; R.N. GOODMAN, Z. KIRALY, and M. ZAITLIN, *The Biochemistry and Physiology of Infectious Plant Disease* (1967), a modern and authoritative consideration of diseases of plants; I.A. MERCHANT and R.A. PACKER, *Veterinary Bacteriology and Virology*, 7th ed. (1967), a useful and thorough discussion of the diseases of animals of interest to man.

(W.Bu.)

Disease, Human

This article presents a general consideration of human disease and of necessity deals with health and its maintenance. A survey of the various forms of human diseases and their classification is presented, to provide some understanding of the complexities involved in the maintenance of health and the scope of the diseases that can affect it.

This article is divided into the following sections:

I. Maintenance of health
  Homeostasis
  Adaptation
  Defense against biotic invasion
  Repair and regeneration
  Hemostasis
  Interrelationship of defensive mechanisms
II. Disease: signs and symptoms
III. The causes of disease
  Diseases of genetic origin
  Congenital malformations
  Heredity and environment
  Chemical and physical injury
  Diseases of immune origin
  Diseases of biotic origin
  Abnormal growth of cells
  Diseases of metabolic–endocrine origin
  Diseases of nutrition
  Diseases of psychogenic origin
  Diseases of senescence
IV. Classifications of diseases

Before human disease can be discussed, the meanings of the terms health, physical fitness, illness, and disease must be considered.

The meaning of health

Health could be defined theoretically in terms of certain measured values; for example, a person having normal body temperature, pulse and breathing rates, blood pressure, height, weight, acuity of vision, sensitivity of hearing, and other normal measurable characteristics might be termed healthy. But what does normal mean and how is it established? It is well-known that if the temperatures are taken of a large number of active, presumably healthy, individuals the temperatures will all come close to 98.6° F (37° C). The great preponderance

of these values will fall between 98.4° F and 98.8° F. Thus health could in part be defined as having a temperature within this narrow range. Similarly, a normal range can be established for pulse, blood pressure, and height. In some healthy individuals, however, the body temperature may range below 98.4° F or above 98.8° F. These low and high temperatures fall outside the limits defined above as normal and are instances of biological variability.

Biological criteria of normality are based on statistical concepts. Body height may be used as an example. If the heights of every individual in a large sample were plotted on a graph, the many points would fall on a bell-shaped curve. At one end of the curve would be the very short people and at the other extreme the few very tall people. The majority of the points of the sample population would fall on the dome of the bell-shaped curve. At the peak of the dome would be those individuals whose height approaches the average of all the heights. Scientists use curves in determining what they call normal criteria. By accepted statistical criteria, 95 percent of the population measured would be included in the normal range, that is, 47.5 percent above and 47.5 percent below the mean at the very centre of the bell. Looked at in another way, in any given normal biologic distribution 5 percent will be considered outside the normal range. Thus the seven-foot basketball player would be considered abnormally tall, but that which is abnormal must be distinguished from that which represents disease. The basketball player might be abnormally tall but still have excellent health. Thus, in any statistical analysis of health the possibility of biological variation must be recognized.

A better example than height of how problems can arise with biological variability is heart size. If the heart is subjected to a greater than normal burden over a long period, it can respond by growing larger (the process is known as hypertrophy). This occurs in certain forms of heart disease, especially in long-standing high blood pressure. A large heart, therefore, may be a sign of disease. On the other hand, it is not uncommon for athletes to have large hearts. Continuous strenuous exercise requires a greater output of blood to the tissues, and the heart meets this demand by becoming larger. The physician may have his diagnostic abilities taxed in deciding whether an abnormally large heart represents evidence of disease or is simply a biological variant.

The effects of age

The effects of age introduce yet another difficulty in the attempt to define health in theoretical measured norms. It is well-known that muscular strength diminishes in the advanced years of life, the bones become more delicate and more easily fractured, vision and hearing become less sharp, and a variety of other changes occur. There is some basis for considering this general deterioration as a disease, but, in view of the fact that it affects virtually everyone, it can be accepted as normal. Theoretical criteria for health, then, would have to be set for virtually every year of life. Thus one would have to say that it is normal for a man of 80 to be breathless after climbing two flights of stairs, while such breathlessness would be distinctly abnormal in an agile child of ten years of age. Moreover, an individual's general level of physical activity significantly alters his ability to respond to the ordinary demands of daily life. The amount of muscular strength possessed by an 80-year-old man who has remained physically active would be considerably more than that of his fragile friend who has led a confined life because of his dislike of activity. There are, therefore, many difficulties in establishing criteria for health in terms of absolute values.

Health might better be defined as the ability to function effectively in complete harmony with one's environment. Implied in such a definition is the capability of meeting—physically, emotionally, and mentally—the ordinary stresses of life. In this definition health is interpreted in terms of the individual's environment. Health to the construction worker would have a dimension different from health to the bookkeeper. The healthy construction worker expects to be able to do manual labour all day, while the bookkeeper, although perfectly capable of performing his own sedentary work, would be totally incapable of such heavy labour and indeed might collapse from the physical strain; yet both individuals might be termed completely healthy in terms of their own way of life.

Physical fitness

The term physical fitness, although frequently used, is also exceedingly difficult to define. In general it refers to the state of optimal maintenance of muscular strength, proper function of the internal organs, and youthful vigour. The champion athlete prepared to cope not only with the commonplace stresses of life but also with the unusual, illustrates the concept of physical fitness. To be in good physical condition is to have the ability to swim a mile to save one's life or to slog home through snow drifts when a car breaks down in a storm. Some experts in fitness insist that the state of health requires that the individual be in prime physical condition. They prefer to divide the spectrum of health and disease into (1) health, (2) absence of disease, and (3) disease. In their view, those who are not in prime condition and are not physically fit cannot be considered as healthy merely because they have no disease.

Health involves more than physical fitness, since it also implies mental and emotional well-being. Should the angry, frustrated, emotionally unstable person in excellent physical condition be called healthy? Certainly he could not be characterized as effectively functioning in complete harmony with his environment. Indeed, such an individual is incapable of good judgment and rational response. Health, then, is not merely the absence of illness or disease but involves the ability to function in harmony with one's environment and to meet the usual and sometimes unusual demands of daily life.

The definitions of illness and disease are equally difficult problems. Despite the fact that these terms are often used interchangeably, illness is not to be equated with disease. A person may have a disease for many years without even being aware of its presence. Although he is diseased, he is not ill. Similarly, the diabetic person who has known disease and has received adequate insulin treatment is not ill. The cancer victim is often totally unaware of his disorder and is not ill until after long years of growth of the tumour, during which time it causes no symptoms. The term illness implies discomfort or inability to function optimally. Hence it is a subjective state of lack of well-being produced by disease. Regrettably, many diseases remain submerged for long years before they produce discomfort or impair function and, thus, escape detection and possible cure.

Illness and disease

Disease, which can be defined at the simplest level as any deviation from normal form and function, may either be associated with illness or be submerged (latent). In the latter circumstance, the disease will either become apparent at some later time or will render the individual more susceptible to illness. The person who fractures an ankle has an injury—a disease—producing immediate illness. Both form and function have been impaired. The illness occurred at the instant of the development of the injury or disease. The child who is infected with measles, on the other hand, does not become ill for approximately ten days after exposure (the incubation period). During this incubation period the child is not ill but has a viral infectious disease that is incubating and will soon produce discomfort and illness. Some diseases render a person more susceptible to illness only when he is put under stress. Some diseases may consist only of extremely subtle defects in cells that render the cells more susceptible to injury in certain situations. The blood disease known as sickle-cell anemia, for example, results from a hereditary abnormality in the production of the red pigment (hemoglobin) of the red cells of the blood. The child of a mother and father both of whom have sickle-cell anemia will probably inherit an overt form of sickle-cell anemia and will have the same disease as his parents. If only one parent has sickle-cell anemia, however, the child may inherit only a tendency to sickle-cell anemia. This tendency is referred to by physicians as the sickle-cell trait. Individuals having such a trait are not anemic but have a

greater likelihood of developing such a disease. When they climb a mountain and are exposed to lower levels of oxygen in the air, red blood cells are destroyed and anemia develops. This can serve as an example of a disease or a disease trait that renders the affected person more susceptible to illness.

Disease, defined as any deviation from normal form and function, may be trivial if the deviation is minimal. A minor skin infection might be considered trivial, for example. On the eyelid, however, such an infection could produce considerable discomfort or illness. Any departure from the state of health, then, is a disease, whether health be measured in the theoretic terms of normal measured values or in the more pragmatic terms of ability to function effectively in harmony with one's environment.

I. Maintenance of health

Health is not a static condition but represents a fluid range of physical and emotional well-being continually subjected to internal and external challenges such as worry, overwork, varying external temperatures, bacteria, and viruses. These constantly changing conditions require the adjustment of the function of the various systems within the body. Mechanisms are continually at work to maintain a constant internal environment called by the French scientist Claude Bernard the *milieu intérieur*. The maintenance of this relatively constant internal environment is known as homeostasis. On a hot summer day, for example, the body is challenged to maintain its normal temperature of 98.6° F. Sweating represents a mechanism by which the skin is kept moist. By the evaporation of the moisture, heat is more rapidly lost. The hot day, therefore, represents a challenge to homeostasis. On a cold day gooseflesh may develop, an example of a homeostatic response that is a throwback to mechanisms in lower animals. In man's furbearing ancestors, cold external environments caused the individual hair shafts to rise and, in effect, produce a heavier, thicker insulation of the body against the external chill. Man still develops this primitive gooseflesh response but regrettably does not have the luxuriant pelt to protect himself.

Definition of homeostasis

Bacteria, viruses, and other microbiologic agents are obvious challenges to health. The body is able, to a considerable extent, to protect itself and adjust to challenges, and to the extent that it is successful, the state of health is maintained. While health is often thought of as fragile and at the mercy of all the evils to which man is exposed, it is, in fact, a ruggedly guarded state protected by a host of internal mechanisms. Few machines made by man could tolerate the abuse heaped on the human body and still be capable of functioning.

Some of the mechanisms vital to the maintenance of health include (1) the maintenance of the internal environment—*i.e.*, homeostasis; (2) adaptation to stress situations; (3) defense against microbiologic agents, such as bacteria and viruses; (4) repair and regeneration of damaged tissue or cells; and (5) clotting of the blood to prevent excessive bleeding. Each of these areas will be discussed briefly. Despite these separate considerations, the commonality of purpose—the preservation and maintenance of health—must not be lost sight of. In so far as each of these mechanisms works to maintain a constant internal environment, it can be considered as a homeostatic mechanism. Later, when disease is discussed, it will be apparent that to a considerable extent disease represents a failure of homeostasis and the other defensive responses listed above.

HOMEOSTASIS

The term homeostasis refers to the maintenance of the internal environment of the body within narrow and rigidly controlled limits. The major functions important in the maintenance of homeostasis are fluid and electrolyte balance, acid–base regulation, thermoregulation, and metabolic control.

Fluid and electrolyte balance. This term refers to the controlled partition of water and major chemical constituents among the cells and the extracellular fluids of the body. The human body is basically a collection of cells grouped together into organ systems and bathed in fluids, most notably the blood. The intracellular fluid is the fluid contained within cells. The extracellular fluid—the fluid outside the cells—is divided into that found within the blood and that found outside of the blood; the latter fluid is known as the interstitial fluid. These fluids are not simply water but contain varying amounts of solutes (electrolytes and other bio-active molecules). An electrolyte (sodium chloride, for example) is defined as any molecule that in solution separates into its ionic components and is capable of conducting an electric current. Cations are electrolytes that migrate toward the negative pole of an electric field; anions migrate toward the positive pole. The electrolyte composition of the various fluid compartments is summarized in Table 1.

Intracellular and extracellular fluid

Table 1: Principal Electrolytes of the Body Fluids

extracellular fluid*		intracellular fluid†	
Cations (+ electrical charge)			
Sodium ($Na+$)	142 mEq/l‡	sodium ($Na+$)	10 mEq/l
Potassium ($K+$)	4 mEq/l	potassium ($K+$)	160 mEq/l
Calcium ($Ca++$)	5 mEq/l	magnesium ($Mg++$)	35 mEq/l
Magnesium ($Mg++$)	3 mEq/l		
Total	154 mEq/l	total	205 mEq/l
Anions (− electrical charge)			
Chloride ($Cl-$)	103 mEq/l	chloride ($Cl-$)	2 mEq/l
Bicarbonate (HCO_3-)	27 mEq/l	bicarbonate (HCO_3-)	8 mEq/l
Phosphate (PO_4^{3-})	2 mEq/l	phosphate (PO_4^{3-})	140 mEq/l
Sulfate (SO_4^{2-})	1 mEq/l		
Protein	16 mEq/l		
Organic acid	5 mEq/l	protein	55 mEq/l
Total	154 mEq/l	total	205 mEq/l

*Approximate values in the blood plasma. †Approximate values for the muscle cells. ‡mEq/l = milliequivalents per litre.

It is apparent from this table that the ionic compositions of the intracellular and extracellular fluids are significantly different. The major cation of extracellular fluid is sodium. The major anion of the extracellular fluid is chloride, while bicarbonate is the second most important. In contrast, the major cation of the intracellular fluid is potassium, and the major anions are proteins and organic phosphates. The intracellular and extracellular compartments are closely integrated and interdependent: changes in one have immediate effects on the other. It is extremely difficult to measure the ionic concentrations within cells; consequently, in clinical medicine, most measurements of electrolyte concentration are performed on the extracellular-fluid compartment, notably the blood serum. The values given in Table 1 remain fairly constant on a day-to-day basis, in spite of various dietary intakes of food and water.

It is the primary task of the kidneys to regulate the various ionic concentrations of the body. Any abnormality in these concentrations can produce serious disease; for instance, the normal sodium concentration in the serum (the blood minus its cells and clotting factors) ranges from 136 to 142 milliequivalents per litre, while the normal potassium level in the serum is kept within the narrow range of 3.5 to five milliequivalents per litre. A rise in the serum potassium to perhaps 6.2 milliequivalents per litre could cause serious abnormalities in the performance of the heart by disturbing the regularity of the nervous impulses that maintain the heart's rhythm.

The state of hydration. The total amount of body water is also maintained at fairly constant levels from day to day by the combined action of the central nervous system and the kidneys. If one were to refrain from drinking any water for a few days, the thirst centre, located in the hypothalamus deep within the brain, would send out messages that would be translated into the feeling of thirst. At the same time a hormone from the pituitary gland known as antidiuretic hormone (ADH) would be secreted. This hormone, released into the bloodstream, reaches the kidneys, where it signals the kidney to retain water and not excrete it. Should too much water be ingested, ADH secretion would be turned off, and the kidney would promptly excrete the excess amount.

Antidiuretic hormone

Acid–base equilibrium. The acidity of the body fluids is maintained within narrow limits. This acidity is expressed in terms of the pH of a solution, values exceeding 7 representing alkalinity and less than 7 acidity. The pH of a solution is an expression of the amount of hydrogen ion present. Increases in hydrogen-ion concentration cause a lowering of the pH, and, conversely, decreases in the hydrogen-ion concentration raise the pH. Any abnormal process raising the hydrogen-ion concentration in the body fluids produces a state of disease referred to as acidosis; one that causes the concentration to be lowered results in alkalosis.

In health, the blood is slightly alkaline, being kept at a pH of 7.35 to 7.45, a narrow range, the maintenance of which is required for the optimum operation of the many chemical reactions that go on constantly in the body. Alterations in the blood pH occur in many diseases, particularly of the lungs and kidneys, organs one of whose functions is the regulation of the body pH.

Thermoregulation. As has been said above, the temperature of the body is kept nearly constant at 98.6° F. Fluctuations within a few tenths of a degree are perfectly compatible with health. Wider swings in temperature are usually indicative of disease, and thus, body temperature is an important factor in assessing health. Body temperature is regulated by a thermostatic control centre in the hypothalamus. A rise in body temperature initiates a chain of events leading to an increase in the rate of breathing and in sweating, two processes that serve to lower the body temperature. Similarly, a decrease in body temperature, perhaps occasioned by a chilly winter walk, leads to increased heat-producing activity such as the muscular contractions of shivering—again mediated by the thermostatic control centre in the hypothalamus.

Metabolic control. In essence, metabolism involves all of the physical and chemical processes by which cells are produced and maintained. Included under this broad umbrella are the regulation of fluid and electrolytes, the maintenance of plasma–protein levels adequate for the building and repair of cells, and control of the amounts of sugar (glucose) and fats (lipids) in the blood so as to provide sufficient amounts for all of the energy-producing activities of the cells. (The main treatment of this subject is contained in the article METABOLISM.)

Blood-glucose level

The control of blood-glucose levels is a good example of homeostasis. Most of the glucose utilized by the body is derived from the dietary intake of various forms of sugars and starches. These are digested within the intestinal tract into the simplest forms of carbohydrate (monosaccharides). Glucose, galactose, and fructose are the principal monosaccharides. These are absorbed from the intestines into the blood and thence drain to the liver. Here all are eventually converted to glucose. The glucose may be utilized by the liver cells in part as a source of energy, but most of it enters the general circulation of the body and contributes to the blood-glucose level. Blood glucose may also be derived in times of need by the conversion of the stored glycogen into glucose.

When food is eaten there is a temporary rise in the blood-glucose level known as alimentary hyperglycemia (high blood-glucose level). Mechanisms are activated that stimulate the pancreas to secrete the hormone insulin. This hormone makes it possible for the cells of the body to utilize the glucose by facilitating its transport (carriage) across the membranes of cells into their interior, where it can enter the complex chemical reactions that ultimately provide the cell with energy. By virtue of insulin secretion, the cells receive adequate amounts of glucose, and the blood-glucose levels are returned to the normal range, somewhere between 80 and 120 milligrams per 100 millilitres of blood.

Metabolic controls are exerted similarly for fats and proteins. As will be noted later on, derangements of these controls can lead to serious disease. The state of health implies proper, smooth-running metabolic machinery.

ADAPTATION

Adaptation refers to the ability of cells to adjust to severe stresses and achieve altered states of equilibrium while preserving their state of health. The process can be likened to adjusting the air supply to an engine. If a car normally travels at 40 miles per hour it requires a certain amount of air to be mixed with the gasoline for the most effective development of power. If the car now begins to travel at 60 miles per hour, it consumes more gasoline and must have additional air. If the air valve can open and provide adequate quantities the car may function perfectly well at this higher speed. In the human body, the large bulging muscles of the man engaged in heavy labour are good examples of cellular adaptation. Because of the heavy demand on these muscles each of the individual muscle cells within the labourer's arms and legs becomes larger (hypertrophies). They become larger by the formation of increased numbers of tiny fibres (myofilaments) that provide the contractile power of muscles. Thus, while the normal cell might have 2,000 myofilaments, for example, the hypertrophied cell might have 4,000 myofilaments. The work load can now be divided among twice as many myofilaments, and no one filament is overloaded. The cells are completely normal, and, in fact, are more robust than their fragile cousins. The individual can do heavy work all day without excessive fatigue, and no cell injury results from heavy work load. A new level of equilibrium has been achieved by the process of cellular hypertrophy, a form of adaptation. A person with this type of muscular development, in fact, can be considered to be in excellent physical condition, since he is capable of meeting emergency situations, such as running from a fire or catching a train, without the dangers that might be encountered by a person who has not undergone such a development.

The enlarged muscle cell

Inhabitants of high altitudes adapt to the lowered amounts of oxygen within the air by developing an increased number of red blood cells (a condition called secondary polycythemia). The greater number of red cells in the blood are capable of absorbing more oxygen from the air breathed into the lungs, and thus, the person who lives in high altitudes makes better use of the slender oxygen content of the air.

Other examples of adaptation can be given, such as the ability of the liver cells to increase their supply of enzymes, thereby enabling them to adapt to increased levels of chemicals to which they may be exposed.

Thus adaptation is a mechanism by which the body preserves and maintains its health by adjusting to alterations in the conditions under which it functions.

DEFENSE AGAINST BIOTIC INVASION

Man swims every day of his life in a sea of various types of microbiologic (biotic) organisms. Most of the various classes and types of these organisms are entirely harmless for man, and, indeed, some play an important role in helping man as, for example, by trapping nitrogen from the air and aiding in the decomposition of waste materials. Organisms capable of producing disease are known as pathogens. The maintenance of health requires defense against such possible biotic invasion. There are four levels of defense in the body: (1) the intact skin and linings of the various orifices of the body (such as the mouth, nose, throat), (2) a widely dispersed system of cells capable of destroying invaders, (3) the capability of mounting an inflammatory reaction that destroys offenders, and (4) the capability of developing an immune response that helps to bring about further neutralization and destroy any attackers.

Four levels of defense

Maintenance of the integrity of skin and mucosal linings. With rare exception, pathogenic organisms cannot penetrate the intact covering and linings of the body. Indeed, if one were to take samples of the bacteria found on the skin, one would find large numbers of potentially harmful organisms that represent no threat unless the skin is punctured or the linings of the body are in some way injured. There are exceptions to this generalization, and a few biotic agents can probably penetrate intact mucosal surfaces. The bacteria (*Salmonella typhosa*) that cause typhoid are thought to penetrate the normal lining of the gastrointestinal tract. Nevertheless, the intact skin and mucosal linings are primary protective

barriers in the maintenance of health. For centuries, mothers have wisely admonished their children to wash a cut. Potentially harmful bacteria can be introduced into a cut, which thus provides a portal of entry for organisms that may then cause an infection. By adequate washing at least sufficient numbers of bacteria are flushed out to prevent the infection. Irritation of the skin from any cause or irritation of the throat by habitual smoking impairs the integrity of these barriers and predisposes to invasion by potentially harmful organisms. The body has ingeniously contrived to place further roadblocks in the way of invaders. The saliva and the secretions in the stomach contain enzymes and acids that also destroy organisms. Thus, man has an effective enclosing barrier that protects him against biotic attack.

Phagocytic cells of the body. The term phagocytosis means "cell eating." In biologic terms it defines the capability of certain cells within the body of engulfing particulate material. When a phagocytic cell comes in contact with some particle such as a bacterium or even inert particulate matter such as a small splinter, the cytoplasm of the cell (the cell substance outside its nucleus) flows around the object and literally incorporates it within the body of the cell. Once engulfed, the intruder is subjected to the action of destructive enzymes within the cell. If the chemical composition of the foreign substance permits degradation by the enzymes, it is destroyed. In the case of bacteria, protein-digesting enzymes (proteases) digest and thereby destroy the microorganism. Phagocytic cells abound in the body; they serve as a second line of defense against most biotic invasion.

There are two groups of phagocytic cells, white blood cells—polymorphonuclear neutrophils and monocytes—and tissue cells. The white blood cells are able to migrate through blood-vessel walls in areas of inflammation or infection, where they may phagocytize foreign material such as bacteria. Moreover, in inflammatory and infectious states, the total number of white cells in the body increases (leukocytosis). Thus the population of phagocytic cells is expanded when it is needed in the body's defense.

The second group of phagocytic cells found in the body are usually firmly fixed within tissues and are known as the reticuloendothelial system. The cells in this system go by a variety of names depending on their location (*e.g.*, Kupffer cells in the liver, macrophages or histiocytes in loose connective tissue). They are particularly abundant in the spleen, liver, lymph nodes, and bone marrow but are also scattered throughout the blood vessels and virtually all of the other tissues of the body. If, for example, bacteria do find a portal of entry, but the bacterial invasion is not too massive and the organisms are not too powerful, these phagocytic cells are capable of engulfing and destroying them before they can cause injury (see RETICULOENDOTHELIAL SYSTEM, HUMAN).

The inflammatory response. Whenever cells are damaged or destroyed in the body, a series of vascular and cellular events known as the inflammatory response is set in motion. This response is protective of health in that it destroys or walls off injurious influences and paves the way for the restoration of the state of normality. The sequence of events is as follows: in an area of injury (as in a bacterial infection), substances are released from cells that cause the small blood vessels in the affected area to become larger and thus increase the amount of blood flowing to the injured area; at the same time, clear fluid exudes out of the vessels into the area; this fluid tends to dilute any injurious substances in the area of injury; next, white corpuscles from the blood flow out of the blood vessels (emigrate) into the damaged area and begin to phagocytize the bacteria and destroyed cells; the resulting mixture of dead cellular debris and white blood cells is known as pus.

Major signs of inflammation

The major signs of inflammation are redness and increased heat (caused by blood-vessel engorgement and enlargement), swelling (resulting from the accumulation of fluid), and pain. The last of these signs (pain) is somewhat of a mystery. Although there is considerable doubt as to its cause in inflammation, there is unanimous agreement that it is one of the cardinal signs of all inflammatory responses. Inflammation can be classified as either acute or chronic. Acute inflammation, such as seen around the skin cut, lasts for only a few days and is characterized microscopically by the presence of a certain type of white blood cell—the polymorphonuclear leukocyte. Chronic inflammation is of a longer duration and is characterized microscopically by the presence of lymphocytes, monocytes, and plasma cells and, in general, is associated with little fluid exudation.

Because of the pain and swelling, the inflammatory response is often viewed as an unwelcome event following injury. Yet it is important to recognize that it is the first step in the healing process and represents an important protective response in the maintenance of health (see INFLAMMATION).

The immune response. The immune reaction is one of the most important defense mechanisms against biotic invasion and is therefore vital to the preservation of health. It is a relatively recent development from an evolutionary point of view, being found only in vertebrates. This complex system, covered at length in the article INFLAMMATION, has, as its participants, antigens, antibodies, complement, and white blood cells. The interaction of these participants can perhaps be suggested by the following analogy: a potential criminal appears but is apprehended by the forces of law and order in the nick of time and is securely impounded in the local jail. In the analogy the villain represents an antigen, any substance capable of evoking an immune response. Most often the antigen is a protein or a polysaccharide (a complex carbohydrate), both of which are important constituents of many bacteria or viruses. In the analogy the police correspond to antibodies or immunologically reactive cells. Antibodies, complex protein structures that have the ability to bind to (form compounds with) antigens and neutralize them, are formed by plasma cells and released into the circulation; thus they are available throughout the body for reaction with antigens wherever they are present. Certain antigens do not evoke circulating antibodies but instead cause the generation of lymphocytes that are capable of an immune response and are known as immunologically competent cells. This special form of immune response is provoked largely by certain types of bacteria, such as those that cause tuberculosis (*Mycobacterium tuberculosis*). In either event, the intruder is neutralized by either antibodies or immunologically competent cells and so is prevented from doing further damage. Complement, a complex group of proteins found in the blood, facilitates the immune response by attracting phagocytic white cells to the area of the reaction and by creating a firmer union between the antigens and the antibodies. The subsequent phagocytosis (imprisonment) of the offending antigen by white cells completes the analogy.

Action of complement

Two of the remarkable qualities of the immune system are specificity and "memory." When an antigen enters the body it evokes a specific antibody or specific immunologically competent cells; that is, the antibody or the cells will neutralize only the antigen that evokes them. Furthermore, the system exhibits what appears to be memory: once challenged by an antigen, the body "remembers" it for years and perhaps for life. The child who has an attack of measles is immune for the rest of his life. When the specific antigen (such as the measles virus) appears at a later date, the immune system recognizes it and responds and thereby prevents a reinfection. Indeed, these two characteristics of the immune system, specificity and memory, serve as the basis for preventive immunization. By inoculation of infants or children with inactivated or attenuated biotic agents, the immune system is made alert to such an antigen should it appear at a later date. Poliomyelitis, for example, once dreaded as a cause of paralysis and death, has recently come under control with the development of the polio vaccine.

What has been said will aid in understanding why certain illnesses (such as measles, mumps, and chickenpox) seem to affect only children—and in fact, are referred to

as the childhood diseases. While these viral diseases can affect persons of any age, most adults have had previous exposure to the antigens (viruses) and are thus immune. Children with no previous exposure have no specific immunity to these invaders and consequently are able to develop the diseases.

Thus the immune system is a vital part of the defense against biotic invasion. The immune system may also be a cause of disease (see below *Diseases of immune origin*).

REPAIR AND REGENERATION

By the replacement of damaged or destroyed cells with healthy new cells, the processes of repair and regeneration work to restore an individual's health after injury and protect his future health. Unlike the salamander, which is capable of totally regenerating a limb if it is destroyed or cut off, man cannot regenerate whole organs or limbs. If one of the kidneys is totally destroyed by a disease process, it is permanently lost. Moreover, the many cell types of the body have varying capacities for regeneration.

What regeneration is

Regeneration is the production of new cells exactly like those destroyed. Of the three categories of cells in man—(1) the labile cells, which continue to multiply throughout life; (2) the stable cells, which do not multiply continuously but have the capacity to do so when called upon to replace cells lost by injury or some destructive process; and (3) the permanent cells, incapable of multiplication in the adult—only the permanent cells are incapable of regeneration. These are the brain cells and the cells of the skeletal and heart muscles.

Labile cells comprise those of the bone marrow, the lymphoid tissues, the skin, and the linings of most of the ducts and hollow organs of the body.

Stable cells are found in the liver, in many of the glands of the body, such as the pancreas and salivary glands, in the lining of the kidney tubules, and in the connective tissues. Normally these cells do not divide unless some are destroyed by disease or injury and must be replaced.

If only a small area of liver (made up of stable cells) is damaged or destroyed, the cells can be replaced by regeneration of the still-living cells around the area of injury. When large areas of liver are destroyed, cellular regeneration cannot occur, and the area of cell loss is replaced by new healthy connective-tissue cells, which thus produce scars. When a person has a heart attack, a certain number of heart-muscle cells (permanent cells) are killed because of loss of their blood supply. Since heart-muscle cells cannot regenerate, the area of injury is replaced by a scar (if the patient survives). Such repair is by no means perfect, but it nonetheless permits restoration of reasonable heart function with perhaps only a slightly reduced level of health.

Cellular regeneration in man is limited by many other factors, such as the availability of blood supply and a supporting connective tissue. When the blood vessels and supporting cells (connective tissue) are destroyed in the liver along with the liver cells, perfect reconstitution of the liver is not possible. There may be some regrowth of liver cells, but they do not form the normal liver architecture, and the newly regenerated cells cannot function because they do not have an appropriate orientation to the blood vessels and bile ducts.

Repair of a cut

A review of the events that occur after a simple cut in the skin provides a good example of the processes of regeneration. At first, the area becomes red, swollen, and painful, because of the inflammatory reaction. A scab forms. Beneath the scab, while the inflammatory process is going on, the cells from the adjacent healthy skin begin to regenerate by dividing and growing over the damaged area. If the damage is minor, perfect reconstruction of the skin and its appendages is likely to result. If the damage has extended below the skin surface, deeper connective-tissue cells, notably the fibroblasts, proliferate and fill the area. These cells lay down collagen (connective-tissue protein) composed of tough, durable fibrils (minute fibres), and, eventually, scar formation ensues. Once scarring has occurred, it cannot be reversed, although considerable shrinking of the scar may occur.

HEMOSTASIS

Another mechanism of defense is hemostasis, the prevention of loss of blood from the blood vessels by formation of a clot. (This process is covered more at length in the article BLEEDING AND BLOOD CLOTTING.) Simply stated, a break in a blood vessel leads to the activation of a complex sequence of events leading to the formation of a solid plug of fibrin and blood cells (fibrin is a fibrous protein formed from fibrinogen). This plug, or clot, seals the damaged vessel and prevents further loss of blood (hemorrhage). The numerous components of the fluid blood called clotting factors contribute in sequential fashion to the formation of the clot. (The clotting factors are commonly referred to by a roman numeral rather than by name. Fibrinogen, for example, is clotting factor I.) These many factors are each important because abnormalities, such as the absence of one of the clotting factors (VIII), can lead to disorders of bleeding, such as that found in hemophiliacs.

INTERRELATIONSHIP OF DEFENSIVE MECHANISMS

The homeostatic and defensive mechanisms involved in maintaining a constant internal environment are complex and yet beautifully coordinated. Thus, the normal state of health is not a static condition but exists rather within a narrow range maintained by the coordinated responses of many systems and mechanisms. Health requires the proper function of all of these controls. Disease may begin in a single organ or system, but the interdependence and close coordination of the many bodily functions, which cooperate so beautifully in health, may be upset by a chain reaction when one breaks down. A disease of the kidney leading to abnormal retention of sodium, for example, can lead to hypertension (high blood pressure). Prolonged hypertension in turn can induce heart failure, and this can result in the abnormal collection of fluid in the lungs. The impairment of respiratory function may then result in a sudden rise in the level of carbon dioxide in the blood, which brings with it further complications. Similarly, if the normal inflammatory response malfunctions, a trivial skin infection (popularly known as a pimple) can enlarge into a boil (a furuncle). The responsible bacterial agents may proliferate in the local site and penetrate small blood vessels to seed the bloodstream, thus causing a generalized infection (septicemia or bacteremia). Such a generalized infection is extremely serious and may cause secondary infections of the heart (endocarditis) or of the coverings of the brain (meningitis).

Chain reactions in disease

Thus health implies the proper functioning of the homeostatic mechanisms just described, including those systems involved in the defense of health. The state of disease basically represents a failure of these mechanisms. Although one tends to think of disease in terms of offending agents, these agents are only able to produce disease by their ability to disrupt normal homeostasis, and it is precisely those disruptions that are the manifestations of disease.

II. Disease: signs and symptoms

Disease may be acute, chronic, malignant, or benign. Of these terms, chronic and acute have to do with the duration of a disease, malignant and benign with its potentiality for causing death.

An acute disease process usually begins abruptly and is over soon. Acute appendicitis, for example, is characterized by the sudden onset of nausea, vomiting, and pain usually localized in the lower right side of the abdomen. It usually requires immediate surgical treatment. The term chronic refers to a process that often begins very gradually and then persists over a long period. Peptic ulcer, for example—an ulcer usually of the stomach or duodenum—is a chronic disease. It may begin at any age and tends to persist indefinitely, although strict regulation of the diet and antacid therapy may result in good control.

The terms benign and malignant, most often used to describe tumours, can be used in a more general sense. Benign diseases are generally without complications, and a

good prognosis (outcome) is usual. A wart on the skin, for example, is a benign tumour caused by a virus; it produces no illness and usually disappears spontaneously if given enough time (often many years). Malignancy implies a process that, if left alone, will result in fatal illness. Cancer is the general term for all malignant tumours.

Each disease entity has a constellation of signs and symptoms more or less uniquely its own; individual symptoms such as fever, however, may be found in a great number of diseases. Some of the common manifestations of disease—as they relate to an imbalance of normal homeostasis—are taken up in this section. They are covered more at length in the article DIAGNOSIS.

Fever is an abnormal rise in body temperature. It is most often a sign of infection but can be present whenever there is tissue destruction, as, for example, from a severe burn or when large amounts of tissue have died because of lack of blood supply. Body temperature is controlled by the thermostatic centre in the hypothalamus. Certain protein and polysaccharide substances called pyrogens, released either from bacteria or viruses or from destroyed cells of the body, are capable of raising the thermostat and causing a rise in body temperature. Fever is a highly significant indicator of disease.

Increased number of white blood cells

Mention was made earlier of the phagocytic white blood cells and the process referred to as leukocytosis. The expansion of the number of circulating phagocytic white blood cells is one of the more common manifestations of disease. The stimulus for such an event may be any inflammatory process in the body, such as is caused by bacteria, viruses, or any process that leads to the destruction of cells. Such leukocytosis is reflected in the white-blood-cell count, which may be substantially elevated above the normal upper value of 10,000 cells per cubic millimetre of blood.

The pulse rate is another easily obtainable and important piece of information. The heart rate varies with the level of physical activity, beating faster during exercise and more slowly during rest. An inappropriate heart rate (or pulse) may be indicative of disease. The heart rate increases in the feverish patient. A weak, rapid pulse may be a sign of severe blood loss or of disease within the heart itself. Irregularity of the pulse (arrhythmia) is an important indicator of malfunction of the heart.

The respiratory rate (the rate of breathing) is modified by disease. People with fever have an increased respiratory rate (hyperventilation), which serves to lower body temperature (this rapid breathing is analogous to the panting of the dog). Hyperventilation is a common response to painful stress. Any condition leading to acidosis (lowering of body pH) similarly drives the respiratory rate upward. Diseases of the lungs—with accompanying inability to oxygenate the blood adequately—have a similar effect.

Thus it can be seen that temperature, pulse, and respiratory rate—called the vital signs—may be important manifestations of disease and are easily obtainable information for the physician. A fourth vital sign, blood pressure, is equally significant. Among other things, it indicates the amount of blood in circulation. A decrease in circulating blood volume, as is seen with severe bleeding, lowers the blood pressure and deprives the tissues of adequate blood flow. Reflexes are initiated that compensate for the reduced blood volume and blood pressure. The heart rate increases and compensates to some extent for the sudden reduction in blood volume and pressure; at the same time, peripheral blood vessels in such areas as the abdomen constrict, tending to divert the reduced blood volume to the more vital areas such as the brain and head. Unusual elevation of pressure (hypertension) is a disease by itself.

Causes of edema

Fluid and electrolyte imbalances may be further consequences of homeostatic failure and additional significant manifestations of disease. The causes of these abnormalities are complex. Edema, or swelling, results from shifts in fluid distribution within body tissues. Edema may be localized, as when the leg veins are narrowed or obstructed by some disease process. The pressure of the blood in the distended veins rises, and fluid is driven out of the vessels into the tissues, causing swelling of the extremity. Generalized edema is seen in renal (kidney) disease that causes abnormal retention of sodium and water. Heart failure is an additional cause of generalized edema, usually most manifest in the feet and ankles. Alterations such as dehydration, hyperventilation, and tissue destruction can all lead to varying fluid and electrolyte derangements. The levels of the serum electrolytes (sodium, potassium, carbonate, chlorine), determined relatively easily in the laboratory, provide the physician with valuable clues to deranged homeostasis induced by disease.

Finally, the determination of body pH and a number of blood tests designed to evaluate adequate (or inadequate) metabolic regulation provide diagnostic clues of homeostatic failure. These tests include determination of the levels of the blood sugar, blood urea nitrogen, and serum protein.

The disease diabetes mellitus provides an excellent example of such a failure of the homeostatic mechanisms. Diabetes is a common disease of metabolic–endocrine (ductless gland) origin involving a relative or absolute deficiency of insulin, a hormone that plays an important role in carbohydrate metabolism. Any or all of the above-mentioned homeostatic derangements can be found in this disease. Patients with a severe form of diabetes may at one time be dehydrated, because of obligatory excretion of water (osmotic diuresis); be acidotic, because of formation of increased amounts of keto acids; be hyperventilating, as a result of the acidosis; be comatose, because of high levels of blood sugar; have a weak pulse, because of severe dehydration; have electrolyte abnormalities; and so on. The signs and symptoms are numerous, all illustrating the interdependence of the homeostatic mechanisms, which, when not functioning properly, provide the manifestations of disease.

At the most elemental level, disease develops when any disruptive or adverse influence overcomes the homeostatic and defensive controls of the body. As will be seen, there are numerous influences that can tip the scales of health toward disease. Viruses and bacteria are obvious threats to health. There are a great many others, some so subtle as to be poorly understood. The following section focusses on the causes of disease rather than on a detailed description of each entity. It represents one method of classification. There is a considerable overlap in categories; certain diseases grouped as metabolic–endocrine in origin could also be classified as diseases of genetic origin. Indeed, the interdependence of the organ systems, the metabolic pathways, and the defense systems renders finite classification in medicine difficult. The human body acts as a unit—an individual—both in health and in disease.

Overlap in causal categories

III. The causes of disease

The search for the causes (etiologies) of the many diseases of man goes back to antiquity. Hippocrates, a Greek physician of the 4th and 5th centuries BC, is credited with being the first to adopt the concept that disease is not a visitation of the gods but rather is caused by earthly influences. Scientists have since continually searched for the causes of disease and, indeed, have discovered the causes of many.

In the development of a disease (pathogenesis) more is involved than merely exposure to its agent. A room full of people may be exposed to a sufferer from a common cold. One or two of those exposed may later contract the cold, but the others do not. A number of individual host factors determine whether the cause will induce disease or not. Thus, in the pathogenesis of disease, the resistance, immunity, age, and nutritional state of the person exposed, as well as virulence and level of exposure, all play a role in determining whether the person develops a disease.

In the following sections the many types of human disease will be divided into a number of categories, and in each category only a few examples will be given to establish the nature of the process. These categories are divided on the basis of the presumed etiology of the dis-

ease. Many diseases are still of unknown (idiopathic) origin. With others the cause may be suspected but not yet definitively proved. In a few instances, the discovery of the etiology of a disease represents the individual achievement of a solitary investigator who may have worked long years on the problem. The story of Louis Pasteur and the discovery of the cause of anthrax is a classic example of this. More often the individual investigator who makes the final breakthrough stands on the shoulders of hundreds of earlier workers who provided bits and pieces of knowledge vital to the final understanding. Thousands of research workers have spent millions of man-hours attempting to elucidate the cause of cancer. Regrettably, cancer must still be listed as being of unknown cause.

DISEASES OF GENETIC ORIGIN

Certain diseases of man are traceable to an abnormality (mutation) in genetic constitution (genome). The mutation, for example, may have altered the genetic coding necessary for the synthesis of a certain enzyme of vital importance in normal metabolism. Most genetic disorders can be detected at birth; the child is born with the defect or defects. Thus these abnormalities are congenital (at birth) genetic disorders. Surprisingly, a few genetic derangements do not become manifest until later life. An example is Huntington's chorea, a nervous-system disease in which the genetic defect requires years to become manifest. Hence it may be said that most but not all genetic diseases are congenital.

Congenital genetic disorders

Conversely, some congenital diseases are not genetic in origin; they may arise from some direct injury to the developing fetus. When a mother contracts the viral disease german measles (rubella) during pregnancy, the virus may infect the fetus and alter its normal development, leading to some malformations, principally of the heart. These malformations constitute a congenital disease that is not genetic (see BIRTH DEFECTS AND CONGENITAL DISORDERS).

Further confusion often arises over the terms genetic and familial. A familial disease is hereditary, passed on from one generation to the next. It resides in a genetic mutation that is transmitted by mother or father (or both) through the gametes (sperm and egg) to their offspring. Not all genetic disorders are familial. The mutation may occur in the formation of the gametes or may arise during the early development of the fetus. Such an infant will have some genetic abnormality, though the parents themselves are completely normal. Perhaps the most illustrative disease is Down's syndrome (mongolism). Most of these unfortunate children are born of normal parents. During the formation of the gametes an error in the division of the chromosomes (see below) leads to the infant's having one too many chromosomes (known as trisomy). All of the defects of Down's syndrome arise out of this extra chromosome. It is extremely important to recognize that genetic disease need not be familial, because other children of the same parents may be entirely normal.

Each cell in the body contains 22 homologous (alike) pairs of chromosomes called autosomes and one pair of sex chromosomes (XX in the female and XY in the male), yielding a total of 46 chromosomes. Each chromosome is comprised of long double strands of deoxyribonucleic acid (DNA) wound in spiral fashion and forming a double helix. The smallest functional unit of the DNA strand is known as a gene. Specific hereditary traits are determined by one or more genes located at precisely corresponding points on each of the pair of homologous chromosomes. A pair of genes on homologous chromosomes controlling a single trait (such as eye colour) is referred to as an allele. Should the alleles on a given pair of chromosomes be identical for a trait, the individual is said to be homozygous for that trait. Should the alleles be different, the individual is said to be heterozygous. Furthermore, one of the genes in an allelic pair is said to be dominant if the trait controlled by that gene is expressed (*i.e.*, is manifested) in a heterozygous individual and recessive if it fails to express itself in a heterozygote. For example, with respect to eye colour, if an individual were heterozygous, having received one brown allele from the mother and one blue allele from the father, he would have brown eyes because the brown allele is dominant and the blue allele is recessive. A blue-eyed person must have two blue alleles, because that trait must be homozygous to be expressed. Similarly, a disease resulting from a recessive gene (or allele) would be manifest only in the homozygous individual. Most diseases are inherited as recessive traits. Diseases controlled by dominant genes are much rarer, because persons carrying such a dominant gene would manifest the disease and so, in some instances, die very young, while those who survive tend not to marry and reproduce. This is an example of Darwin's theory of natural selection. The most desirable traits are selected for, and the most undesirable perish with their host.

Recessive inheritance of disease

Diseases of genetic origin can be divided conveniently into those with large mutations that produce demonstrable chromosomal abnormalities and those in which the mutation is at the level of a gene and is so subtle as to be inapparent with currently available methods of study of the chromosomes. This division, of course, is artificial, because the basic mechanism, alteration in genetic code, is common to both groups. Diseases with demonstrable chromosomal abnormalities have large genetic alterations and, in general, have more dramatic clinical presentations. On the other hand, diseases without demonstrable chromosomal abnormalities have very small genetic defects and, therefore, cause clinical derangements that are often more subtle.

The demonstration of chromosomal abnormalities involves the following process. Use is made of colchicine, a substance that is capable of stopping cell mitotic division when all the chromosomes are fully formed (this stage is called metaphase). By use of special stains, the chromosomes can readily be seen with a light microscope. Furthermore, they can be photographed and arranged pictorially into groups based on chromosome size and shape. Such an arrangement of chromosomes is referred to as a karyotype.

If no chromosomal abnormality can be seen, a disease can be determined to be genetic in origin only by a systematic study of the family of which the diseased person is a member. Successive generations in a family tree must be studied to determine whether the disease appears in the family pedigree and is presumably familial. Careful retrospective studies of such family trees not only can determine the genetic basis of a disease but also can illuminate the pattern of inheritance, such as dominance and recessiveness.

Factors relating to genetic injury. The causes of mutations are still poorly understood. Certain factors, however, are thought to be important. It is an accepted fact that maternal age plays an important role in predisposing toward genetic injury. The frequency of Down's syndrome and of congenital malformations increases with the age of the mother. All ova are formed during the fetal development of the female infant. They persist throughout the active reproductive life of the woman. It is hypothesized that the ova are continuously exposed to all the adverse influences, such as infections and nutritional deficiencies, suffered by the woman. The longer the exposure to such injury (*i.e.*, the older the mother), the greater the chance of genetic injury to the ova. No such relationship has been shown with the father's age. From a statistical point of view, the safest time for a woman to have children is before the age of 30.

Congenital defects and age of mother

Radiation is a well-recognized cause of chromosomal damage. In man, the survivors of the atomic-bomb blasts in Japan in 1945 have shown definite chromosomal abnormalities in certain types of their circulating white blood cells. Indeed, a higher incidence of leukemia (a form of cancer of white cells) has been reported in this population, suggesting that the chromosomal changes may have played some role in the induction of the disease (see also RADIATION INJURY).

Viruses have been shown to cause mutations in human cells when the cells are grown in a tissue-culture flask, but there is no clear evidence that viral infections of man have caused genetic injury. It should be re-emphasized

that the virus of German measles (rubella) causes congenital malformations, but it has not been established that there is a genetic role to these disorders.

The role of drugs and chemicals in the induction of genetic mutations is a complicated one. Currently, there is great concern about the possible ill effects of lysergic acid diethylamide (LSD) on the offspring of habitual users. While there are many agents that cause chromosomal alterations in the human cells grown in culture, there is not yet any definitive evidence of drug- or chemical-induced genetic injury in the living host.

Diseases associated with abnormalities in the karyotype. Demonstrable chromosomal abnormalities may arise through several pathways. These abnormalities may consist of an alteration in the structure of one or more chromosomes or an abnormal number of chromosomes. The mechanisms leading to alterations in number are nondisjunction and simple loss of a chromosome.

The normal chromosome count of man is 46. This number is referred to in genetics as the diploid mode. Before mitosis (division of cells into double the number of cells, each with 46 chromosomes) actually begins, there is synthesis of new deoxyribonucleic acid (DNA) until the cell about to divide contains double the normal amount of DNA and each of the 46 chromosomes also possesses twice the normal amount. During mitosis each of these large chromosomes divides in half; each half, called a chromatid, possesses as much DNA as a usual chromosome. During mitosis the chromatid doublet separates and the two halves migrate to the opposite poles of the cell about to be divided. In this fashion each daughter cell receives the usual diploid number of chromosomes. If a chromatid doublet fails to divide (nondisjunction), the resulting daughter cells receive an inappropriate number of chromosomes (*i.e.*, one cell might receive 47 chromosomes and the other 45) instead of the usual 46 chromosomes.

Nondisjunction, translocation, and deletion

Simple loss of a chromosome occurs during metaphase, when a chromosome fails to line up properly along the equator of the cell. When cell division occurs the aberrant chromosome is lost, and daughter cells have unequal chromosomal composition.

Mechanisms leading to alterations in chromosomal structure include translocation and deletion. Translocation occurs when a chromosome fractures and a resulting fragment attaches itself to another chromosome. When a fragment fails to reattach itself to a chromosome, that segment is lost from the genetic pool, a process referred to as deletion (Figure 1).

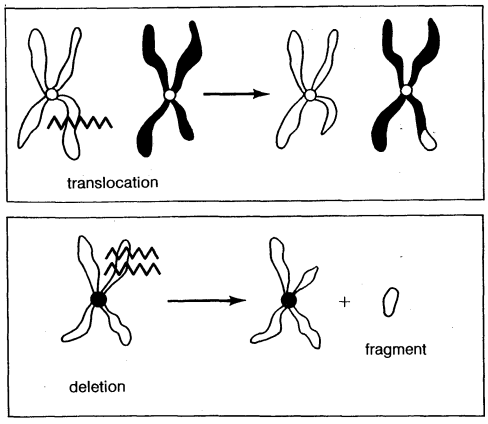


Figure 1: Alterations in chromosome structure.

Diseases with demonstrable chromosomal abnormalities can involve either autosomal (nonsex) chromosomes or the sex chromosomes. Down's syndrome is an example of the former; Turner's syndrome is an example of the latter.

Down's syndrome (mongolism) is a relatively common genetic disorder, affecting approximately one out of every 700 births. Nowhere is the effect of maternal age on chromosomal abnormalities better illustrated than in this disorder. When the mother is less than 30 years of age the incidence of this disorder is approximately one in every 3,000 births. In the age group 30–34 the incidence rises to one in 600. Mothers between the ages of 45 and 50 have a one in 40 chance of having a mongoloid child.

The basis of this disorder remained a mystery until 1959, when it was demonstrated that affected children have 47 chromosomes instead of the normal 46. The extra chromosome was found to be one of the smaller autosomal (nonsex) chromosomes. A rarer version of the same disorder is caused by a translocation of one chromosome to another (Figure 1).

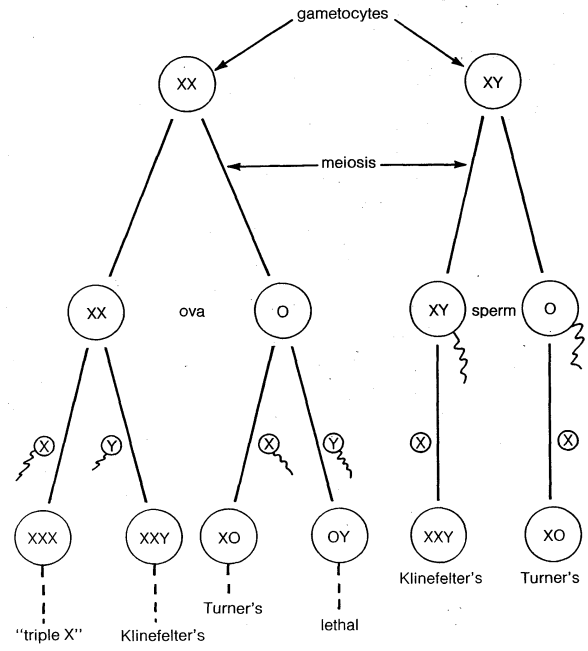


Figure 2: Nondisjunctional errors in gamete formation (see text).

The term mongolism is a misnomer, for while the child with Down's syndrome may seem to the casual observer to resemble somewhat the child of Mongolian descent, the latter bears none of the signs of the genetically defective child. These signs include mental retardation, below-average stature, slow reflexes, and abnormalities of the tongue, hands, feet, and eyes. Years ago, before the use of antibiotics, these children perished of pneumonia in infancy. Medical advances have provided them with a far better chance of surviving; indeed, they now account for a significant percentage of institutionalized children. Characteristically they are sweet, gentle children, and many, with training, can be significantly helped.

In the formation of sperm and ova—the sex cells, or gametes—a reduction division (meiosis) occurs. It differs from mitosis in that the daughter cells each contain exactly one-half the normal chromosomal number. Thus, each sperm or egg contains 23 single chromosomes (22 autosomes and one sex chromosome) instead of 46. During fertilization, when the gametes unite, they form a zygote (fertilized egg), which has 46 chromosomes, half from the father and half from the mother. Just as the 22 pairs of autosomes are divided in meiosis, so are the pair of sex chromosomes. In the production of spermatozoa, meiotic division can result in sperm containing either an X-chromosome or a Y. The production of female gametes, or ova, yields only cells containing an X.

Abnormalities of sex chromosomes

Demonstrable abnormalities of the sex chromosomes are almost always the result of nondisjunction (see above) during meiosis. Thus one gamete may possess two sex chromosomes, while the other contains none. Repeated nondisjunctional errors in meiosis may yield gametes with three, four, or even five sex chromosomes contributed by one or both gametes. Thus zygotes may have too few or too many chromosomes and even two or three Y-chromosomes. A Y-chromosome determines the male sex of the fetus, however many X-chromosomes may be

present. It is the absence of a Y that specifies the female sex. The multi-X zygotes are female, as are those with a single X and no Y. The absence of both sex chromosomes or the presence of only a single Y-chromosome is incompatible with life. One of the more frequent patterns of abnormal meiotic division is illustrated in Figure 2. In this figure, during male spermatogenesis, nondisjunction has occurred, yielding one gamete with two sex chromosomes instead of one and the other gamete without sex chromosomes. The mating possibilities for these gametes would thus yield either an XXY zygote or an X0. This indeed does occur. People with an XXY chromosomal pattern have a disorder known as Klinefelter's syndrome. People with only one X-chromosome (X0) have a disease known as Turner's syndrome.

Turner's syndrome (X0) occurs in approximately one out of every 3,000 births. These female children lacking a second sex chromosome often are not recognized as suffering from some disorder until puberty, at which time the conspicuous absence of secondary sex characteristics (*i.e.*, pubic hair and breast formation) becomes evident. Children with Turner's syndrome are quite short in stature with a broad, weblike neck. The breasts are poorly developed and the internal sex organs are functionless. These girls fail to menstruate and usually are sterile. Klinefelter's syndrome is a disorder of persons who, in general, are male in physical configuration but have small testes, do not produce sperm, and display varying degrees of eunuchoidism.

Diseases without demonstrable chromosomal abnormalities. Enzymes are specialized proteins that facilitate the chemical reactions in the body. The synthesis of these proteins is dictated by the sequence of certain units (nucleotide bases) in the deoxyribonucleic acid (DNA) molecule within the chromosome. Mutations involving a portion of the DNA molecule can result in the defective production, or total absence, of one or more enzymes. More often than not, these mutations produce little or no disease. On occasion the absence of a critical enzyme can have far-reaching results. At present, several thousand distinct diseases have been ascribed to mutations that lead to the absence or relative lack of critical enzymes. Collectively these disorders are referred to as inborn errors of metabolism (see METABOLISM, DISEASES OF).

Disease due to lack of an enzyme

Table 2: Examples of Genetic Diseases with No Demonstrable Chromosomal Abnormalities

metabolic defect	disease	mechanism
Defective enzyme synthesis		
Tyrosinase	albinism	blockage of synthesis of melanin
...	gout	excess production of uric acid
Defective synthesis of proteins not classified as enzymes		
Hemoglobin	sickle-cell anemia	abnormal hemoglobin leading to tendency to hemolysis
Factor VIII (AHF)	hemophilia	defective hemostasis leading to bleeding disorder
No metabolic defect known		
—	polydactylism	—
—	cleft palate	—

While most studies of these submicroscopic genetic abnormalities have concerned themselves with enzyme abnormalities, there are important exceptions. There are several diseases caused by probable mutations at the gene level in which no biochemical abnormality has been identified. The addition of one or more digits to the hand (polydactylism), for example, is a genetic disorder not associated with any known biochemical abnormality.

Finally, genetic mutations can produce disease by abnormal synthesis of proteins not classified as enzymes. Disorders involving hemoglobin, the protein responsible for carrying oxygen in the blood, have been clearly described (*e.g.*, sickle-cell anemia). Hemophilia is a sex-linked (carried on the sex chromosome) disorder similarly involving a protein not classified as an enzyme. Table 2 summarizes these classifications.

CONGENITAL MALFORMATIONS

Congenital malformations, any structural defects present at birth, range from those that have serious effect upon health to trivial alterations, such as an extra nipple (supernumerary nipple) or a "strawberry mole." Some of the more serious forms of malformation include blindness, the absence of a limb, cleft palate, and mental retardation. As many of the serious childhood diseases have been brought under control, congenital malformations have become major causes of death in newborn infants, accounting for about 15 percent of all mortality during the first year of life. Progress in preventing birth defects has been painfully slow because the mechanisms and derangements that lead to such malformations are largely unknown. Nonetheless, the study of birth defects (teratology) and their causes has become a major area of current medical–biological research.

Incidence of congenital defects

The overall incidence of congenital malformations lies in the range of 1–10 percent of all births. This wide range is partially the result of differences in defining the term malformation. Another variable is the age range employed in the analysis of congenital malformations. Some reports confine their analyses to the first month or first year of life, while others extend the time span of the study. The incidence of malformations varies also with the time span embraced, because some malformations are not discovered immediately after birth.

Congenital structural abnormalities must be viewed as the postnatal emergence of developmental defects that may have originated at any time in the development of the fetus, beginning with the fertilization of the ovum. The malformations encountered in living children represent the mildest forms of developmental defects. The more severe malformations are not compatible with survival of the fetus. It is estimated by some that approximately 20 percent of fertilized eggs fail to develop because of serious developmental anomalies. Less severe aberrations are compatible with shortened fetal survival. There is a natural process of selection in which the greater the structural defect, the less the likelihood of survival.

The causes of congenital malformations are poorly understood at this time. It is believed that approximately 10 percent can be attributed directly to genetic defects, while another 10 percent represent the influence of environmental factors on the developing fetus. The remaining 80 percent may have both genetic and environmental factors involved in their causation. Most of the following material focusses on the 20 percent of malformations whose causes are better understood.

The genetic factors that contribute to congenital malformation have been previously cited in the discussion on genetics. There are other structural anomalies that are suspected but not proven to have genetic origins. Cleft palate and harelip are examples of these defects. Although no well-defined pattern of inheritance has been identified and no demonstrable abnormalities have been discerned in the chromosomes, there is a tendency for these effects to appear in families, suggesting a hereditary genetic origin. Certain environmental influences have been proven to be capable of causing congenital malformations. One would imagine that the fetus would be protected from its environment securely tucked away inside the mother's womb. It has been well documented, however, that radiation, drugs and chemicals, viruses, hormones, and nutrition may reach inside the womb to induce developmental anomalies in the fetus.

Radiation as cause of defects

Ionizing radiation is a well-documented potential cause of congenital malformations. Pregnant women exposed to radiation have an increased number of fetal deaths as well as an increase in the number of congenital defects in their children. Most of these defects are found above the neck and include microcephaly (small head), skull defects, blindness, and mental retardation.

Although few drugs have been proven to be capable of causing malformation (teratogenic), there are some notable exceptions. Thalidomide, a drug taken as a sedative, was found to cause the formerly rare congenital defect known as phocomelia (seal-flipper arms and legs). Other

compounds that cause or may cause malformations include aminopterin and busulfan (both anticancer agents), quinine, and lysergic acid diethylamide (LSD).

Several infectious diseases cause malformations when they occur during pregnancy. Although it is considered a benign childhood disease, German measles (rubella) in the early weeks of pregnancy directly injures the fetus and so leads to certain birth defects in the infant. The nature of these defects is related to the period of pregnancy in which the disease occurs. For instance, cataracts (eye defects) are found in children when the rubella-virus infection occurs in the sixth week of pregnancy. Viral infections in the ninth week of pregnancy result in a strong chance of deafness in the child. Similarly, heart and dental abnormalities are noted in children of mothers affected during the fifth through tenth week of gestation. After the tenth week of pregnancy, rubella infections have little effect on the fetus.

Syphilis, a bacterial venereal disease, during pregnancy is a well-recognized cause of deformities in the offspring. These include thickening of the shins, notching of the teeth, collapse of the bridge of the nose, and eye defects. Since syphilis is an easily treatable disease, such malformations are avoidable.

Neither hormones nor nutritional factors are considered important causal agents in the production of gross structural defects at birth. Diabetes mellitus, a disorder of carbohydrate metabolism caused by a lack of the hormone insulin, is a possible exception. Studies of the offspring of diabetic mothers have revealed a high incidence of stillbirths and neonatal deaths. It has also been noted that the average weight at birth of infants of diabetic mothers is considerably higher than that of normal children. Some investigators claim an increased incidence of congenital defects in the offspring of diabetic mothers, but the validity of these studies is disputed (see also BIRTH DEFECTS AND CONGENITAL DISORDERS).

HEREDITY AND ENVIRONMENT

Both genetic and environmental factors appear to play major roles in producing congenital malformations. Other diseases can be spread across a wide spectrum, with predominantly genetic diseases at one extreme of the spectrum and diseases largely environmental at the other. In the genetic part of the spectrum are diseases such as Turner's syndrome; in the environmental part are infectious diseases and chemical poisoning. Lying at some point between these two extremes are the bulk of diseases of man—those with both genetic and environmental causative influences that are significant. Indeed, even at the very extreme ends of the spectrum both factors play some role. The genetic constitution dictates in part the host's response to environmental challenges. Similarly, environmental factors play significant roles in the manifestation of genetically induced disease. Sickle-cell anemia, for example, an inherited disease characterized by abnormal red blood cells and hemoglobin, is seriously exacerbated by low levels of oxygen in the air.

Familial tendencies to disease

Furthermore, there are many disorders in which there is a familial tendency to develop the disease, but no formal pattern of inheritance has been delineated. Many forms of cancer, high blood pressure, arthritis (inflammation of the joints), and obesity, for example, seem to have a familial tendency. Although the exact roles of environmental and genetic factors are unknown in all these diseases, it is strongly felt that both factors contribute in some part to the disease process.

CHEMICAL AND PHYSICAL INJURY

Chemical injury: poisoning. A poison is any substance that can cause illness or death when ingested in small quantities. This definition excludes the multitude of substances that can produce harm if ingested in large quantities. Even sugar solutions of high enough concentration can cause damage to cells.

There are several considerations to keep in mind when one discusses poisoning. The first of these, as already suggested, is the degree of toxicity. A substance with a very high toxicity (such as hydrogen cyanide gas) need only be taken in minute amounts to cause serious harm.

A second consideration is the mechanism by which a poison exerts its harm. Some chemicals act solely on the kidney or liver, while others affect principally the central nervous system. Each poison has a characteristic ability to cause damage at particular sites within the body.

A third factor is the body's ability to eliminate the substance. Some chemicals, rapidly excreted in the urine, must exert their harm rapidly while they remain transiently in the body. Others are poorly eliminated. With these, chronic ingestion of nontoxic amounts leads to a buildup in the body that can eventually reach toxic levels. Lead poisoning is a good example of this phenomenon.

The route of entry is also important. Many substances are harmless when eaten but become deadly if injected into a vein. Similarly, the chemical form of the substance affects its action on the body. Metallic mercury, as found in thermometers, is harmlessly excreted, whereas the chloride salt of the same substance is deadly.

Finally, the condition of the host, the recipient of the poison, is an important consideration. A dose of aspirin harmless to an adult may be poisonous to an infant. Similarly, an elderly person's tolerance of a substance may be much lower than that of a healthy young adult.

There is a wide variety of poisons. As was already noted, almost any substance in sufficient quantity is capable of causing harm. Among the many poisons, a few stand out as being the most commonly encountered in medical practice. Some are of relatively low toxicity, but have importance because of their widespread use. Many physicians consider aspirin (acetylsalicylic acid) the most dangerous poison because of its commonplace use and abuse and because it is the leading cause of poisoning in children. In the following paragraphs three groups of agents will be presented: (1) organic chemicals, (2) inorganic chemicals, and (3) drugs.

Organic chemicals. Among the organic chemicals commonly encountered in instances of poisoning are two forms of alcohol, ethyl alcohol (ethanol) and methyl alcohol (methanol). Ethyl alcohol is the form found in most alcoholic beverages. Methyl alcohol, or wood alcohol, is used for a variety of household purposes.

Ethyl alcohol poisoning

Acute ethyl alcohol poisoning is encountered after ingestion of large quantities over a relatively short time. The alcohol is quickly absorbed from the gastrointestinal tract, and high blood levels can be achieved in a remarkably short time. Ethyl alcohol acts principally as a central-nervous-system depressant and, fortunately, stupor usually results before fatal doses can be reached. The difference in blood levels between intoxication and fatal stupor is very slight, however, and death may result with the ingestion of large quantities of alcohol from depression of the respiratory centre in the brain.

Chronic alcoholism is considered by some as a form of poisoning, but most view it as a sociomedical disease. It may lead to serious liver injury (cirrhosis) over the course of years, but there is debate as to whether it is the alcohol itself that is injurious to the liver or the inadequate food intake of the chronic alcoholic.

Methyl alcohol (methanol) is usually ingested either by accident or with suicidal intent. Once inside the body it is transformed to formic acid, an extremely toxic substance that selects the nerves in the eye as its target. Without treatment, blindness results. Methyl alcohol can also affect the brain tissue itself, causing the brain to swell and leading to death in some instances. Prompt treatment can prevent lasting damage to the brain and the eyes.

Carbon monoxide is a nonirritating, inert gas, without colour, taste, or odour. A poison responsible for a large number of suicidal deaths, it is one of the chemical products of any combustion of organic material. Inhalation of a 1 percent concentration can be fatal within ten to 20 minutes. Once absorbed, carbon monoxide develops a firm union with hemoglobin; thus chronic inhalation of nontoxic concentrations can lead to a cumulative buildup of the substance in the blood. The formation of the compound with hemoglobin prevents the normal carriage of oxygen by the hemoglobin. Thus carbon monoxide acts as an internal asphyxiant by leading to oxy-

gen starvation of tissues. It suffocates the cells of the body and so causes death or serious disease. It should be noted that exposure to even low concentrations can result in the slow accumulation of this poison over hours, days, or weeks, leading very gradually to toxic or fatal levels.

Inorganic chemicals. The inorganic chemicals most commonly responsible for poisonings in the United States are cyanide, mercury, arsenic, and lead. It should be remembered that these substances often appear in chemical forms that are quite harmless. In general, it is the soluble salts of the substances that are poisons.

Cyanide poisoning

Cyanide is a dangerous substance in any form. It may occur in the form of hydrocyanic gas or as solid compounds such as potassium cyanide. One of the most lethal poisons known to man, in amounts less than even 0.1 gram it may cause death within minutes. Like carbon monoxide it is an internal asphyxiant, but it differs from carbon monoxide in its site of action. Whereas carbon monoxide prevents the delivery of oxygen to cells, cyanide prevents the utilization of oxygen by cells. Unable to utilize oxygen, the cell is deprived of its energy-generating biochemical reactions, vital to its function and preservation. The result is cell death on the microscopic level and often death of the whole organism.

Mercury in the pure metallic form is rather harmless, but the salt of the same substance, notably mercuric chloride, is a deadly poison. As little as 0.1 gram is enough to cause damage to body tissues and one gram can cause death. This agent causes extensive tissue damage wherever high concentrations of the poison are encountered. When the substance is swallowed, the stomach represents the portal of entry. The mercuric chloride is partially absorbed into the blood, and this portion is excreted through the urine. The remainder affects organs in the digestive tract, principally the stomach, the kidneys, and the colon. Mercuric salts can cause death of cells by precipitating the proteins within the cells, a form of cell injury called coagulative necrosis. The damage to the kidneys leads to kidney failure, abnormally high levels of urea in the blood, and even death. The poison induces ulcerations in the colon and bloody diarrhea. With careful treatment affected persons survive with full recovery. Chronic ingestion of smaller amounts of mercuric salts, as is seen in some industrial settings, can result in disease involving the mouth, skin, and nervous system.

The favourite poison (that is, for homicidal purposes) is arsenic. Both odourless and tasteless compounds of arsenic are found in some rat poisons, plant sprays, paints, and other household preparations. Many of these household staples are ingested accidentally by children. Principally affected by arsenic are the blood vessels and the central nervous system. Indeed, vascular collapse and depression of the central nervous system can be followed by coma and death within hours after ingestion.

Lead salts

The soluble salts of inorganic lead are also strong systemic poisons. They may accumulate within the body over a long period until toxic levels are reached, and cell damage ensues. These salts were at one time commonly found in paints, and lead poisoning was frequently seen in children who chewed on their painted cribs or woodwork. Recent legislation in many countries has outlawed the use of lead-base paints for infants' furniture. Other forms of poisoning are incurred through industrial exposure and ingestion of water from lead pipes. The major effects of lead poisoning are found in the blood-forming organs, the gastrointestinal tract, and the central nervous system. Lead poisoning damages red blood cells and leads to hemolysis (rupturing of red blood cells) with resulting anemia. In the brain, lead accumulation causes the degeneration of nerve cells. This produces such manifestations as mental depression, psychoses, convulsions, and even coma and death. Peripheral nerves are damaged as well, sometimes causing weakness or even paralysis of hands and feet. Within the mouth, a darkish lead line can be seen along the dental margins of the gums. If an early fatality does not occur, the lead is slowly excreted and complete recovery may be anticipated.

Drugs. Drugs are another important cause of poisoning. It is a pharmacological principle that for any therapeutic gain derived from a drug, a price is paid. There are few drugs used today that have no side effects (*i.e.*, effects unintended when the drug is administered). Although these side effects may be harmless and inconsequential, certain drugs have side effects that are potent. Similarly, a drug may be useful in a certain dose range but harmful when larger doses are taken. Morphine, for example, is an excellent drug for the control of severe pain, but morphine can depress respiration, and too much of it can cause death. The following material is concerned with barbiturates and salicylates, drugs commonly found to cause serious illness from overingestion. In the pill-oriented society of modern developed nations, there is an alarming increase in drug-induced disease and drug-induced death. All drugs are potentially harmful.

Barbiturates are pharmacologically classed as hypnotics, because they are used as sedatives and sleeping pills. Besides their therapeutic use, they are also commonly employed in suicidal attempts, which require far larger amounts than the therapeutic dose. Barbiturates affect the central nervous system almost exclusively. With toxic levels, the vital centres located within the midbrain are depressed; this leads to profound coma, depression of respiration, oxygen starvation of the tissues, and even shock. The identification of barbiturate poisoning relies almost exclusively on finding the substance in the blood or urine, because there is little anatomical change. Treatment is directed toward getting the drug out of the system as quickly as possible, either by inducing copious urinary excretion of the drug or by the use of the artificial kidney—a process called hemodialysis.

Aspirin and other salicylates

Aspirin, or acetylsalicylic acid, is a drug that deserves special mention because it is such a common household item and therefore within the reach of small children. In the United States approximately 27,000,000 pounds of aspirin are consumed yearly, enough to treat about 17,000,000,000 headaches. Approximately ten to 30 grams of aspirin can be fatal in adults, and much smaller amounts can be fatal in children. (A single aspirin tablet of standard size contains approximately one-third gram.) Doses of methyl salicylate (wintergreen oil) as small as three grams have been known to be responsible for infant death. There are many signs and symptoms associated with salicylate poisoning, including headaches, drowsiness, dyspepsia, nausea, vomiting, sweating, and thirst. Salicylate poisoning is an acute medical emergency. Rigorous medical treatment is demanded and use of the artificial kidney is often required.

Heroin is rapidly becoming one of the leading killers among the adolescent population. Not only can overdoses of the drug result in severe respiratory depression and even death, but its injection may bring on such disorders as serum hepatitis—a dangerous form of liver disease—from a contaminated needle and reactions to impurities such as fungi and bacteria. (Heroin and its use and abuse are taken up in a number of other articles; *e.g.*, NARCOTIC; DRUG PROBLEMS.)

Physical injury. Physical injuries include those caused by mechanical trauma, heat and cold, electrical discharges, changes in pressure, and radiation. Mechanical trauma is an injury to any portion of the body from a blow, crush, cut, or penetrating wound. The complications of mechanical trauma are usually related to fracture, hemorrhage, and infection. They do not necessarily have to appear immediately after occurrence of the injury. Slow internal bleeding may remain masked for days and lead to an eventual emergency. Similarly, wound infection and even systemic infection are rarely detectable until many days after the damage. All significant mechanical injuries must therefore be kept under observation for days or even weeks.

Injuries from cold or heat. Among physical injuries are injuries caused by cold or heat. Prolonged exposure of tissue to freezing temperatures causes tissue damage known as frostbite. Several factors predispose to frostbite, such as general weakness, lack of adequate clothing, and any type of insufficiency of the peripheral blood vessels. Frostbite tends to occur in the most exposed areas (ears, nose, fingers) and in areas of poor circulation (feet). The

first symptom is usually the sensation of a sharp prickling feeling coming from an area of numb, hard skin. Since the freezing produces local anesthesia, the victim rarely, if ever, suffers severe pain and is often unaware of any harm being done until hours or even days after the exposure. Eventually the affected area becomes swollen, blistered, and, as tissue dies and sloughs off, extremely susceptible to infection. This stage is accompanied by extreme pain (see FROSTBITE).

When the entire body is exposed to low temperatures over a long period, the result can be alarming. At first blood is diverted from the skin to deeper areas of the body, resulting in anoxia (oxygen lack) of the skin and the tissues under the skin, including the walls of the small vessels. This small-blood-vessel damage leads to swelling of the tissues beneath the skin as fluid seeps out of the vessels. When the exposure is prolonged it leads eventually to cooling of the blood itself. Once this has occurred the results are catastrophic. All the vital organs become affected, and death usually ensues.

Three categories of burns

Burns may be divided into three categories depending on severity. A first-degree burn is the least destructive and affects the most superficial layer of skin, the epidermis. Sunburn is an example of a first-degree burn. The symptoms are pain and some swelling. A second-degree burn is a deeper and hence more severe injury. It is characterized by blistering and often considerable edema (swelling). A third-degree burn is extremely serious; the entire thickness of the skin is destroyed, along with deeper structures such as muscles. Because the nerve endings are destroyed in such burns, the wound is surprisingly painless in the areas of worst involvement.

The outlook in burn injuries is dependent upon the age of the victim and the percent of total body area affected; for instance, a 20-year-old man with third-degree burns over 50 percent of his body has a 50-percent chance of survival, while a 50-year-old person with the same burn has a 20-percent chance.

Loss of fluid and electrolytes and infection associated with loss of skin provide the major causes of burn mortality. Much progress has been made in the treatment of burns with the advent of bacteriocidal creams, better understanding of fluid and electrolyte control, and the improvement of skin grafting techniques (see BURNS).

Electrical injuries. The injurious effects of an electrical current passing through the body are determined by its voltage, its amperage, and the resistance of the tissues in the pathway of the current. It must be emphasized that exposure to electricity can only be harmful if there is a contact point of entry and a discharge point through which the current leaves the body. If the body is well insulated against such passage, either at the point of entry or discharge, no current flows and no injury results. The voltage of current refers to its electromotive force, the amperage to its intensity. With high-voltage discharges, such as are encountered when an individual is struck by lightning, the major effect is to disrupt nervous impulses; death is usually caused by interruption of the regulatory impulses of the heart. In low-voltage currents, such as are more likely to be encountered in accidental exposure to house or industrial currents, death is more often due to the stimulation of nerve pathways that freeze muscles and may in this way block respiration. If the electrical shock does not produce immediate death, serious illness may result from the damage incurred by organs in the pathway of the flow of the electricity through the body (see SHOCK, ELECTRICAL).

Pressure-change injuries. Physical injuries from pressure change are of two general types: (1) blast injury and (2) the effects of too rapid changes in the atmospheric pressure in the environment. Blast injuries may be transmitted through air or water; their effect depends upon the area of the body exposed to the blast. If it is an air blast, the entire body is subject to the strong wave of compression, which is followed immediately by a wave of lowered pressure. In effect the body is first violently squeezed and then suddenly overexpanded as the pressure waves move beyond the body. The chest or abdomen may suffer injuries from the compression, but it is the negative pressure following the wave that induces most of the damage, since overexpansion leads to rupture of the lungs and of other internal organs, particularly the intestines. If the blast injury is transmitted through water, the victim is usually floating, and only that part of the body underwater is exposed. An individual floating on the surface of the water may simply be popped out of the water like a cork and totally escape injury.

Disease caused by sudden changes in atmospheric pressure is encountered principally in two situations: in those persons who work below the surface of the earth or underwater, where increased atmospheric pressures are used, and in those who enter high altitudes as in high-altitude flights or in travel by rockets. In the digging of tunnels, the shaft is maintained at a higher than normal atmospheric pressure in an effort to support the outside walls and to prevent seepage of water. The worker adapts to this increased pressure by entering air chambers, where he is slowly pressurized. Similarly, the air pressure within the suits of deep-sea divers is slowly pressurized for the same reasons. In both situations, as the individuals return to normal atmospheric pressures they must slowly be depressurized. If the pressure is reduced too quickly some of the gases that are dissolved in the blood at the higher pressure come out of solution to create small bubbles within the bloodstream. The oxygen in these bubbles is rapidly dissolved, but the nitrogen, which is a significant component of air, is less soluble and persists as bubbles of gas that block small blood vessels. These individuals suffer excruciating pain, principally in the muscles, which causes them to bend over in agony, and this disorder is popularly known as the bends. The same principle applies in those who sally high above the earth's surface. The normal atmospheric pressure of earth progressively diminishes at higher altitudes. If astronauts or those who fly in planes ascend too rapidly bubbles of nitrogen and oxygen likewise appear as they enter rarefied atmospheres. The interior of the plane is pressurized to maintain normal atmospheric levels. If such pressurization were to fail, the victims might suffer the bends. The three cosmonauts in the Russian spaceship Soyuz 11 were said to have died from such a failure, on June 30, 1971, during re-entry into the earth's atmosphere. In usual air flights, however, the altitudes are not high enough to make the bends a serious threat.

Development of the bends

Radiation injury. Radiation can result in both beneficial and dangerous biological effects. It is one of the effective forms of treatment for certain types of cancer. On the other hand, with the spectre of nuclear warfare ever present, the deleterious effects of radiation exposure are remembered. There are basically two forms of radiation: particulate, comprised of very fast moving particles (alpha and beta particles, neutrons, and deuterons), and electromagnetic radiation in the form of gamma rays and X-rays. From a biological point of view, the most important attribute of radiant energy is its ability to cause ionization—to form positively or negatively charged particles in the substance that it encounters. It is this ionizing effect that appears to cause destruction and even death of cells. The ionization of compounds within the cells theoretically leads to an alteration in the chemical composition of the cells. The damage may affect a wide range of chemical structures from simple inorganic molecules to complex structural proteins and enzymes. Deoxyribonucleic acid (DNA)—the genetic material of the cell—is highly susceptible to ionizing radiation. Cells and tissues may therefore die because of damage to enzymes, because of the inability of the cell to survive with a defective complement of DNA, or as a result of its inability to reproduce viable daughter cells. The cell is most susceptible to irradiation during the process of division.

Factors affecting severity of radiation injury

The severity of radiation injury is dependent upon the penetrability of the radiation, the area of the body exposed to radiation, and the duration of exposure. These variables all add up to the important concept of the total amount of radiant energy absorbed. A quantitative unit of radiation absorbed is called a rad. With total body irradiation, as occurred in the atomic blasts, or in accidents occurring at sites where radioactive material is being

handled, large areas are exposed to intense doses of radiation. In these circumstances, the body may receive within seconds enough radiant energy to cause either acute radiation sickness or death. It has been estimated that a dose as small as 500 rads absorbed instantaneously is sufficient to cause acute radiation sickness. This takes the form of nausea, vomiting, loss of appetite, headache, and a general feeling of weakness. Later, serious deficiencies in the blood cells and platelets may become evident as a result of radiation injury to the bone marrow, leading to hemorrhages or increased vulnerability to infections. Total body irradiation of 1,000 rads is almost certain to be fatal. At this level all of the cells of the body are seriously injured and there is total breakdown of normal fluid and electrolyte balance.

When the radiation exposure is confined to only a part of the body and is delivered in divided doses, a frequent practice in the treatment of cancer, the effect of the radiation depends on the vulnerability of the cell types in the body to this form of energy. Some cells, such as those that divide actively, are particularly sensitive to radiation. In this category are the cells of the bone marrow, spleen, lymph nodes, sex glands, and lining of the stomach and intestines. Their vulnerability may be related to the impact of radiant energy on the chromosomal material involved in mitotic division. In contrast, certain cells of the body such as nerve and muscle cells are peculiarly resistant to radiation. Here, too, the fundamental basis may be the fact that these cells never divide in adult life and so never enter a mitotic cycle. It should be pointed out that these general principles underlie the use of radiation in the treatment of cancer. In malignant tumours the cells are actively dividing and so are more sensitive to radiation. The goal of radiation therapy of tumours is to deliver a dosage to the tumours that is sufficient to destroy the cancer cells without injuring too severely the normal cells in the pathway of the radiation. Obviously, when an internal cancer is treated, the skin, underlying fat, muscles, and nearby organs are unavoidably exposed to the radiation. The possibility of delivering effective doses of radiation to the unwanted cancer depends on the ability of the normal cells to withstand the radiation.

Finally, there are probable deleterious effects of radiation in producing congenital malformations, certain leukemias, and possibly some genetic disorders (see RADIATION INJURY).

DISEASES OF IMMUNE ORIGIN

Types of immune disorders

The same immune system that protects against infectious disease is also capable of causing disease. These disorders of immunity fall into two broad categories: (1) hyperreactivity (increased reactivity) of the immune system to external or environmental antigens and (2) abnormal reactivity of the immune system to one's own tissues and cells—called auto-immunity.

The immune mechanism involves two quite separate types of response. In the first type an antigen stimulates the production of circulating antibodies that are elaborated by certain cells, called plasma cells, in the lymph nodes and bone marrow. The antibodies, known as immunoglobulins, pass from their sites of origin into the bloodstream and circulate freely, ready to neutralize antigens wherever they are found. Each antigen evokes its own specific antibody and the specificity is exquisitely precise. The antigen to its antibody is as specific as a key to its lock. The antibodies produced by antigens are protein molecules of different sizes. It has been possible to segregate these antibody molecules on the basis of size into classes of immunoglobulins, which have been designated as immunoglobulin A, immunoglobulin G, immunoglobulin M, immunoglobulin E, and, indeed, several other less important classes. The second type of immune response does not yield humoral-circulating antibodies but instead generates lymphocytes (a form of white blood cell) that are reactive against specific antigens. It is not yet understood why certain antigens evoke antibodies while other antigens evoke reactive lymphocytes.

Three pathways have been identified for the production of disease by circulating antibodies and reactive lymphocytes: (1) the release of vasoactive substances (substances that affect the blood vessels), (2) the deposit of antigen–antibody complexes in tissues, and (3) direct destruction of cells and tissues.

Release of vasoactive substances results from the union of certain forms of immunoglobulins with certain cells in the body. Immunoglobulins of the E class are principally implicated in this pathway. Certain antigens such as plant pollens, dusts, and products found within food evoke in the sensitized individual the formation of immunoglobulin E. When this form of antibody is produced it becomes fixed to one of two types of cells within the body, mast cells or basophilic leukocytes. Mast cells are scattered throughout the supporting tissues of the body. The basophilic leukocytes (white blood cells that stain readily with basic dyes) are found in the circulating blood. Presumably, when an individual first reacts to one of the specific types of antigens, the antibody so formed becomes adherent to these cells. These cells are rich in histamine, one of the major vasoactive substances found in the body. When released, histamine has the ability to dilate small blood vessels and capillaries, causing both redness in the tissue affected and leakage of fluid from the blood into the tissues. When a person is exposed a second time to the same antigen, the antigen combines with the antibody now fixed to these particular cells, and histamine is released. Localized areas of redness and swelling (hives) may occur on the skin. If a more severe reaction occurs, the affected person may have profound circulatory disturbances. Most common food allergies, such as sensitivity to strawberries or egg proteins, are mediated by the release of vasoactive substances.

Immune complex diseases

The second major pathway by which the immune response causes disease is through the formation of antigen–antibody complexes in the circulating blood. The diseases so produced are often referred to as immune-complex diseases. The antibody in these instances is not fixed to certain body cells but circulates freely in response to some antigenic challenge. The antigen and antibody in the circulation unite, producing immune complexes. When a particular concentration of antigen and antibody is present in the blood, the antigen–antibody complexes formed are of such a size as to become deposited in blood-vessel walls. Once so deposited the immune complex attracts a substance referred to as complement.

Complement is actually comprised of 11 separable components. Several of these components have the capability of attracting leukocytes. Other components are directly injurious to tissue cell, causing rupture of cell membranes. In diseases mediated by the formation of circulating immune complexes, it is actually the complement that does most of the damage. As the immune complexes become deposited in blood-vessel walls and attract complement, the complement in turn gathers leukocytes about it. Certain of these leukocytes, the polymorphonuclears, release powerful enzymes that damage the blood-vessel wall. The blood vessels may either become filled with blood clot or rupture. This type of immune-complex disease is particularly characteristic of immune reactions to antigens contained in certain types of bacteria. Specific strains of streptococci are especially likely to induce immune-complex disease, and, for reasons that are obscure, the immune complexes in streptococcal reactions tend to become deposited in the tiny blood vessels within the kidney and so induce a disease known as poststreptococcal glomerulonephritis. The kidney is actually an innocent bystander and is the victim of immune complexes caused by the bacterial infection.

Reactive lymphocytes

The third pathway by which the immune response causes disease is the direct destruction of cells and tissues by reactive lymphocytes. This pathway is principally involved in the immune reaction against transplanted organs. Each individual's cells have a specific constellation of antigens. These antigens are likely to be proteins and related compounds on the surface of the cell wall. No two individuals have the same precise combination of tissue antigens, with the possible exception of identical twins formed from a single ovum. Even the mother and the father of a child will have tissue antigens distinct

from those of their offspring. These tissue antigens, also known as transplantation antigens, are powerful stimuli to the immune response—in this particular setting mediated by reactive lymphocytes. They are responsible for the rejection reaction that follows transplantation of foreign organs into genetically dissimilar recipients; for example, the transplantation of a kidney from a donor to a recipient who is not an identical twin. How these lymphocytes actually cause cell and tissue destruction is somewhat obscure. There is some evidence that the lymphocytes elaborate a variety of factors that cause cell damage. The intensity of the immune response depends to a considerable degree on the extent of the antigenic differences between donor and recipient. This antigenic difference is sometimes referred to as the antigenic barrier, because it constitutes an obstacle to transplantation (see IMMUNITY; ALLERGY AND ANAPHYLACTIC SHOCK).

DISEASES OF BIOTIC ORIGIN

General considerations. *Types of agents.* Biotic agents include living forms that range in size from the smallest virus, measuring approximately 20 millimicrons in diameter (the finest human hair is about 250 times as thick), to tapeworms that achieve lengths of ten metres (about 33 feet). These agents are commonly grouped as viruses, rickettsia, bacteria, fungi, and parasites. The disease that these organisms cause is only incidental to their struggle for survival. Most of these agents do not require a human host for their life cycles. Many survive readily in soil, water, or lower animal species and so are harmless to man. In fact, some play a vital role in the economy of nature, decomposing organic material and returning it to the soil. They thus contribute to the enrichment of the soil, for example, and so contribute to man's survival. Other living organisms, which prefer or may require the temperature range of warm-blooded animals, may flourish on the skin or in the secretions of fluids of the mouth or intestinal tract but do not invade tissue or cause disease. Thus there is a distinction to be made between infection and disease. All animals are infected with biotic agents. Organisms that do not cause disease are termed nonpathogenic, or commensals. Organisms that invade and cause disease are termed pathogenic. The *Streptococcus viridans* organisms, for example, are found in the throats of over 90 percent of healthy persons. In this area they are not considered pathogenic. The same organism cultured from the bloodstream is highly pathogenic and usually indicates the presence of the disease subacute bacterial endocarditis (chronic bacterial invasion of the lining of the heart). In order for such nonpathogenic agents to achieve pathogenicity, they must obviously overcome the defenses of the host. Most biotic agents require a portal of entry through the intact skin or mucosal linings of the body. They must be present in sufficient number to escape the phagocytes. They must be capable of surviving the inflammatory and immune response. Ultimately, to induce disease, they must have sufficient virulence and invasiveness to cause significant tissue injury.

Invasiveness

Factors that affect invasiveness. Invasiveness is the capability of penetrating and spreading throughout tissues. Remarkably, little is known of the factors that condition it. In a few instances enzymes produced by biotic agents have been identified that are capable of breaking down the integrity of the supporting tissues of the body, thereby preparing a pathway for the spread of the organism. These enzymes, known as hyaluronidases, are found in certain highly pathogenic forms of streptococci that characteristically cause rapidly spreading infections. Streptococci may also produce streptokinases, substances that dissolve blood clots and further facilitate the spread of the organisms. Only very few bacteria elaborate such enzymes, however, and the general attribute of invasiveness is still poorly understood. There are marked differences in invasiveness to be found among the various types of bacteria. The organism that causes diphtheria (*Corynebacterium diphtheriae*), for example, is capable of invading only the surface cells of the mouth and throat. The disease that it produces results from the elaboration of a powerful exotoxin (a toxin that is produced by the organism and released into the surrounding tissues) that is absorbed into the bloodstream from the local infection within the throat. This exotoxin evokes the major damage in the heart and the nervous system. The diphtheria bacillus, therefore, is an example of a serious infection in which the organism has low invasiveness. In contrast, the organism that causes syphilis (*Treponema pallidum*) has a high degree of invasiveness. It is one of those rare biotic agents that are capable of penetrating intact skin and mucosal linings of the body. Generally, it is acquired by a normal host during sexual intercourse with an infected individual. It creates a small local lesion, a shallow ulcer, known as a chancre, on the penis or on the external genitalia of the female. From this lesion the organism soon (one to two weeks) penetrates into the bloodstream and disseminates throughout the body. The small local lesion, the chancre, is known as the primary stage of syphilis. The spread throughout the bloodstream is known as the secondary stage of syphilis and is characterized by the appearance of a rash. This phase is followed in months to years by the devastating tertiary stage of syphilis, which may cause serious cardiovascular or central-nervous-system disease.

The invasiveness of viruses undoubtedly is facilitated by their extremely small size, but, because of this size, the exact mechanism is difficult to study. In the case of fungi and parasites, the invasiveness is related to the life cycle of the organism. The formation of tiny spores by fungi and the smaller reproductive forms of the parasites provide vehicles by which infection may be drawn into the lungs or may pass through tiny defects in the skin or mucosal linings of the various openings and tracts of the body.

Virulence

In general, virulence is the degree of toxicity or the injury-producing potential of a micro-organism. The words virulence and pathogenicity are often used interchangeably. The virulence of bacteria usually relates to their capability of producing a powerful exotoxin or endotoxin. Invasiveness also adds to an organism's virulence by permitting it to spread. But there are additional considerations, poorly understood, that modify the virulence of bacteria. Indeed, specific strains within a single genus may have greatly varying virulence. There are 32 types of pneumococci, for example, all of which may cause pneumonia (inflammation of the lungs). Among these, three types cause more severe disease than do the others. To some extent these differences are related to the powerful toxic polysaccharide capsule of the most virulent strains of this genus. In this particular instance, the polysaccharide is a powerful antigen and therefore imparts virulence to the organism. In many other instances, however, the attribute of virulence is poorly understood.

So far, diseases caused by biotic agents have been considered in terms of the role of the invader. Equally important is the role of the host, the individual who contracts the disease. Any infectious disease is a battle between the invader and the defender. Virulent organisms may be capable of inducing serious illness even in the most robust. The converse is perhaps more important. The weak host is prey to many forms of biotic infection, even those of low virulence and invasiveness. Some of the more important of the many factors that condition the level of resistance to biotic infection in the individual are touched upon in the following paragraphs.

Factors that affect predisposition. Age is one of the important influences in the predisposition to biotic disease. The two extremes of life, infancy and old age, are periods of maximal vulnerability. At birth, the infant possesses antibodies of maternal origin that provide some protection for the first months of life. When these are lost there is a period of high vulnerability until the infant develops his own immune responsiveness, about halfway through the first year of life. Later in childhood, when the child has an adequately functioning immune system, he is still somewhat vulnerable because he has not yet had exposure to many organisms and so has many immunologic deficits. Thus such infections as mumps, measles, and whooping cough are known as childhood diseases. The increased

vulnerability of old age to infectious disease may be related to nutritional deficiencies, which are common in the elderly. Poor blood supply because of progressive arteriosclerotic disease (hardening of the arteries) may also play a role. Diseases already present, such as heart failure and cancer, facilitate biotic infections.

The physiologic condition of the host is thought in some way to influence the vulnerability to biotic disease. There is a prevailing opinion, difficult to substantiate by objective data, that the robust, well-conditioned individual is less vulnerable than his more fragile colleague. Perhaps the underlying variables are the state of nutrition, the adequacy of blood supply to the tissues, and the total homeostatic responsiveness of the person in good physical condition.

Role of nutrition

Nutrition plays an important and well-documented role in predisposition or resistance to infectious disease. There are abundant observations from both clinical and animal studies that dietary factors such as proteins and vitamins A, D, C, and the B complex are necessary in adequate amounts to maintain a high level of resistance to infections. It has been shown in animals that deficiencies of any of these nutrients increase susceptibility to infection caused by pneumococci, rickettsia, and *Salmonella* organisms. Paradoxically, however, deficiencies of the vitamin B fractions in mice decrease susceptibility to the virus that causes poliomyelitis. It may well be that when this vitamin is not present in sufficient amounts, nerve cells cannot support the growth of the polio virus. Proteins play an important role in protecting against infections by providing an adequate substrate for the synthesis of antibodies. Proteins undoubtedly also play a role in the cell growth necessary for repair and regeneration of injury. Vitamin C is also necessary in the inflammatory process that occurs in infectious disease. War-torn and malnourished populations are frequently decimated by such infections as tuberculosis, bubonic plague, and malaria. Nutrition is then of unchallenged importance.

There is a well-known group of genetic disorders characterized by deficient function of the immune system. Some persons with these disorders are incapable of forming antibodies. Others cannot generate lymphocytes that are involved in the immune response. These immunologically handicapped persons are exceedingly vulnerable to infections; many die in early life because of unresisted biotic disease.

A host of metabolic defects predispose to biotic infections. Diabetes mellitus (a metabolic disease in which inadequate secretion of insulin leads to high blood sugar and a large number of other effects) is associated with vulnerability to infectious diseases. It is not certain that diabetic persons develop more infections, but, in such persons, once an infection begins it tends to become more severe. A number of factors may be responsible for the diabetic's vulnerability. He has a marked predisposition to arteriosclerosis (hardening of the arteries), and such vascular disease reduces the blood supply to many of the tissues and organs of the body and so impairs their defensive capability. Additionally, blood with high levels of sugar may serve as a better growth medium for the bacteria than normal blood. The diabetic is subject to acidosis (*e.g.*, the blood and the body tissues are less alkaline than normal) and nutritional imbalances, and these may also contribute to the predisposition to infection.

Increased levels of adrenal hormones (steroids), whether produced by the glands of a person or given as treatment for some other disease, markedly increase the susceptibility to infections. Tuberculosis often becomes activated or disseminated in such circumstances. The steroids also suppress the inflammatory response.

Drugs as factors predisposing to disease

Therapeutic agents, paradoxically, have become important factors in predisposing to disease of biotic origin and indeed in altering the incidence patterns of infectious disease. The drugs that are principally involved include those used to suppress the immune response intentionally, as well as the host of antimicrobial and antibiotic agents now employed in the treatment of infectious disease. Immunosuppressive drugs are used for the treatment of persons about to receive an organ transplant, to block the immune response. Similarly, these drugs are employed in the treatment of the autoimmune diseases. Whatever the reason, immunosuppressive therapy is a two-edged sword. While it may be vital for the purposes of supporting the transplantation or treating the autoimmune disease (disease involving an immune response to one's own tissues), it renders the patient vulnerable to attack by biotic agents. Indeed, these immunologically crippled persons become susceptible to organisms of extremely low virulence.

Antimicrobial drugs have also been two-edged swords. A patient suffering from a streptococcal disease, for example, may be appropriately treated with penicillin. Certain strains of staphylococci, however, are resistant to penicillin. Although the streptococcal organisms, as well as other commensals, may be eradicated by the antibiotic, the staphylococci begin to proliferate, possibly because the competition with other bacteria for nutrients and food supply has been removed. In this noncompetitive situation they may cause disease. More powerful antibiotics may destroy all bacteria, including staphylococci, but permit the unrestrained proliferation of fungi and other agents of low virulence that are nonetheless resistant to the antibiotic. Thus antibiotics have changed the entire frequency pattern of biotic disease. Organisms that are susceptible to antibiotics have become relatively infrequent causes of disease and rare causes of death. Such infections when recognized can be promptly controlled. But organisms that have proved to be more resistant to antibiotics have become the more common causes of serious clinical infection. For this reason certain forms of drug-resistant bacteria referred to as gram-negative rods (*Escherichia coli, Aerobacter aerogenes, Pseudomonas aeruginosa,* and strains of *Proteus*) as well as fungi have emerged as the important biotic causes of death.

Types of viral disease

Viral diseases. Of the many existing viruses a few are of great importance as causes of human sickness, being responsible for such diseases as smallpox, poliomyelitis, encephalitis, influenza, yellow fever, measles, mumps, and such minor disorders as the common cold, warts, and cold sores of the lips (herpes simplex). At present, viruses are also highly suspect as possible causes of cancer in man. Indeed, there is a large body of evidence pointing to viruses as causes of cancer in lower animals, but, while there are many hints of similar phenomena in man, the data are insufficient for solid conclusions.

Viruses may survive for some time in the soil, in water, or in milk, but they cannot grow and divide unless they parasitize or invade living cells. These agents are principally made up of nucleic acids (deoxyribonucleic or ribonucleic acid) enclosed within a protein coat. It is the outer coat that provides their antigenic specificity and their ability to penetrate cells, but it is the nucleic acid component that replicates (proliferates) within the cell. How these nucleic acids in the course of their replication cause cell injury is still not entirely clear.

Certain viruses proliferate within the host cells and accumulate in sufficient number to cause rupture of the cells. Others multiply within the cell body and compete with the host for nutrition or vital constituents of the cell's metabolism. Both types of viruses are said to be cytocidal (cell killing). The virus causing smallpox is such an agent. This disease, fortunately largely eradicated by protective immunization (vaccination) of entire populations, is closely related to chickenpox. It is, however, a far more serious disorder and when it was prevalent caused many deaths. Smallpox is characterized by the appearance of large vesicles (blisters) on the skin, producing a severe generalized skin eruption over the entire surface of the body. Blisters are formed where the cells of the skin have been killed and fluid has accumulated. The virus simultaneously attacks cells in the internal organs, such as the liver, and is often lethal.

Poliomyelitis

Poliomyelitis is another example of a viral disease in which the virus grows within cells and causes their destruction. The polio virus is present in the saliva and

stools of infected individuals. The next host presumably becomes infected by swallowing or inhaling virus-contaminated material such as foods, water, or airborne droplets produced by coughing or sneezing. The virus penetrates the tissues of the mouth and throat and grows in cells there without causing much disability, but, from this vantage point, it is apparently spread through the bloodstream to the brain, where it finds its most desired environment—nerve cells within the brain and spinal cord that innervate muscles. By destroying these nerve cells the virus causes paralysis of muscles. When it affects the muscles required for breathing, it may cause death.

Certain viral agents, particularly those capable of producing tumours in lower animals, flourish within cells and stimulate the cells to active growth. These viruses are referred to as being oncogenic (tumour producing). The number of virus-induced tumours in lower animals is now large. While the only proven virus-induced tumour of man is the skin wart, there is much more evidence relating viruses to some of the tumours of man (see CANCER).

Most viral infections occur in childhood. This age distribution has been explained on immunologic grounds. Viruses usually induce a firm and enduring immunity. On first exposure to a virus, children may or may not contract the disease, depending on their resistance, the size of the infective dose of virus, and many other variables. Those who contract the disease, as well as those who resist the infection, develop a permanent immunity to any further exposure. By either pathway, as children grow older they progressively gather protection against viral infections. Consequently, the incidence of these infections falls in adult and later life. The frequency of common colds is now explained on the grounds that a host of different viral agents all induce similar respiratory infections and, while a single attack confers immunity against the specific causative agent, it provides no protection against the rest.

Viral diseases are resistant to the usual forms of drug therapy now available, such as penicillin and the other antibiotics. This point is made because of a distressing tendency among individuals to take penicillin or another antibiotic for a common cold.

Rickettsial diseases. The rickettsial diseases of man are caused by micro-organisms that fall between viruses and bacteria in size. These minute agents are barely visible under the ordinary light microscope. Like viruses, they multiply only within the cells of susceptible hosts. They are found in nature in a variety of ticks and lice, and when transmitted to man by the bite of one of these arthropods usually cause acute febrile (fever-producing) illnesses, most of which are characterized by skin rashes. When these agents invade the cells of man they apparently subvert the metabolism of the host cell to the purpose of the micro-organism and so cause death of the cell.

Typhus

Epidemic typhus is an acute febrile infection caused by *Rickettsia prowazeki*. The disease is characterized by involvement of many organs, particularly the brain. This agent is found in nature in the human body louse. The louse itself becomes infected by biting a person who has the acute disease. The rickettsia flourish within the intestinal tract of the louse and are transmitted to the new host by the bite of the louse. The louse itself, however, may be carried from man to man by rats or other vermin. Typhus may therefore become epidemic when living conditions deteriorate. It has been estimated that perhaps 25,000,000 cases occurred in Russia during World War I.

Bacterial diseases. The diseases produced by bacteria are the most common of infectious biotic disease. They range from trivial skin infections to such devastating disorders as bubonic plague and tuberculosis. Various types of pneumonia; infections of the cerebrospinal fluid (meningitis), the liver, and the kidneys; and the venereal diseases syphilis and gonorrhea are all forms of bacterial infection. Bacteria fall into three large groups, spherical cocci, rod-shaped bacilli, and spiral forms known as spirochetes. Some of the large bacilli are as long as ten microns, while some of the small cocci and coccobacillary forms are as small as 175 millimicrons.

All bacteria induce disease by one of three methods: (1) the production of a harmful chemical substance that is secreted or excreted by the bacterium (exotoxin), (2) the elaboration of a harmful chemical substance that is liberated only after disintegration of the micro-organism (endotoxin), or (3) the induction of sensitivity within the host to antigenic properties of the bacterial organism. A few brief examples will serve to illustrate these methods of producing disease.

Exotoxic infection. Mention has already been made of diphtheria as a disease caused by a powerful exotoxin elaborated by the *Corynebacterium diphtheriae*. An additional dramatic example of an exotoxic disease is botulism, caused by *Clostridium botulinum*. This organism is anaerobic; that is, it can only grow in the absence of air. It is therefore a particularly dangerous contaminant of improperly canned or preserved foods. If the organism has not been destroyed by the methods of sterilization, it will grow under the anaerobic conditions obtained in the canning or preserving and release powerful exotoxins into the canned food. Food poisoning appears, therefore, within hours of the ingestion of these preformed toxins. The organism itself contributes nothing directly to the production of the disease; indeed, as soon as the tin or jar is opened the organism is rendered innocuous, because it cannot grow when exposed to air. The absorbed toxins paralyze nerve endings paralyzing all muscles, and affected persons are unable to speak, swallow, or breathe. Death is usually caused by respiratory-muscle paralysis.

Endotoxic infection. An example of an endotoxic infection is typhoid, caused by *Salmonella typhosa*. The infection is almost always acquired by the ingestion of micro-organisms in contaminated food, milk, or water. Once swallowed, the organism creates a local infection within the intestinal tract and then spreads rapidly through the lymphatic system into the blood. The endotoxin is largely concentrated in the cell wall of the bacteria; as bacteria are destroyed by normal body defenses the endotoxin is released to produce the manifestations of a systemic infection. Fever, headache, the development of a skin rash, diarrhea, and muscular weakness are the predominant symptoms. Because the organism is susceptible to antibiotic therapy, deaths are usually infrequent.

Carrier state

Typhoid provides an opportunity to touch upon another bacteriologic phenomenon, the development of a carrier state. About 10 percent of persons who recover from the acute typhoid infection continue to harbour organisms and excrete them in their stools for eight to ten weeks, or more. These carriers constitute reservoirs of infection and may spread the disease to others. When engaged in food-handling occupations, they represent an obvious threat to the public health.

Sensitization. Tuberculosis exemplifies a disease resulting from the development of sensitization to the products of an organism. The organism, *Mycobacterium tuberculosis*, is of low virulence, but once the patient is infected with the organism he develops a state of sensitization to certain products formed by the bacteria. It is this adverse immune reaction that is responsible for the death of tissues and for the serious manifestations of the disease. Tuberculosis is usually contracted as a respiratory infection by the inhalation of organisms from an infected individual. Such an individual may have obvious signs of tuberculosis such as cough, fever, and night sweats, or he could be remarkably free of signs and symptoms of the disease and yet be capable of spreading it. If infective particles are coughed up he may contaminate his environment or persons with whom he comes into contact and so spread the disease. The new host inhales the organism. At first these produce only a trivial inflammatory reaction within the lungs, but in the course of a few weeks sensitization occurs to certain of the protein and fatty components of the bacterial cell. Once this sensitivity emerges it induces necrosis (death) of the infected tissues, usually portions of the lung. On first exposure to the organism the individual, usually a child, develops what is called primary tuberculosis. Generally the organisms tend to local-

ize, at first in the periphery of the lungs. This primary stage may become quiescent or even heal. Weeks to months or even years later, a reactivation or reinfection will cause secondary tuberculosis, in which the organism spreads via the lymphatic channels to localize in the uppermost portions of the lung, known as the apices. It is in the secondary state that the most damage occurs, and the infection may lead to extensive necrosis of regions of the lung, with the formation of large cavities. If the infection spreads into the bloodstream, seeding of any organ or tissues of the body may occur. These disseminated localizations, known as miliary spread, are most often encountered in the brain, spleen, liver, bone marrow, kidneys, and male and female genital tracts. Pulmonary (lung) tuberculosis alone is a serious infection; miliary tuberculosis is even more threatening and is often fatal. Fortunately, when the disease is identified, effective therapy is available for the cure or at least the control of the disease, which was once devastating.

Cavitation in tuberculosis

Fungi and other parasites. Diseases caused by fungi and parasites are relatively uncommon. Fungal infections, also known as mycotic infections, may affect the skin surfaces or the internal organs of the body. The superficial mycotic infections are generally not serious and include such well-known disorders as athlete's foot (caused by *Tinea pedis*) and ringworm (caused by the agents *Microsporum* and *Trichophyton*). Deep mycotic infections such as histoplasmosis, blastomycosis, coccidioidomycosis, and candidiasis are potentially life threatening.

The other parasites that attack man range from unicellular organisms such as *Entamoeba histolytica* to such multicellular forms as tapeworms and roundworms. Most parasitic infestations are encountered in the less privileged areas of the world, where sanitation is not optimal. Indeed, parasitic infestations constitute major causes of death in regions of Central and South America, Africa, India, and Asia. Parasitic infestations are uncommon in the Northern Hemisphere, but on occasion certain forms such as amebiasis, tapeworms, and trichinosis are encountered even in the industrialized cities of the world. Trichinosis is an excellent example. Trichina worms (*Trichinella spiralis*) parasitize pigs, and their larvae are encysted within the muscles of these animals. When pork is inadequately cooked the cysts are not destroyed and so the disease may be contracted. When such pork is eaten the cysts develop into trichina worms in the intestinal tract of man. Here the worms reproduce and spread to the muscles of man. The infestation causes muscle aches and pains, fever, and weakness. When the heart and the muscles of respiration are severely affected, death may result. Trichinosis is still encountered sporadically throughout North America and Europe. In every instance that can be traced, the disease comes from inadequately cooked pork or pork products.

ABNORMAL GROWTH OF CELLS

Among the various diseases of man, those resulting from abnormal cell growth are the second most common cause of death. Cancer, for example, a form of abnormal cell growth, accounted for approximately 16 percent of all mortality in the United States in 1969. Only heart disease outranked it. Besides cancer, a malignant tumour, there are benign tumours, which rarely produce serious disease. The two forms of tumours are collectively spoken of in medicine as neoplasms (new growths), and their study is known as oncology. In addition to neoplasms, there are other forms of abnormal cell growth, known as hyperplasias (increased growth of cell), that cause less serious disease. Hyperplasia causes enlargement of tissues and organs, but the abnormal cell growth does not go out of control to produce a tumour.

Tumours and hyperplasias

Hyperplasia. Hyperplastic states result from the controlled proliferation of normal cells, in contrast to tumours, in which the growth is uncontrolled and the cells are abnormal. Hyperplasia of the cells of the thyroid gland, for example, causes some enlargement of the gland (known as goitre), but the hyperplastic cells are essentially normal in their appearance and in their manner of mitotic division. Because the new cells, like their antecedents, are capable of producing thyroid hormone, the increased number of cells causes excess activity of the thyroid (hyperthyroidism).

Most hyperplasias, wherever they occur, are the result of some stimulatory influence, such as hormonal stimulation or chronic irritation. The increased thickness of the skin that is represented by the ordinary callus or corn on the foot, for example, is due to controlled growth of the cells of the skin in the immediate area of chronic irritation or pressure from ill-fitting shoes. In contrast to the irritation that causes the callus or the corn, the cause of hyperplasia encountered in the thyroid gland is currently thought to be some thyroid-stimulator substance that appears in some individuals for obscure reasons. Hyperplasia of the prostate gland in elderly males is attributable to a relative excess of steroid hormones derived from the adrenal. These adrenal hormones (estrogens) are produced throughout life in both the female and the male. During a man's active reproductive life, the estrogens are more than counterbalanced by the male sex hormones that are produced by the testes. With aging, the testes become functionally less active and the levels of male sex hormones fall. As a consequence, the relative excess of estrogens of adrenal origin causes hyperplasia of the prostate.

It is still not known why hormones and irritation cause cells to grow. Evidence suggests that cell growth, whether normal or hyperplastic, is a fundamental attribute of all cells and is kept under control by mechanisms either within the cell or in the immediate vicinity of the cell. In other words, the normal growth of man from the time of conception is a regulated sequence of control mechanisms. When the controls are turned off, cells immediately begin to proliferate by division. Growth necessary to achieve full adult size is the consequence of release of controls. Hyperplasia is a consequence of the turning down of these controls by hormones or irritation. Cancer results when the controls over certain cells are totally lost.

The cause of hyperplasia

Neoplasms. Cancers are by far the most important form of abnormal cell growth. When cells become cancerous they appear to have escaped from control mechanisms. They grow in a haphazard, disorderly manner and come to have a more or less wild, anarchic appearance. They vary widely in size and shape and no longer resemble the cells from which they arose. Although little is known about the causes of loss of growth controls that permit the development of neoplasms in man, in experimental animals cancers can be induced by a number of agents, including (1) certain chemicals, (2) certain viruses, (3) radiation, and (4) a number of other less well defined influences such as hormones and chronic irritation.

Many chemicals have been shown to be capable of causing cancer in animals. Some of them, known to react with the deoxyribonucleic acid of cells, are suspected of affecting control mechanisms of growth. A few chemicals have been implicated in the production of cancer in man. Workers in rubber industries and dye factories formerly suffered a higher incidence of bladder tumours than normal control populations. It was discovered that some of the solvents used in these industries, the naphthylamines, are capable of causing bladder cancer in animals and were responsible for the increased frequency of these cancers in man. Asbestos fibres such as are used in the manufacture of shingles for houses and brake linings are another hazard to man. When inhaled these fibres may cause certain forms of cancer in the lungs. The mold *Aspergillus flavus*, which grows on peanuts and grains, elaborates a substance that induces liver cancer in animals and is suspected of causing an analogous tumour in man. There is a very high incidence of this form of tumour in the underdeveloped nations of the world, where grains are poorly preserved and where they comprise a large part of the diet.

There is also strong evidence implicating the chemicals in cigarette smoke in the production of lung cancer. The chemicals may act in concert with other influences such as bacterial or viral lung infections.

Table 3: Major Hormones of Man

hormone	source	target organ	action
Growth hormone	anterior pituitary	all cells	promotes growth
Thyrotropin	anterior pituitary	thyroid gland	controls secretion of thyroxine
Corticotropin	anterior pituitary	adrenal glands	controls secretion of adrenal hormones
Antidiuretic hormone (ADH)	posterior pituitary	kidney	promotes retention of water
Oxytocin	posterior pituitary	breast and others	promotes lactation
Aldosterone	adrenal gland	kidney	regulates body fluids
Cortisone	adrenal gland	all cells	manifold; plays central role in sugar and fat metabolism
Thyroxine	thyroid gland	all cells	speeds up metabolic machinery
Insulin	pancreas	all cells	facilitates uptake of glucose into cells
Parathormone	parathyroid gland	gastrointestinal tract, bone, kidney	controls serum calcium level
Erythropoietin	kidney	bone marrow	stimulates red-blood-cell production
Testosterone	testis	many tissues	plays role in growth and development of reproductive system in males
Estrogens	ovary	many tissues	in females, similar to testosterone

Mention was made in an earlier section of the possible role of viruses in the induction in man of tumours and, particularly, cancer. It was pointed out that certain viruses cause cells to divide rapidly. How viruses produce this effect is unknown, but it is believed that the DNA and RNA of the virus are incorporated within the genetic apparatus of the cell and in some way modify it so that normal repressor control mechanisms are lost.

Viral tumours in man

The only definitely established viral tumour of man is the common wart, and this is hardly a significant neoplasm. Another tumour of man strongly suspected of having a viral causation is the Burkitt lymphoma, a rare form of cancer most often encountered in South Africa but also occasionally encountered in other areas of the world. When originally recognized in South Africa, it was found in a geographic locale with a high median temperature and a high level of humidity. Moreover, this tumour tended to occur in young children, a distinctly unusual age group to have a high incidence of cancer. All of these observations led to the suspicion that a virus might be the cause. The geographic localization of temperature and humidity raised the possibility that some insect spreads the virus from an infected person to a normal susceptible host. Children might particularly be affected because they had no previous exposure to this agent and would therefore be completely susceptible, or nonimmune. Intensive study of the Burkitt lymphoma has disclosed that it indeed does contain a viral agent and that this virus is capable of transforming cells in tissue culture from normal growth patterns to abnormal cancerous growth patterns. Final proof is lacking, however, because it has not been possible to inoculate this virus into human beings.

Radiant energy is a well-known cause of cancer in animals and man. The pioneer workers in the development of X-rays did not realize the danger involved and so were unduly exposed to this form of energy. Many developed cancers of the skin and arms where they had received excessive exposure. The survivors of the atomic bombs dropped on the Japanese cities of Hiroshima and Nagasaki in 1945 have suffered various forms of cancer, particularly the leukemias, cancers that involve the white blood cells.

It was clear that many influences may induce cancer, but what role these play in the overall problem of cancer in man is still uncertain. The environment contains many potential cancer-producing influences: chemicals, viruses, and radiant energy from the sun; some investigators imply that the occurrence of cancer in the advanced years of life suggests the possibility of the accumulation over decades of enough of these influences to trigger the formation of a malignant tumour.

Cancer may arise in any organ or tissue. Some tumours grow more rapidly and wildly than others, but it is believed that all begin as small aggregations of abnormal cells that take years to develop into tumorous masses. Indeed, there are good reasons for believing that it may take ten years or more for a cancer to evolve from its early origins to a clinically recognizable disease. Cancers do not explode into being; during the long years of their development they can easily be removed by surgery if discovered. They rarely cause symptoms or illness during this developmental phase, and consequently tests must be sought that can be applied to apparently normal individuals to detect these minute cancers. As cancers grow they invade surrounding tissues. Thus a cancer that begins in the thyroid, for example, may spread into the structures of the neck, such as the trachea (the windpipe), or even more widely. Obviously, when such spread has occurred it is much more difficult, and perhaps impossible, to remove the tumour surgically. And, even more regrettably, most cancers have the ability to seed distant sites, a phenomenon termed metastasis. A cancer of the lung may metastasize to the liver, adrenals, and other tissues in the body. When such distant implants develop, it is evident that surgical removal is impossible, because not all of the implants can be found and removed. The hope, then, is to discover the cancer before it has invaded too widely and before it has metastasized (see CANCER).

Slow growth of cancers

DISEASES OF METABOLIC–ENDOCRINE ORIGIN

The term metabolism encompasses all the chemical reactions vital to the growth and maintenance of the body: Derangements in metabolism are found in almost every disease condition. Most of these derangements are secondary; *i.e.*, they result from some other basic disorder (infection, kidney disease, or heart disease, for example). A few primary metabolic disorders were mentioned in the discussion of genetic disease. In these, small genetic mutations were shown to lead to derangements in the synthesis of specific proteins. At this point another group of primary metabolic disorders—those associated with hormonal derangements—will be touched upon.

Hormones are rather large organic molecules secreted in small amounts by specific cells in the various endocrine (ductless) glands. These secretions are carried by the blood to distant sites (target organs), where they act to regulate specific chemical reactions. Table 3 summarizes the major hormones in man, their source, target organ, and principal action.

Cause of endocrine disease

All endocrine disease stems from either an overproduction (hyperfunction) or underproduction (hypofunction) of some hormone-secreting endocrine gland. There are relatively few causes of hormone overproduction. In general it results from hyperplasia, an increase in the number of hormone-secreting cells in a specific endocrine gland. A tumour of an endocrine gland may also cause hypersecretion of a hormone. Although most endocrine neoplasms are benign tumours, the resulting hypersecretion of hormone can have far-reaching effects.

The tiny pituitary gland, for example, tucked into the base of the skull, produces many hormones that have far-ranging effects, mostly controlling the function of the other endocrine glands, such as the adrenals, ovaries, and testes. Acromegaly is a rare endocrine disease caused by excess secretion of pituitary growth hormone in the adult. The disease is characterized by an overgrowth of most of the organs in the body, but especially of bone and connective tissue. A person with the disorder usually has a large head, big, brawny hands, and huge feet. Postmortem examination reveals the presence of small, benign pituitary tumours in 75 percent of the cases. If the excess growth hormone appears during childhood, it causes exaggerated growth, including extreme height. Another example, a pheochromocytoma, is a small, benign tumour of the adrenal gland having far-reaching effects. These tumours elaborate an excess of catecholamine hormones (norepinephrine and epinephrine), resulting often in severe hypertension (high blood pressure) and subsequent vascular disease.

An example of hormone overproduction because of hyperplasia is hyperthyroidism, the disease produced by an excess of thyroid hormone. It is characterized by a rapid pulse, increased sweating, weight loss, heat intolerance, and frequent disturbances in the heart rhythm. A strange substance called long-acting thyroid stimulator (LATS) is found in the blood of persons with Graves' disease, a variant of hyperthyroidism that is marked by exophthalmos (bulging eyeballs) and diffuse goitre (thyroid-gland enlargement). LATS, which is thought to be responsible for the thyroid hyperplasia, has the same physical properties as an antibody. This suggests that the disease may have an immune origin.

Cause of Cushing's syndrome

Cushing's syndrome, an exception to the generalization that hypersecretion of hormones is due to either neoplasia or hyperplasia, results from an overproduction of the adrenal steroid hormones (such as cortisol). Persons with this disorder have a characteristic redistribution of fat centring chiefly around the upper part of the back (buffalo hump) and the face (moon faces). They also suffer from impaired carbohydrate metabolism, psychiatric disturbances, and, commonly, high blood pressure. Although the disease is occasionally caused by tumours or by hyperplasia of the adrenals, in most instances it is not. It has been suggested that the disease results either from excessive adrenal-stimulating hormone from the pituitary or from an enzymatic defect that leads to inadequate metabolic breakdown of the steroid hormones, but neither presumed mechanism has thus far withstood the trial of intensive laboratory investigation.

Underproduction of hormone is most often the result of destruction of hormone-secreting cells. This destruction may be caused by infection, infarction (tissue death due to loss of blood supply), or obliteration of endocrine glands by cancer. Underproduction of hormone may also result from failure of the gland to undergo normal fetal development or may be a feature of a hereditary disease (as in diabetes mellitus).

Destruction of endocrine tissue by infection is most often due to tuberculosis. Addison's disease, also known as adrenal insufficiency, was once due almost exclusively to a tuberculous infection of the adrenal glands.

Cretinism, a condition that includes mental retardation and physical stunting, results from the congenital lack of thyroid hormone. The hormone lack is caused by iodine deficiency during the critical growth periods of fetal and neonatal life or, rarely, failure of the thyroid gland to develop at all (thyroid aplasia).

Death of tissue due to inadequate blood supply (infarction) is the most common cause of hypopituitarism and is seen almost exclusively in new mothers (postpartum necrosis of the pituitary). The exact nature of this process is not known, but it is assumed to be related to severe hemorrhages occurring during childbirth.

Treatment of endocrine disease involves either hormone supplementation, in the case of hypofunction, or destruction of endocrine gland tissue, in cases of hyperfunction, by surgery or radiation (see ENDOCRINE SYSTEM DISEASES AND DISORDERS).

DISEASES OF NUTRITION

Diseases of nutrition include the effects of undernutrition, prevalent in less developed areas but present even in affluent societies, and the effects of nutritional excess.

Diseases of nutritional excess. Obesity, perhaps the most important nutritional disease in the United States and in Europe, results, usually, from excessive caloric intake, although emotional, genetic, and endocrine factors may be present.

Obesity predisposes toward several serious disorders, including a state of chronic oxygen deficiency, called the hypoventilation syndrome; high blood pressure; and atherosclerosis, a degenerative condition of the blood vessels that is discussed further below.

Excessive intake of certain vitamins, especially vitamins A and D, can also produce disease. Vitamins A and D are both fat soluble and tend to accumulate to toxic levels in the bodily tissues when taken in excessive quantities. Vitamin C and the B vitamins, soluble in water, are more easily metabolized or excreted and, therefore, rarely accumulate to toxic levels.

Types of nutritional deficiency

Diseases of nutritional deficiency. Nutritional deficiencies may take the form of inadequacies of (1) total caloric intake, (2) protein intake, or (3) certain essential nutrients such as the vitamins and, more rarely, specific amino acids (components of proteins) and fatty acids.

Protein–calorie malnutrition is still prevalent in certain areas. It has been estimated that about two-thirds of the world's population has less than enough food to eat. Whereas the average adult in economically developed countries consumes 3,000 or more calories per day, in less developed areas the underprivileged may have less than 1,000 calories of food available per day. Not only is the quantity inadequate but the quality of the food is nutritionally deficient and usually lacks protein. In these deprived areas the malnutrition has its greatest impact on the young; the death rate in children up to the age of five may reach 50 percent, in contrast to a rate of about one per 100,000 in the United States. The deaths from protein–calorie malnutrition result from the failure of the child to thrive, with progressive weight loss and weakness, leading to an infection, usually some form of gastrointestinal bacterial or parasitic disease. In other circumstances adequate calories may be available, but a deficiency of protein induces a disorder known as kwashiorkor. This is seen especially in children who are weaned because of the birth of another child. They are fed on rice water, which may contain enough carbohydrate calories but is lacking in protein calories. Weakness, wasting, liver disease, and bacterial and parasitic infestations follow. Even if these children survive to adulthood they rarely achieve a state of reasonable health.

Vitamin deficiencies, the most important forms of selective malnutrition, may arise in a variety of ways, the most common and the most important being an improper, inadequate diet. When the total caloric intake is inadequate, vitamin deficiencies may also occur, but in these circumstances the more profound lack of calories and proteins masks the lack of vitamins.

Secondary vitamin deficiencies

Vitamin deficiencies may also be encountered despite a diet that is apparently adequate nutritionally. One source of such a deficiency, called secondary, is interference with absorption of the vitamin. Pernicious anemia is a classic example of this phenomenon. The absence of the substance known as intrinsic factor, normally found in the stomach lining, leads to the inability to absorb and store one of the B vitamins (vitamin B_{12}), necessary for the normal formation of red blood cells. The basis of pernicious anemia, then, is a lack of absorption of vitamin B_{12}. The absence of certain digestive enzymes, as is found in pancreatic disease, can lead to the inability to digest and absorb fats and the fat-soluble vitamins (A, D, E, and K). Impaired uptake of vitamins may be encountered in gastrointestinal diseases. Some of these diseases reduce the absorptive function of the bowel. Similarly, diseases associated with severe, prolonged vomiting may interfere with adequate absorption.

Avitaminoses (vitamin lack) may be encountered when there are increased losses of vitamins such as occur with

chronic severe diarrhea or excessive sweating or when there are increased requirements for vitamins during periods of rapid growth, especially during childhood and pregnancy. Fever and the endocrine disorder hyperthyroidism are two additional examples of conditions that require higher than the usual levels of vitamin intake. Unless the diet is adjusted to the increased requirements, deficiencies may develop. Lastly, artificial manipulation of the body and its natural metabolic pathways, as by certain surgical procedures or the administration of various drugs, can lead to avitaminoses. (Diseases involving deficiencies of particular vitamins are discussed in NUTRITIONAL DISEASES AND DISORDERS.)

DISEASES OF PSYCHOGENIC ORIGIN

Great numbers of persons are afflicted at some time with diseases of psychogenic origin. It is estimated, for example, that approximately one of every two hospital beds in the United States is occupied by someone who is suffering from mental illness. The large sums of money spent on liquor, tranquillizing drugs, and other anxiety-reducing remedies are further evidence of the widespread prevalence of psychological stress.

One of the major difficulties in assessing mental illness is the establishment of criteria for normal behaviour. Indeed, there are many who feel that everyone is affected at some time with varying degrees of mental illness. These aberrations often appear as minor idiosyncrasies and, as such, rarely pose serious threats to the individual. It is only when such "abnormal" behaviour interferes directly with the ability to adapt and interferes with one's feelings of well-being that it is properly termed mental illness.

Definition of mental illness

Mental disease may be organic—may arise from a physical abnormality—or may be nonorganic. The latter type of disorder is often called functional. For some of the severe psychiatric disorders, such as schizophrenia, an organic basis is postulated, but thus far an exact cause has not been established.

Organic psychogenic disease. There is strong evidence that heredity plays an important role in the genesis of mental illness. This is particularly true in the case of schizophrenia, a severe form of mental illness characterized by withdrawal from reality, disturbances in thought processes, and emotional distortion. There are other genetic diseases that demonstrate severe disturbances in the mental processes. Among these are Down's syndrome (mongolism), Huntington's chorea (a disorder characterized by spasmodic movements of the face, arms, and legs, and by gradual loss of the mental faculties), and phenylketonuria (PKU). The latter occurs in infants who lack a specific enzyme needed to break down phenylalanine, one of the amino acids found in proteins. This disorder is believed to be transmitted by a recessive autosomal gene (*i.e.*, it is not sex-linked and must be inherited from both parents to be fully manifest) and is associated with severe mental retardation.

Infections, particularly syphilis, are known to be the cause of certain brain disorders. Late stage, or tertiary, syphilis is known to lead slowly to destruction of brain tissue and to be responsible for bizarre psychological manifestations in affected individuals.

Alcohol ingestion over a number of years can cause significant organic brain damage.

Drugs and chemicals are a common source of psychiatric disturbances. Lead poisoning has already been discussed. Steroid hormones, used in the treatment of a variety of disorders, are known to be responsible for unusual psychic changes in a small percentage of users. Lysergic acid diethylamide (LSD) has been reported to precipitate psychiatric distress in predisposed users of the drug.

Dietary factors have been clearly established as involved causally in the development of psychiatric illness. There is a clear association between mental illness and deficiency of pantothenic acid and vitamin B_{12}. Furthermore, children raised under starvation conditions are believed to have impaired mental development.

Finally, any serious illness carries with it profound psychological stress, which may be accompanied by symptoms such as denial or depression. The patient's emotional reaction to his illness may make objective appraisal of physical impairment difficult.

Nonorganic psychogenic diseases. Just as man the biological organism must defend himself against such physical hazards as bacteria, viruses, and trauma, so man the psychological organism must defend himself against the equally hostile but somewhat different psychic hazards. Here he must deal with the loss of real and fantasied objects, maintain control over biological urges, endure frustration, tolerate injury and disease, and meet demands as a working, interacting member of society. Individuals faced with these stresses may exhibit symptoms of psychiatric illness.

Symptoms of psychiatric illness

These symptoms include any of a perplexing variety of manifestations. There may be disturbances of thought: compulsions, obsessions, inhibition of thinking, phobias, delusions, or dissociated thinking. Disorders of perception include illusions and hallucinations. Disturbances in human relationships and distortions of self-concept are commonly observed in many psychiatric disorders. Feelings are also often disturbed, anxiety and depression being the most common symptoms. Finally, the most frequent symptoms of psychiatric illness are of a physical nature, pain being the most common. Such symptoms can include virtually every system and may mimic any organic disease. Insomnia, loss of appetite, diarrhea, constipation, nausea, difficulties in swallowing, urinary problems (such as bed-wetting), fainting, seizures, palpitations, weakness, and memory loss are common. Blindness, deafness, mutism, paralysis, and false pregnancy are somewhat more bizarre. Such physical manifestations comprise the group referred to as psychosomatic diseases.

Classification of mental illness. Psychological disturbances are usually classified as neuroses, psychoses, and character disorders. There are also categorizations for a whole host of deviant sexual behaviour such as transvestism, voyeurism, sadism, and masochism. These categories are not mutually exclusive and are not distinct entities as understood in organic illness. Rather, they are designations for complex patterns of response to stress. The terms anxiety and depression describe certain affects (feelings) that may be found in any of the aforementioned categories. Anxiety, the feeling of fear in anticipation of an unwanted event, is often useful and may prepare one for a difficult task. When anxiety becomes excessive or inappropriate, however, it can severely hamper one's ability to live in harmony with one's environment. Similarly, depression can entirely disrupt normal existence. Depression can be defined as a feeling of sadness, usually related to unexpressed anger and resulting in feelings of apathy, listlessness, and inability to do work or function in a social setting.

Neurosis and psychosis in their extremes are easily differentiated, but they share many features and can be viewed as a continuum in a spectrum of mental disease. Both the neurotic and the psychotic individual struggle to maintain control of unacceptable impulses related to developmental trauma or conflicts, and yet, both are capable at times of behaviour that is completely normal. What differentiates the neurotic from the psychotic is the psychological mechanisms used to deal with these unacceptable impulses. The neurotic attempts to control his impulses by a variety of mental mechanisms and struggles to maintain interpersonal relations and to adapt to the realities of his external environment. When the neurotic is unsuccessful in keeping sexual, aggressive, or other unacceptable impulses from his consciousness, he is overwhelmed by guilt, shame, and anxiety and experiences himself as ill. The psychotic, on the other hand, distorts or changes his perception of the environment or detaches himself from it to deny or distort the forbidden impulses. He strives to make the environment one in which the forbidden impulses become acceptable to him. The psychotic is less able to maintain socially acceptable behaviour, to carry on interpersonal relationships, or to carry on effective work. In addition, he is usually un-

aware of these shortcomings, whereas the neurotic is acutely aware of deficiencies in these areas.

Character disorders

Character disorders, the third broad class of nonorganic psychiatric disease, are characterized by rigidity and inflexibility in dealing with stresses. Whereas the neurotic is capable of living a relatively normal existence and only becomes symptomatic when dealing with a particular stressful situation, the person described as having a character disorder incorporates such neurotic symptoms into his life-style in a permanent way. Thus the neurotic symptoms become indistinguishable from the personality itself. The obsessive–compulsive type, for example, demands strict order in his daily routine. Though such a person may lead a highly useful and productive life, interpersonal relationships often suffer as a result (see PSYCHOSES; PSYCHONEUROSES).

DISEASES OF SENESCENCE

The process of aging begins at the time of conception. Throughout life the body undergoes a series of changes that can be considered as manifestations of aging. During the first half of life these changes are generally referred to as maturation; during the last half of life, as progressive senescence. Visual acuity, sensitivity of hearing, and muscular vigour begin to deteriorate after the third decade of life. These changes, although they may begin at different ages and progress at differing rates, are universal among all individuals and must therefore be considered as the normal aging process.

It is extremely difficult to draw a sharp line between the deleterious effects of normal aging and the deleterious effects of diseases of aging. The diseases most commonly manifested in the elderly are disorders of the heart, blood vessels, and joints. The heart disease of the elderly is related to the generalized vascular disease known as arteriosclerosis, which frequently attacks the major coronary arteries of the heart. Arteriosclerosis and arthritis will therefore be briefly touched upon here. More extended discussions may be found in CARDIOVASCULAR SYSTEM DISEASES AND DISORDERS and in JOINT DISEASES AND INJURIES. These problems and other aspects of aging are also considered in AGING, HUMAN.

Arteriosclerosis is not a specific disease. The term is applied to all diseases that cause hardening of the arteries. Several minor processes can induce hardening of the arteries, but the overwhelming preponderance of cases of arteriosclerosis are caused by atherosclerosis. This disorder, which eventually affects all individuals to varying degrees, begins relatively early in life in most persons. There are great variations, however, in the severity of this disease among individuals and among racial, national, and ethnic populations.

Atherosclerosis

Atherosclerosis is characterized by the deposition of fats (cholesterol and other complex lipids) in the linings (intima) of the arteries. These deposits cause bulging plaques referred to as atheromas. As the disease progresses with age, the atheromas enlarge and progressively narrow the arterial channels (lumens). Atheromas have the further distressing potential of becoming solidified by the deposition of calcium salts, and in this manner the blood vessels are converted into rigid pipelike structures. The atheromas may also rupture into the lumen of the vessel and thus induce blood clotting. In this way the artery may become totally obstructed. Atherosclerosis has a predilection for the aorta, the major artery of the body, and the arteries of the heart, brain, and legs. It is atherosclerosis of the arteries of the heart (the coronaries) that is the cause of heart attacks, known technically as myocardial infarction. In most industrialized nations today heart attacks alone cause 20–25 percent of all mortality. When atherosclerosis narrows but does not totally block the coronary arteries, the heart is also injured by lack of adequate blood supply and nutrition and becomes progressively smaller and weaker; even though this disease is not as life threatening as a heart attack, it nonetheless frequently causes heart failure, an inability of the heart to deliver an adequate supply of blood to the tissues. Atherosclerosis of the arteries of the brain is the usual cause of "strokes." When the arteries to the legs become affected in this way, gangrene may develop.

Regrettably, the exact cause of atherosclerosis has not been determined. There is almost unanimous agreement, however, that the most significant causal factors in the development of this disease are obesity, high blood pressure, and high levels of blood lipids (fats). There is also a widely held suspicion that too much food and excesses of carbohydrates and animal fats (such as are derived from eggs, butter, and meats) contribute to the onset and progression of this disease. Rural peasant populations who live on fish and vegetable oils have less atherosclerosis than the city dwellers of the U.S. and Europe.

Arthritis, probably the second most common and distressing disease among the elderly, is a disease of the joints. It causes considerable pain, discomfort, and lack of mobility and so makes life burdensome. Moreover, arthritic individuals are more subject to other illnesses. Degenerative arthritis (osteoarthritis) is common to all elderly people to a lesser or greater degree.

The ends of the bones that move over each other in the joints of the body are covered by cartilage. This, being more flexible and rubbery than bone, forms a smooth surface, relatively free of friction. In degenerative joint disease these smooth surfaces gradually become rough. The resulting friction between the rough surfaces causes the cartilage to wear away. The bone surfaces then become exposed, and the joint becomes painful and less mobile. As the disease progresses bony spurs form along the ends of the bone, and these may fuse across the joint space and completely immobilize the joint. Osteoarthritis usually begins in the fourth decade of life and slowly progresses with increasing age. Coinciding with these changes in joints are changes involving the bone itself. The bone of elderly persons is known to be less dense and more brittle; it tends, therefore, to fracture more easily. It also heals with greater difficulty.

Degenerative joint disease

There are many subtle changes that occur with the normal aging process. These may include degenerative changes in the brain, leading to impaired mental ability and even senility. As this damage is usually accompanied by atherosclerosis of the arteries of the brain, it is difficult to know how much of the change is the result of impaired blood flow and how much is related to normal aging. Finally, but of no less significance, is the general decline in the body's ability to defend itself against disease. Thus elderly persons are more susceptible to infections, trauma, and a number of other bodily derangements. Simple, uncomplicated pneumonia, which might be easily tolerated by the young, healthy adult, may be fatal for an elderly, weakened person.

IV. Classifications of diseases

Value of disease classification

Classifications of diseases become extremely important in the compilation of statistics on causes of illness (morbidity) and causes of death (mortality). It is obviously important to know what kinds of illness and disease are prevalent in an area and how these prevalence rates vary with time. By classifying diseases it became apparent, for example, that the frequency of lung cancer was entering a period of alarming increase in mid-20th century. Once a rare form of cancer, it had become the single most important form of cancer in males. With this knowledge a search was instituted for possible causes of this increased prevalence. It was concluded that the occurrence of lung cancer was closely associated with cigarette smoking. Classification of disease had helped to ferret out an important, possibly causal, relationship.

The most widely used classifications of disease are (1) topographical, by bodily region or system; (2) anatomic, by organ or tissue; (3) physiological, by function or effect; (4) pathological, by the nature of the disease process; (5) etiological (causal); (6) juristic, by speed of advent of death; (7) epidemiological; and (8) statistical. Any single disease may fall within several of these classifications.

In the topographical classification, diseases are subdivided into such categories as gastrointestinal disease, vascular disease, abdominal disease, and chest disease. Various specializations within medicine follow such topo-

graphic or systemic divisions, so that there are physicians who are essentially vascular surgeons, for example, or clinicians who are specialized in gastrointestinal disease. Similarly, some physicians have become specialized in chest disease and concentrate principally on diseases of the heart and lungs.

In the anatomic classification, disease is categorized by the specific organ or tissue affected; hence, heart disease, liver disease, and lung disease. Medical specialties such as cardiology are restricted to diseases of a single organ, in this case the heart. Such a classification has its greatest use in identifying the various kinds of disease that affect a particular organ. The heart is a good example to consider. By the segregation of cardiac disease it has been made apparent that heart disease is now the most important cause of death in the United States and in most other industrialized nations. Moreover, it has become further apparent that disease caused by atherosclerosis of the coronary arteries is by far the most important form of heart disease. In making a diagnosis of cardiac disease in an elderly patient, the cardiologist must first determine whether this disease of the coronary arteries is responsible for the heart's failure to function normally.

Physiological classification of disease

The physiologic classification of disease is based on the underlying functional derangement produced by a specific disorder. Included in this classification are such designations as respiratory and metabolic disease. Respiratory diseases are those that interfere with the intake and expulsion of air and the exchange of oxygen for carbon dioxide in the lungs. Metabolic diseases are those in which disturbances of the body's chemical processes are a basic feature. Diabetes and gout are examples.

The pathological classification of disease considers the nature of the disease process. Neoplastic and inflammatory disease are examples. Neoplastic disease includes the whole range of tumours, particularly cancers, and their effect on human beings. Physicians who specialize in neoplastic disease are called oncologists.

The etiologic classification of disease is based on the cause, when known. This classification is particularly important and useful in the consideration of biotic disease. On this basis disease might be classified as staphylococcal or rickettsial or fungal, to cite only a few instances. It is important to know, for example, what kinds of disease staphylococci produce in human beings. It is well-known that they cause skin infections and pneumonia, but it is also important to note how often they cause meningitis, abscesses in the liver, and kidney infections. The venereal diseases syphilis and gonorrhea are further examples of diseases classified by etiology.

The juristic basis of the classification of disease is concerned with the legal circumstances in which death occurs. It is principally involved with sudden death, the cause of which is not clearly evident. Thus, on a juristic basis, some deaths and diseases are classified as medical–legal and fall within the jurisdiction of coroners and medical examiners. A person living alone is found dead in his bed—did he die of natural causes or was he killed? Had the person who dropped dead on the street been given some poison that took a short time to act? Much less dramatic, but perhaps more common, are disease and death caused by exposure of the individual to some unrecognized danger to health in his working or living conditions. Could the illness or disease be attributable to fumes or dusts in a factory? These are examples of the many types of disease and death that fall properly in this classification.

Epidemiological classification

The epidemiological classification of disease deals with the incidence, distribution, and control of disorders in a population. To use the example of typhoid, a disease spread through contaminated food and water, it first becomes important to establish that the disease observed is truly caused by *Salmonella typhosa*, the typhoid organism. Once the diagnosis is established it is obviously important to know the number of cases, whether the cases were scattered over the course of a year or occurred within a short period, and what the geographic distribution is. It is critically important that the precise address and activities of the patients be established. Two widely separated locations within a city might be found to have clusters of cases of typhoid all arising virtually simultaneously. It might be found that each of these clusters revolved about a family unit including cousins, nephews, and other friends, suggesting that in some way personal relationships might be important. Further investigation might disclose that all of the infected persons had dined at one time or at short intervals in a specific home. It might further be found that the person who had prepared the meal had recently visited some rural area and had suffered a mild attack of the disease and was now spreading it to family and friends by unknowing contamination of food. This hypothetical case suggests the importance of the etiologic, as well as the epidemiologic, classification of disease.

Epidemiology is one of the important sciences in the study of nutritional and biotic diseases around the world. The United Nations now supports, in part, a World Health Organization, whose chief function is the worldwide investigation of the distribution of disease. In the course of this investigation, many observations have been made that help to explain the cause and provide approaches to the control of many diseases.

Statistical classification

The statistical basis of classification of disease employs analysis of the incidence (the numbers of new cases of a specific disease that occur during a certain period) and the prevalence rate (number of cases of a disease in existence at a certain time) of diseases. If, for example, a disease has an incidence rate of 100 cases per year in a given locale, and, on the average, the affected persons live three years with the disease, it is obvious that the prevalence of the disease is 300. Statistical classification is an additional important tool in the study of possible causes of disease. These studies, as well as epidemiologic, nutritional, and pathologic analyses, have made it clear, for example, that diet is an important consideration in the possible causation of atherosclerosis. The statistical analyses drew attention to the role of high levels of fats and carbohydrates in the diet in the possible causation of atherosclerosis. The analyses further drew attention to the fact that certain populations that do not eat large quantities of animal fats and subsist largely on vegetable oils and fish have a much lower incidence of atherosclerosis. Thus statistical surveys are of great importance in the study of human disease.

The state of health, then, is jealously and effectively guarded by a constellation of mechanisms that immediately come into play to correct any deviation from normality. Diseases abound; they emanate from a host of causes. Some are as subtle as an infinitesimal change in the structure of a protein, others are as gross as the massive injuries that follow an automobile accident. But despite all of these threats to health, it must be remembered that the normal longevity in the developed countries of the world has now exceeded the biblical "three score and ten" years through man's increased understanding—and control—of disease.

BIBLIOGRAPHY. P.B. BEESON and W. MCDERMOTT (eds.), *Cecil-Loeb Textbook of Medicine*, 13th ed., pp. 24–76 (1971), a discussion of man, his environment and disease, including a new section on air and water pollution; L.S. GOODMAN and A. GILMAN (eds.), *The Pharmacological Basis of Therapeutics*, 4th ed. (1970), a comprehensive text on drugs; T.R. HARRISON *et al.* (eds.), *Principles of Internal Medicine*, 6th ed., pp. 43–202 (1970), a detailed discussion of the cardinal manifestations of disease, under such headings as fever, cough, headache, etc.; T. LIDZ, *The Person: His Development Throughout the Life Cycle* (1968), an excellent insight into man—the psychological organism; S.L. ROBBINS and M. ANGELL, *Basic Pathology* (1971), a general consideration of disease from both a clinical and morphologic approach, including an up-to-date discussion of the etiology and pathogenesis of the many diseases of man; J.S. and M.W. THOMPSON, *Genetics in Medicine* (1966), a well-illustrated and clearly written text on basic genetic principles and their relation to the genesis of human disease.

(S.L.R./J.H.Ro.)

Diseases of Animals

Man's concern with diseases that afflict animals dates from his earliest contacts with them and is reflected in

early views of religion and magic. Diseases of animals remain a concern of man principally because of the economic losses they cause and the possible transmission of the causative agents to humans. The branch of medicine called veterinary medicine deals with the study, prevention, and treatment of diseases not only in domesticated animals but also in wild animals and in animals used in scientific research. The prevention, control, and eradication of diseases of economically important animals are agricultural concerns. Programs for the control of diseases communicable from animals to man, called zoonoses, especially those in pets and in wildlife, are closely related to human health. Further, the diseases of animals are of increasing importance, for a primary public-health problem throughout the world is animal-protein deficiency in the diet of humans. Indeed, both the United Nations Food and Agricultural Organization (FAO) and the World Health Organization (WHO) have been attempting to solve the problem of protein deficits in a world whose human population is rapidly expanding.

This article is divided into the following major sections:

I. General features of animal diseases
Historical background
Importance
Role of ecology
II. Detection and diagnosis
Reactions of tissue to disease
Methods of examination
Tests as diagnostic aids
III. Survey of animal diseases
Infectious and noninfectious diseases
Zoonoses
Disease prevention, control, and eradication

I. General features of animal diseases

HISTORICAL BACKGROUND

Historical evidence, like that from currently developing nations, indicates that veterinary medicine originally developed in response to the needs of pastoral and agricultural man along with human medicine. It seems likely that a veterinary profession existed throughout a large area of Africa and Asia from at least 2000 BC. Ancient Egyptian literature includes monographs on both animal and human diseases. Evidence of the parallel development of human and veterinary medicine is found in the writings of Hippocrates on medicine and of Aristotle, who described the symptomatology and therapy of the diseases of animals, including man. Early Greek scholars, noting the similarities of medical problems among the many animal species, taught both human and veterinary medicine. In the late 4th century BC, Alexander the Great designed programs involving the study of animals, and medical writings of the Romans show that some of the most important early observations on the natural history of disease were made by men who wrote chiefly about agriculture, particularly the aspect involving domesticated animals.

Parallel development with human medicine

It is of interest that most of the earliest suggestions of relationships between human health and animal diseases were part of folklore, magic, or religious practice. The Hindu's concern for the well-being of animals, for example, originated in his belief in reincarnation. From the pre-Christian Era to about 1500, the distinctions between the practices of human and veterinary medicine were not clear-cut; this was especially true in the fields of obstetrics and orthopedics, in which animal doctors in rural areas often delivered babies and set human-bone fractures. It was realized, however, that training in one field was inadequate for practicing in the other, and the two fields were separated.

Veterinary literature from the civilizations of Greece and Rome contains reference to "herd factors" in disease; contagion within groups of animals kept together, therefore, was recognized, and both quarantine and slaughter were used to control outbreaks of livestock diseases. Rinderpest (cattle plague) was the most important livestock disease from the 5th century until control methods were developed. Serious outbreaks of the disease prompted the founding of the first veterinary college (École Nationale Vétérinaire), in Lyons, France, in 1762. Many aspects of animal diseases are best understood in terms of population or herd phenomena; for example, herds of livestock, rather than individual animals, are vaccinated against specific diseases, and housing, nutrition, and breeding practices are related to the likelihood of illness in the herd.

The work of Pasteur was of fundamental significance to general medicine and to agriculture. Veterinarians became concerned with foods of animal origin after the discovery of micro-organisms and their identification with diseases in man and other animals. Efforts were directed toward protecting humans from diseases of animal origin, primarily those transmitted through meat or dairy products. Modern principles of food hygiene, first established for the dairy and meat-packing industries in the 19th and early 20th centuries, have been generally applied to other food-related industries. The veterinary profession, especially in Europe, assumed a major role in early food-hygiene programs.

Since World War II, the eradication of animal diseases, rather than their control, has become increasingly important, and conducting basic research, combatting zoonoses, and contributing to man's food supply have become indispensable services of veterinary medicine.

IMPORTANCE

Economic importance. About 50 percent of the world's population suffers from chronic malnutrition and hunger. Inadequate diet claims many thousands of lives each day. When the lack of adequate food to meet present needs for an estimated world population of 3,672,000,000 in mid-1970 is coupled with the prediction that the population may increase to 7,000,000,000 by the year 2000, it becomes obvious that animal-food supplies must be increased. One way in which this might be accomplished is by learning to control the diseases that afflict animals throughout the world (see Table 10), especially in the developing nations of Asia and Africa, where the population is expanding most rapidly. Most of the information concerning animal diseases, however, applies to domesticated animals such as pigs, cattle, and sheep, which are relatively unimportant as food sources in these nations. Remarkably little is known of the diseases of the goat, the water buffalo, the camel, the elephant, the yak, the llama, or the alpaca; all are domesticated animals upon which the economies of many developing countries depend. It is in these countries that increased animal production resulting from the development of methods for the control and eradication of diseases affecting these animals is most urgently needed.

Despite the development of various effective methods of disease control, substantial quantities of meat and milk are lost each year throughout the world. The United States Department of Agriculture, for example, has estimated that losses from animal diseases in the United States and other developed countries in 1965 probably amounted to 10 percent of the yearly income from livestock and poultry. It has been estimated that worldwide losses in 1967 as a result of disease included more than 1,000,000,000 each of cattle and sheep; 350,000,000 goats; 100,000,000 buffalo; 11,000,000 camels; 550,000,000 pigs; 119,000,000 horses, mules, and donkeys, and close to 1,000,000,000 fowl. In countries in which animal-disease control is not yet adequately developed, the loss of animal protein from disease is about 30 to 40 percent of the quantity available in underdeveloped areas. In addition, such countries also suffer losses resulting from poor husbandry practices—*e.g.*, the average yield from cattle carcasses in the United States is about 600 pounds (270 kilograms) per animal; that from African cattle is about 325 pounds (150 kilograms).

Role in human disease. Animals have long been recognized as agents of human disease. Man has probably been bitten, stung, kicked, and gored by animals for as long as he has been on earth; in addition, early man sometimes became ill or died after eating the flesh of dead animals. In more recent times, man has discovered that many invertebrate animals are capable of transmitting causative agents of disease from man to man or

from other vertebrates to man. Such animals, which act as hosts, agents, and carriers of disease, are important in causing and perpetuating human illness. Because about three-fourths of the important known zoonoses are associated with domesticated animals, including pets, the term zoonoses was originally defined as a group of diseases that man is able to acquire from domesticated animals. But this definition has been modified to include all human diseases (whether or not they manifest themselves in all hosts as apparent diseases) that are acquired from or transmitted to any other vertebrate animal. Thus, zoonoses are naturally occurring infections and infestations (a term that indicates the presence of invertebrates in or on a host animal) shared by man and other vertebrates. Although the role of domesticated animals in many zoonoses is understood, the role of the numerous species of wild animals with which man is less intimately associated is not well understood. The discovery that diseases such as yellow fever, viral brain infections, plague, and numerous other important diseases involving man or his domesticated animals are fundamentally diseases of wildlife and exist independently of man and his civilization, however, has increased the significance of studying the nature of wildlife diseases. Table 9 contains a partial list of zoonoses, including the causative agents and the animals involved.

Role of domesticated animals

Animals in research: the biomedical model. Although in modern times the practice of veterinary medicine has been separated from that of human medicine, the observations of the physician and the veterinarian continue to add to the common body of medical knowledge. Of the more than 1,200,000 species of animals thus far identified, only a few have been utilized in research, even though it is likely that, for every known human disease, an identical or similar disease exists in at least one other animal species. Veterinary medicine plays an ever-increasing role in the health of man through the use of animals as biomedical models with similar disease counterparts in man. This use of animals as models is important because research on many genetic and chronic diseases of man cannot be carried out using humans.

Importance of animals in research

It has been estimated that as many as 28,000,000 mice and 250,000 monkeys are utilized each year in research laboratories in the U.S. alone. Animal studies are used in the development of new surgical techniques (*e.g.*, organ transplantations), in the testing of new drugs for safety, and in nutritional research. Animals are especially valuable in research involving chronic degenerative diseases because they can be induced experimentally in them with relative ease. The importance of chronic degenerative diseases, such as cancer and cardiovascular diseases, has increased in parallel with the growing number of communicable diseases that have been brought under control. See Table 1 for a list of animals with diseases similar to those that occur in man.

Examples of animal diseases that are quite similar to commonly occurring human diseases include chronic emphysema in the horse; leukemia in cats and cattle; muscular dystrophies in chickens and mice; atherosclerosis in pigs and pigeons; blood-coagulation disorders and nephritis in dogs; gastric ulcers in swine; vascular aneurysms (permanent and abnormal blood-filled area of a blood vessel) in turkeys; diabetes mellitus in Chinese hamsters; milk allergy and gallstones in rabbits; hepatitis in dogs and horses; hydrocephalus (fluid in the head) and skin allergies in many species; epilepsy (a disease of the central nervous system) in dogs and gerbils; hereditary deafness in many small animals; cataracts in the eyes of dogs and mice; and urinary stones in dogs and cattle.

The study of animals with diseases similar to those that affect man has increased knowledge of the diseases in man; knowledge of nutrition, for example, based largely on the results of animal studies, has improved the health of animals, including man. Animal investigations have been used extensively in the treatment of shock, in open-heart surgery, in organ transplantations, and in the testing of new drugs. Other important contributions to human health undoubtedly will result from new research discoveries involving the study of animal diseases.

ROLE OF ECOLOGY

Epidemiology, the study of epidemics, is sometimes defined as the medical aspect of ecology, for it is the study of diseases in animal populations. Hence the epidemiologist is concerned with the interactions of organisms and their environments as related to the presence of disease. The multiple-causality concept of disease embraced by epidemiology involves combinations of environmental factors and host factors, in addition to the determination of the specific causative agent of a given disease. Environmental factors include geographical features, climate, and concentration of certain elements in soil and water. Host factors include age, breed, sex, and the physiological state of an animal as well as the general immunity of a herd resulting from previous contact with a disease. Epidemiology, therefore, is concerned with the determination of the individual animals that are affected by a disease, the environmental circumstances under which it may occur, the causative agents, and the ways in which transmission occurs in nature. The epidemiologist, who utilizes many scientific disciplines (*e.g.*, medicine, zoology, mathematics, anthropology), attempts to determine the types of diseases that exist in a specific geographical area and to control them by modifying the environment.

Diseases in animal populations are characterized by certain features. Some outbreaks are termed sporadic diseases because they appear only occasionally in individuals within an animal population. Diseases normally present in an area are referred to as endemic, or enzootic, diseases, and they usually reflect a relatively stable relationship between the causative agent and the animals affected by it. Diseases that occasionally occur at higher than normal rates in animal populations are referred to as epidemic, or epizootic, diseases, and they generally represent an unstable relationship between the causative agent and affected animals.

Sporadic and endemic diseases

The effect of diseases on a stable ecological system, which is the result of the dominance of some plants and animals and the subordination or extinction of others, depends on the degree to which the causative agents of diseases and their hosts are part of the system. Epidemic diseases result from an ecological imbalance; endemic diseases often represent a balanced state. Ecological imbalance and, hence, epidemic disease may be either naturally caused or induced by man. A breakdown in sanitation in a city, for example, offers conditions favourable for an increase in the rodent population, with the possibility that diseases such as plague may be introduced into and spread among the human population. In this case, an epidemic would result as much from an alteration in the environment as from the presence of the causative agent *Pasteurella pestis*, since, in relatively balanced ecological systems, the causative agent exists enzootically in the rodents (*i.e.*, they serve as reservoirs for the disease) and seldom involves man. In a similar manner, an increase in the number of epidemics of viral encephalitis, a brain disease, in man has resulted from the ecological imbalance of mosquitoes and wild birds caused by man's exploitation of lowland for farming. Driven from their natural habitat of reeds and rushes, the wild birds, important natural hosts for the virus that causes the disease, are forced to feed near farms; mosquitoes transmit the virus from birds to cattle and man.

II. Detection and diagnosis

REACTIONS OF TISSUE TO DISEASE

Disease may be defined as an injurious deviation from a normal physiological state of an organism sufficient to produce overt signs, or symptoms. The deviation may be either an obvious organic change in the tissue comprising an organ or a functional disturbance whose organic changes are not obvious. The severity of the changes that occur in cells and tissues subjected to injurious agents is dependent upon both the sensitivity of the tissue concerned and the nature and time course of the agent. A mildly injurious agent that is present for short periods of time may either have little effect or stimulate cells to increased activity. Strongly injurious agents in prolonged contact with cells cause characteristic changes

Table 1: A Partial Listing of Biomedical Models in Veterinary Medicine

animal disease (model)	animal affected	human counterpart disease
Cardiovascular system diseases		
Hereditary lymphedema	dog	Milroy's disease
Elevated blood pressure	mouse	hypertension
Atherosclerosis	swine	atherosclerosis
Periarteritis nodosa	cattle	periarteritis nodosa
Dissecting aneurysms	turkey	aneurysm
High-altitude disease	cattle	right ventricular hypertrophy
Endocardial fibroelastosis	dog	endocardial fibroelastosis
Heart failure	dog	congestive heart failure
Congenital lymphatic edema	swine	lymphatic edema
Endocrine system diseases		
Diabetes mellitus	Chinese hamster	diabetes mellitus
Antidiuretic-hormone deficiency	mouse	diabetes insipidus
Polyuria	Chinese hamster	diabetes insipidus
Congenital goitre	cattle	goitre
Adrenal cortical hypertrophy	dog	hyperadrenocorticism
Snell's dwarf	mouse	thyrotropin deficiency
Adenohypophyseal aplasia	cattle	adenohypophyseal aplasia
Hyperinsulinism	dog	hyperinsulinism
Familial "adiposity"	mouse	obesity
Acetonemia	cattle	ketosis
Early senility	Syrian hamster	aging
Gastrointestinal system diseases		
Esophageal achalasia	dog	achalasia
Cleft palate	horse	cleft palate
Gastric ulcer	swine	gastric ulcer
Regional ileitis	swine	regional ileitis
Granulomatosis colitis	boxer dog	ulcerative colitis
Acute hemorrhagic colitis	rabbit	hemorrhagic colitis
Megacolon	mouse	megacolon
Pancreatitis	dog	pancreatitis
Liver diseases		
Viral hepatitis	subhuman primate	viral hepatitis
Serum hepatitis	horse	transfusion hepatitis
Dubin–Johnson syndrome	sheep	Dubin–Johnson syndrome
Congenital photosensitivity and hyperbilirubinemia	Southdown sheep	Gilbert's syndrome
Nonhemolytic hyperbilirubinemia	rat	Crigler–Najjar syndrome
Pigmentary liver disease	howler monkey	hepatocellular melanosis
Hepatorenal syndrome	dog	hepatorenal syndrome
Hepatic coma	horse	hepatic coma
Glycogen-storage syndrome	dog	von Gierke's syndrome
Lantana camara poisoning	sheep	kwashiorkor
Pyrrolizidine plant alkaloids	cattle	veno-occlusive disease
Hemopoietic system diseases		
Congenital erythrocytic porphyria (recessive)	cattle	congenital erythrocytic porphyria
Congenital porphyria (dominant)	cat	erythrocytic porphyria
Hereditary leukomelanopathy	mink	Chediak–Higashi syndrome
Pelger–Huët anomaly	cattle	Pelger–Huët anomaly
Cyclic neutropenia	dog	cyclic neutropenia
Aleutian disease	mink	multiple myeloma
Abnormal lipid in lymphoid tumours	mouse	Niemann–Pick disease
Viral leukemia	cat	lymphocytic leukemia
Multiple myeloma	dog	multiple myeloma
Bialbuminemia	swine	bialbuminemia
Hemophilia (factor VIII)	dog	hemophilia
Factor VII deficiency	dog	factor VII deficiency
Hemophilia-B-like disease	dog	Christmas disease
Hertwig's anemia	mouse	macrocytic anemia
Malaria	penguin	malaria
In vitro sickling of erythrocytes	deer	sickle-cell anemia
Muscle diseases		
Hereditary muscular dystrophy	chicken	muscular dystrophy
Muscle diseases (cont.)		
Polymyopathy	Syrian hamster	muscular dystrophy
Muscular dysgenesis	mouse	prenatal muscle degeneration
Nutritional muscular dystrophy	sheep	muscular dystrophy
Paralytic myoglobinuria	horse	paroxysmal myoglobinuria
Myoclonia congenita	swine	myotonia congenita
Nervous system diseases		
Cerebellar hypoplasia	cat	cerebellar hypoplasia
Nigropallidal encephalomalacia	horse	Parkinson's disease
Hydrocephalus	rabbit	hydrocephalus
Leukoencephalosis	mouse	dystrophy of white matter
Globoid leukodystrophy	dog	globoid leukodystrophy
Grand-mal seizures	gerbil	epilepsy
Lipodystrophy	dog	familial amaurotic idiocy
Scotty cramps	dog	neurogenic muscular cramps
Milk fever	cattle	hypocalcemia
Trembler mutation	mouse	tremours
Hereditary ataxia	calf	ataxia
Congenital myotonia	goat	myotonia
Eye and ear diseases		
Hereditary deafness	cat	deafness
Cochlear degeneration	mouse	cochlear degeneration
Hypoplasia of organ of Corti	dog	hypoplasia of organ of Corti
Hereditary glaucoma	rabbit	glaucoma
Inherited cataract	cattle	cataract
Hereditary iridal heterochromia	cattle	iridal heterochromia
Congenital retinal dysplasia	dog	retinal dysplasia
Retinal dystrophy	mouse	pigmented retina
Diabetic microaneurysms	dog	diabetic microaneurysms
Reproductive system diseases		
Toxemia of pregnancy	guinea pig	toxemia of pregnancy
Prolonged gestation	cattle	prolonged gestation
Uterine cystic hyperplasia	mouse	uterine cystic hyperplasia
Prostatic hyperplasia	canine	prostatitis
Cryptorchidism	swine	cryptorchidism
Respiratory system diseases		
Acute pulmonary emphysema	cattle	pulmonary emphysema
Chronic pulmonary emphysema	horse	pulmonary emphysema
Pulmonary adenomatosis	cattle	adenomatosis
Pneumonia	dog	Hecht's pneumonia
Induced lung tumours	mouse	lung tumours
Skeletal system diseases		
Osteodystrophy	primate	fibrous osteodystrophy
Familial osteoporosis	dog	osteogenesis imperfecti
Senile osteoporosis	mouse	senile osteoporosis
Achondroplasia	rabbit	dwarfism
Intervertebral-disk syndrome	dog	disk luxation
Hip dysplasia	dog	acetabular dysplasia
Clubfoot	mouse	clubfoot
Skin diseases		
Baldness, male pattern	stumptail macaque	baldness, male pattern
Albinism	mouse	albinism
Genetic hypotrichosis	cattle	hypotrichosis
Hyperkeratosis	cattle	hyperkeratosis
Cutis hyperelastica	dog	Ehlers–Danlos disease
Seborrheic dermatitis	dog	seborrheic dermatitis
Impetigo	dog	impetigo
Milia	dog	milia
Urinary system diseases		
Diabetes insipidus	mouse	diabetes insipidus
Cystinuria	blotched genet	cystinuria
Chronic interstitial nephritis	dog	uremia
Cystic or absent kidneys	rat	cystic kidneys
Renal amyloidosis	mouse	renal amyloidosis
Cloisonné kidneys	goat	renal hemosiderosis

in them by interfering with normal cell processes. Most causative agents of disease fall into the latter category. Causative agents and some of the symptoms of many of the diseases mentioned in this section are found in Tables 2 through 7.

Characteristics of cell and tissue changes. Changes in cells and tissues as a result of disease include degenerative and infiltrative changes. Degenerative changes are characterized by the deterioration of cells or a tissue from a higher to a lower form, especially to a less functionally active form. When chemical changes occur in the tissue, the process is one of degeneration. When the changes involve the accumulation of materials within the cells comprising tissues, the process is called infiltration. Diseases such as pneumonia, metal poisoning, or septicemia (the persistence of disease-causing bacteria in the bloodstream) may cause the mildest type of degeneration—parenchymatous changes, or cloudy swelling

Degeneration and infiltration

Table 2: Selected Infectious and Parasitic Diseases of Animals

animal(s) affected	name(s) of disease	causative organism	nature of disease
Diseases of bacterial origin			
Most mammals, chickens	necrobacillosis, calf diptheria, bovine foot rot, necrotic hepatitis, dermatitis	*Sphaerophorus necrophorus*	organism invades tissue and causes tissue death (necrosis) after other wounds or infections have occurred; *i.e.*, disease is known as a secondary infection
Cattle, sheep, horses, chickens, man, many other animals	botulism	toxins produced by *Clostridium botulinum*	results from eating toxins released in decayed or spoiled foods; toxins cause rapid paralysis of nerves in throat and all muscles, almost always fatal
Swine, cattle, sheep, goats, rabbits, man, dogs, many other animals	listeriosis, circling disease, meningoencephalitis	*Listeria monocytogenes*	symptoms vary in affected animal; organism may affect the central nervous system (brain, spinal cord) or the membranes surrounding it or cause necrosis of heart muscles, localized tissue death in liver, or a septicemia (persistence of bacteria in the bloodstream)
Swine, cattle, sheep, goats, horses, turkeys, man	erysipelas, diamond-skin disease	*Erysipelothrix insidiosa*	manifestations include septicemia and pathological discontinuities of tissue (lesions) in skin, heart, joints (in swine); arthritis (in sheep, occasionally in cattle, horses, goats); septicemia and death (in turkeys); skin lesions (in man)
Cattle, sheep, goats, horses, mules	anthrax, splenic fever, charbon	*Bacillus anthracis*	spores (inactive forms) of organisms in soil, transmitted through insect bites or food; manifestations include hemorrhage and edema (accumulation of fluid) in tissues
Swine, cattle, sheep, horses, mules	malignant edema, gas gangrene	*Clostridium septicum*	spores enter from dirt into injured tissue, cause severe gangrene (rotting of dead tissue), swelling; prognosis (outlook) poor
Swine, cattle, sheep, goats	pyobacillosis	*Corynebacterium pyogenes*	characterized by multiple abscesses (localized collections of pus) throughout the body; may result in debilitation (including arthritis in swine) and death
Primarily swine, cattle, goats (secondarily in man and other animals)	brucellosis, Bang's disease, contagious abortion; undulant fever, Malta fever (in man)	*Brucella abortus, B. melitensis, B. suis.*	primarily affects genital organs in both sexes; may cause abortion, sterility, infection of fetus in female, local lesions in various tissues; pasteurization of milk has controlled the disease in man
Swine, cattle, sheep	shipping fever, pasteurellosis, hemorrhagic septicemia	*Pasteurella multocida* and *P. hemolytica;* also in conjunction with viral agents	causes of great economic losses throughout the world; manifestations may include acute to chronic respiratory disease; the various causative organisms vary in virulence (degree of pathogenicity); an acute form (*i.e.*, short, severe) affects rabbits
Horses, mules, donkeys (man less susceptible)	glanders, farcy, malleus	*Malleomyces mallei*	organisms enter animal through digestive tract, travel via blood to lungs, trachea, and skin, and form ulcers; an acute form causes death, a chronic type may persist for years; human infections occur from exposure of broken skin to affected animals
Horses (mules, donkeys less susceptible)	strangles, distemper, infectious adenitis	*Streptococcus equi*	most common in young, undernourished horses in crowded conditions; manifested by high temperature, nose infections, and abscesses in lymph glands of neck
Horses	purpura hemorrhagica, petechial fever	unknown, but associated with *Streptococcus equi*	noncontagious, follows acute infections and toxemias; characterized by generalized hemorrhages in tissues and the accumulation of fluid (edema); relapses often occur
Primarily horses	ulcerative lymphangitis or cellulitis	*Corynebacterium pseudotuberculosis*	chronic disease, develops slowly following the entrance of bacteria through skin; affects hindlegs, sometimes severely
Horses (males most susceptible); sometimes swine, cattle; rarely dogs or cats	tetanus, lockjaw	toxins produced by *Clostridium tetani*	bacteria enter tissue at time of injury, produce toxins in necrotic tissue; affected animals become stiff; death results from suffocation
Swine	streptococcal infection	*Streptococcus* species	younger pigs more easily infected; symptoms varied (*e.g.*, septicemia, arthritis, uterine inflammation, middle-ear infection, multiple abscesses)
Swine, accidentally in other animals	salmonellosis, enteritis, swine typhoid	*Salmonella choleraesuis*	may be acute (*i.e.*, have a short, severe course) or chronic (persists for a long time); symptoms include loss of weight, sometimes acute septicemia
Swine, cattle	actinobacillosis, botryomycosis, big head	*Actinobacillus lignieresii*	organism a normal inhabitant of mouth, enters tissues through ulcers or wounds; symptoms include abscesses
Cattle	leptospirosis, hemoglobinuria	*Leptospira* species	symptoms of acute form include abortion, bloody milk, hemoglobin (blood pigment) in urine, kidney disease, and destruction of red blood cells; a milder form also exists
Cattle	infectious bovine pyelonephritis, infectious cystitis	*Corynebacterium renale*	usually observed in pregnant cattle in winter; a slowly developing disease that affects kidneys and bladder
Primarily cattle (rarely man, swine, sheep, horses)	tuberculosis, pearly disease	*Mycobacterium tuberculosis*	a chronic disease characterized by lesions, usually in lungs and lymph nodes, but sometimes in many other organs
Cattle (rarely sheep)	bacillary hemoglobinuria, red water disease	*Clostridium hemolyticum*	spores eaten with food, develop into active cells, migrate to liver, and produce infarcts (tissue death); usually fatal within 36 hours

Table 2: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Cattle, sheep	pinkeye, infectious keratitis, kerato-conjunctivitis	*Moraxella bovis* (in cattle), *Colesiota conjunctivae* (in sheep)	affects eyes; may result in blindness; a very contagious disease whose spread may be influenced by dust irritation or possibly by viral invasion
Cattle, sheep	Johne's disease, paratuberculosis	*Mycobacterium paratuberculosis*	a chronic disease; causes diarrhea, progressive weight loss
Cattle, sheep	vibriosis, epizootic abortion	*Vibrio fetus*	a venereal disease (*i.e.*, transmitted by sexual contact) in cattle; transmitted in contaminated food and water in sheep; commonly results in infertility or abortion in cattle, abortion in sheep
Cattle, sheep (occasionally swine, goats, deer, horses)	blackleg, black quarter, quarter ill	*Clostridium feseri* (*chauvoei*)	spores transmitted from soil to animal through wounds or cuts; symptoms include lameness, gangrene of affected tissues (usually in leg muscles); usually fatal
Sheep	enterotoxemia, overeating disease, pulpy kidney	*Clostridium perfringens* type D	affected lambs usually fat or feeding on rich clover pasture; usually fatal within a day from acute toxemia (absorption of bacterial toxins)
Primarily sheep	pseudotuberculosis, caseous lymphadenitis	*Corynebacterium ovis*	organisms transmitted to animal through breaks in the skin; slowly developing abscesses (usually in lungs or lymph nodes) may rupture and spread throughout body
Sheep, goats (occasionally cattle)	black disease, infectious necrotic hepatitis	*Clostridium novyi*	organisms probably present normally in intestinal tract, associated with fluke (parasitic worm) movements in liver; produce liver damage, toxemia, and death
Diseases of viral origin			
Mammals	rabies, hydrophobia, lyssa, mad dog, le Rage	rabies virus	transmitted primarily through the bite of a rabid animal; wild animals (*e.g.*, skunks, squirrels, bats) a reservoir for infection; disease characterized by central-nervous-system symptoms (*e.g.*, rage, excitability; paralysis of jaw with salivation), general paralysis, and death
Mammals (especially cattle)	bovine warts, papillomatosis	papilloma viruses	warts of variable size develop, usually on sides of head and neck of cattle, sometimes on sex organs
Many mammals (*e.g.*, swine, cattle, sheep, goats, horses)	pox, variola	pox virus	often an acute highly infectious disease, characterized by formation of papules (small solid elevations), vesicles (small liquid-containing sacs), and pustules (small pus-filled elevations) on the skin
Swine, cattle (also rats, dogs, cats)	pseudorabies, Aujeszky's disease, mad itch	pseudorabies virus	affected animals rub body parts, undergo spasmodic muscle contractions; froth at the mouth; and show nervous irritability; usually fatal
Young pigs	transmissible gastroenteritis (TGE)	TGE virus	acute and fatal to pigs less than two to three weeks old; virus attacks absorptive surfaces of small intestine
Swine	hog cholera, swine fever, swine pest	hog-cholera virus	infectious disease; may be acute or chronic; spread by flies, animal contact, garbage, contaminated pastures; symptoms include high fever, severe hemorrhages in skin and organs
Swine	vesicular exanthema (VE)	VE virus	vesicles form on snout, mouth, abdominal wall; foot lesions occur; highly infectious, spread through animal contact or raw pork scraps in garbage; fatal usually only in young pigs
Swine, ferrets, mice	swine influenza, hog flu	swine-influenza virus; bacterium; *Hemophilus suis*	acute contagious disease; virus enters animal through lungworm larvae, acute infection occurs if *Hemophilus* organisms are in lung; symptoms include fever, pneumonia, bronchitis
Swine, cattle, horses	vesicular stomatitis (VS), mouth thrush	VS virus	effects varied; *e.g.*, high fever, salivation, vesicles in mouth region, lack of appetite (in horses); inflammation of mammary glands, vesicles in mouth region, inflammation of feet (in cattle); snout and mouth lesions, severe feet involvement (in swine)
Cattle	sporadic bovine encephalomyelitis (SBE), Buss disease	SBE virus	weakens calves; an infection of brain and membranes of brain and spinal cord; recovery often occurs
Cattle	infectious bovine rhinotracheitis (IBR), red nose, pinkeye, dust pneumonia	IBR virus	acute infection followed by secondary bacterial infections (*Pasteurella multocida*, *Spherophorus necrophorus*) in lungs, sex organs, eye; nostrils swell, become red
Cattle, buffalo, deer	malignant catarrhal fever (MCF), head catarrh, snotsiekte, epitheliosis	MCF virus	numerous symptoms include rapid weight loss, eye lesions, nasal discharge, muscular twitching, convulsions; usually fatal
Cattle, sheep	bluetongue, soremuzzle, catarrhal fever	bluetongue virus	virus transmitted through gnats (small flies); a serious problem in sheep; numerous symptoms; recovery very slow; usually less than 10 percent mortality of a flock
Sheep (also transmissible to cows)	ovine virus abortion (OVA), enzootic abortion	OVA virus	causes economic losses through abortion, weak lambs, and poor breeding efficiency

Table 2: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Sheep, goats, man	contagious ecthyma (CE), sore mouth, doby mouth, orf, pustular dermatitis	CE virus	udder, lips, and nose of sheep affected; secondary bacterial invasion may result in death, but animals usually recover
Horses	equine infectious anemia (EIA), swamp fever, malarial fever	EIA virus	transmitted by mosquitoes, lice, flies, and hypodermic needles; either acute, chronic, or latent (not manifest); varied symptoms include intermittent fever, loss of weight, jaundice, hemorrhages, anemia, and fluid in body cavities
Horses	equine viral arteritis (EVA), infectious arteritis	EVA virus	an acute contagious disease similar to EVR in symptomatology but causes damage to small arteries
Horses, mules, man, laboratory animals	equine encephalomyelitis (EE), sleeping sickness, viral encephalitis	EE virus	many viral strains transmitted by an insect (usually mosquitoes or mites or ticks); disease causes inflammation of brain cells; the many symptoms include death; inapparent infections occur in chickens, pigeons, and pheasants
Horses (also guinea pigs, mice, hamsters)	equine viral rhinopneumonitis (EVR), equine virus abortion	EVR virus	disease highly contagious; symptoms include high fever, mild upper-respiratory involvement, and usually abortion with liver damage of fetus in pregnant mares
Diseases of fungal origin			
Many domestic, laboratory, and wild mammals	histoplasmosis, reticuloendothelial cytomycosis	*Histoplasma capsulatum*	chronic; may resemble tuberculosis; granulomas (tumours) in lungs, liver, and spleen; intestinal involvement in dogs results in diarrhea
Swine, cattle, sheep, horses, fowl, dogs, cats	ringworm, tinea, trichophytosis	*Trichophyton* and *Microsporum* species	infectious skin disease caused by invasion of hair follicles; characterized by round crusty lesions, inflammation
Swine, cattle, horses (man secondarily)	actinomycosis, lumpy jaw, wooden tongue	*Actinomyces bovis*	cattle manifest a bonelike swelling on upper or lower jaw; swine manifest a tumourlike enlargement of udder caused by infections from teeth of suckling pigs
Cattle, horses, cats, dogs, man	cryptococcosis	*Cryptococcus neoformans*	usually caused by inhalation of contaminated dust; lungs affected primarily; disease may spread to almost any organ
Cattle, sheep, dogs, man	coccidioidomycosis coccidioidal granuloma, San Joaquin Valley fever	*Coccidioides immitis*	in the acute respiratory form, symptoms include cough, with recovery in two weeks; in the more serious chronic form, gradual loss in weight, abscesses, and granulomas in various tissues, including skin, occur
Primarily horses and man	sporotrichosis	*Sporotrichum schenckii*	occurs first as ulcers on skin, invasion of the lymph glands occurs, may spread throughout circulatory system
Diseases of rickettsial origin			
Swine	eperythrozoonosis, ictero-anemia, yellow-belly	*Eperythrozoon suis*	organisms cause red-blood-cell destruction; both acute and mild forms; many animals with a mild form act as carriers
Cattle	anaplasmosis, South African gall sickness	*Anaplasma marginale*	infectious disease spread either by blood-sucking ticks, mosquitoes, flies, or by mechanical transmission (*e.g.*, resulting from dehorning, vaccination); symptoms usually include extreme anemia from destruction of red blood cells; recovered animals are immunological carriers
Diseases of protozoal origin			
Most animals (including man)	toxoplasmosis	*Toxoplasma gondii*	transmission not clear but probably occurs by contaminated food or direct contact; symptoms include weakness, respiratory problems, lack of coordination, nodules throughout body, enlarged lymph nodes, and tissue death; treatment difficult
Swine, cattle, sheep, goats	coccidiosis	*Eimeria zürnii, E. bovis*, and ten other species	causative organisms found in most mature animals; symptoms result in loss of large amounts of blood and dehydration; mortality may be as high as 50 percent
Cattle	trichomoniasis	*Trichomonas fetus*	transmitted by sexual contact or by artificial insemination, symptoms include abortion, failure to conceive, inflammation of uterus; no symptoms apparent in infected bull that acts as a carrier
Cattle	Texas fever, cattle-tick fever, babesiasis, piroplasmosis, red water	*Babesia bigemina*	organism, which destroys red blood cells, is transmitted by ticks (*Margaropus* species) and mechanical means (*i.e.*, surgical instruments, needles); symptoms include high fever, severe anemia, hemoglobin in urine; chronic forms occur; some animals act as carriers
Horses	dourine, equine syphilis, breeding disease	*Trypanosoma equiperdum*	transmitted by sexual contact or blood-sucking flies; affects sex organs, causes plaquelike areas on skin, paralysis of muscles, loss of condition
Horses, mules, donkeys	equine piroplasmosis, equine malaria, babesiasis	*Babesia caballi, B. equi*	acute cases may die quickly; animals with less acute forms have varied symptoms (*e.g.*, intermittent fever, jaundice, internal hemorrhages); anemia results from invasion of red blood cells by causative organisms; *B. equi* more pathogenic than *B. caballi*

Table 2: Selected Infectious and Parasitic Diseases of Animals (continued)

animal(s) affected	name(s) of disease	causative organism	nature of disease
Diseases of nematode (roundworm) origin			
Swine	lungworms	*Metastrongylus* species	common symptoms include coughing and lung irritation
Swine	kidney worms	*Stephanurus dentatus*	mature worms live in urinary tract; larvae migrate to liver, produce lesions and weight loss
Swine	intestinal roundworm	*Ascaris suum*	migrations of organisms through lungs cause hemorrhages and pneumonia, may interfere with bile flow and food absorption
Swine	intestinal threadworm	*Strongyloides*	migration of large numbers of larvae cause tissue damage
Cattle	stomach worms	*Ostertagia* and *Trichostrongylus* species	symptoms include anemia, stunted growth, and diarrhea
Cattle	nodular worm	*Oesophagostomum radiatum*	nodules form in tissues; poor intestinal absorption caused by larvae (immature forms of organism) and nodules results in diarrhea
Cattle	verminous pneumonia	*Dictyocaulus viviparus*	ingested infective larvae migrate to lungs, produce coughing, discomfort, and pneumonia
Sheep	lungworms	*Dictyocaulus filaria*, *Muellerius capillaris*	symptoms include formation of lung nodules, collapse of portions of lungs
Sheep	hookworms	*Bunostomum trigonocephalum*	bloodsucking hookworms cause anemia, intermittent diarrhea
Sheep	filarial dermatitis	*Elaeophora schneideri*	symptoms include skin lesions
Horses	oxyuriasis (pinworms)	*Oxyuris equi*	worms cause irritation in the area around the anus
Horses	ascariasis (intestinal roundworms)	*Parascaris equorum*	larval forms cause damage; symptoms include defective intestinal absorption
Horses	strongylosis	*Strongylus* species	large numbers of worms weaken foals (newborn horses); migrating larvae may cause formation of clots in blood vessels and lameness
Horses	habronemiasis (summer sores)	*Habronema* species	habronema larvae may enter skin wounds, causing granulation; eye and stomach inflammations also occur
Diseases of platyhelminth (flatworm) origin			
Sheep	tapeworm	*Moniezia expansa*	results in poor growth
Sheep	fringed tapeworm	*Thysanosoma actinoides*	causes digestive disturbances
Horses	tapeworm	*Anaplocephala perfoliata*	symptoms may include inflammation of gut and ulceration
Diseases of acanthocephalan (spiny-headed-worm) origin			
Swine	spiny-headed worm	*Macracanthorhynchus hirudinaceus*	nodules form on small intestine; may result in peritonitis (inflammation of lining of internal organs)
Diseases of arthropod origin			
Cattle	scabies (mange)	*Chorioptes bovis*, *Psoroptes equi*, *Sarcoptes scabiei* (mites)	contagious skin diseases
Cattle	grubs	*Hypoderma bovis* (heel fly)	migrating larvae produce tissue damage, cysts, and hide damage
Cattle	lice	*Linognathus vituli*, *Solenopotes capillatus*, *Haematopinus quadripertusis* (bloodsucking lice), *Bovicola bovis* (biting louse)	symptoms include dermatitis and anemia
Sheep	screwworm infestations	many fly larvae (*e.g.*, *Cochliomyia hominivorax*, *Chrysomyia bezziana*)	flies lay eggs in open wounds; developing larvae (screwworms) burrow into tissue and destroy it
Sheep	psoroptic mange (sheep scab)	*Psoroptes communis* (mite)	all parts of skin inflamed, particularly those covered with wool

of the cells; the cells first affected are the specialized cells of the liver and the kidney. Serious cellular damage may cause the uptake of water by cells (hydropic degeneration), which lose their structural features as they fill with water. The causes for the accumulation in cells of abnormal amounts of fats (fatty infiltration and degeneration) have not yet been established with certainty but probably involve fat metabolism. Poisons such as phosphorus may cause sudden increases in the accumulation of fats in the liver. An abnormal protein material may accumulate in connective-tissue components of small arteries as a result of chronic pneumonia, chronic bacterial infections, and prolonged antitoxin production (in horses); the condition is known as amyloid degeneration and infiltration. Hyaline degeneration, characterized by tissues that become clear and appear glasslike, usually occurs in connective-tissue components of small blood vessels as a result of conditions that may occur in kidney structures (glomeruli) of animals with nephritis or in lymph glands of animals with tuberculosis. Certain structures (glomeruli) of animals with nephritis or in degeneration.

The condition in which mucus, a secretion of mucous membranes lining the inside surfaces of organs, is produced in excess and accumulates in greater than normal amounts is referred to as mucoid degeneration. Major causes include chronic irritation of mucous membranes and certain mucus-producing tumours. Abnormal amounts of glycogen, the principal storage carbohydrate of animals, may occur in the liver as a result of certain inherited diseases of animals; the condition is known as glycogen infiltration. The abnormal deposition of calcium salts, called hypercalcification, may occur as a result of several diseases involving the blood vessels and the heart, the urinary system, the gallbladder, and the bonelike tissue called cartilage. Pigments (coloured molecules) from coal dust or asbestos dust may infiltrate the lungs of certain dogs in two types of lung disease: anthracosis and asbestosis. Abnormal amounts of iron-containing coloured molecules (hemosiderin) resulting from the

breakdown of hemoglobin, the gas-carrying protein of red blood cells, are often deposited in the liver and the spleen after diseases that involve excessive breakdown of red blood cells. A black molecule (melanin) occurs abnormally in the livers of certain sheep suffering from Dubin–Johnson syndrome and in certain tumours called melanomas. Uric acid infiltration, which occurs in poultry, is characterized by the deposition of uric acid salts.

Necrosis, the death of cells or tissues, takes place if the blood supply to tissues is restricted; poisons produced by microbes, chemical poisons, and extreme heat or electricity also may cause necrosis. The rotting of the dead tissue is known as gangrene.

Atrophy of animal tissue involves a process of tissue wasting, in which a decrease occurs in the size or number of functional cells; *e.g.*, in inherited muscular dystrophy of chickens. Hypertrophy—an increase in the size of the cells in a tissue or an organ— occurs in heart muscle during diseases involving the heart valves, in certain pneumonias, and in some diseases of the endocrine glands. Aplasia is the term used when an entire organ is missing from an animal; hypoplasia indicates arrested or incomplete development of an organ, and hyperplasia an increase in the production of the number of cells; *e.g.*, the persistent callus that forms on the elbows of some dogs. Metaplasia is used to describe the change of one cell type into another; it may occur in chronic irritation of tissues and in certain cancerous tumours.

Characteristics of inflammatory reactions. When tissues are injured, they become inflamed. The inflammation may be acute, in which case the inflammatory processes are active, or chronic, in which case the processes occur slowly and new connective tissue is formed. The reaction of inflamed tissues is a combination of defensive and repair mechanisms. Acute inflammation is characterized by redness, heat, swelling, sensitivity, and impaired function. Several types of acute inflammation are known. Mild acute inflammations of mucous membranes resulting in the production of thin watery material (exudate) are called catarrhal inflammations; parenchymatous inflammations occur in organs undergoing degeneration. If the exudate formed in response to an injury is of a serious nature—that is, resembling blood plasma—the process is called serous inflammation. In fibrinous inflammation, a protein (fibrin) forms on membranes, including those in the lungs. In suppurative inflammation, dead tissue is replaced with pus composed of colourless blood cells (leucocytes) and tissue juices.

During the inflammatory reaction, the injured tissue is surrounded by an area of rapidly dividing cells. Specialized cells called macrophages enter the tissue and remove blood and tissue debris. Other cells, called neutrophils, ingest disease-causing bacteria and other foreign material. In chronic inflammations, the connective tissue contains fibroblasts, cells that divide and form new connective, or scar, tissue (see INFLAMMATION).

Hemorrhage and edema

Characteristics of circulatory disturbances. An increase in the rate of blood flow to a body part, referred to as congestion, or hyperemia, occurs during inflammation; a diminished blood flow to tissues is referred to as ischemia, or a local anemia. Examples of hemorrhage, the escape of blood from vessels, include epistaxis, or nosebleeds, in racehorses; hematemesis, or regurgitation of blood, in dogs with uremia; hemoptysis, or blood loss from lungs; hematuria, or blood in urine, of cattle with inflammation of the urinary bladder. Edema, a condition characterized by abnormal accumulations of fluid in tissues, occurs not only in a tissue during inflammation but over the entire body if the concentration of blood-serum proteins, especially albumin, is low. A thrombosis, or a blood clot in a blood vessel, may block or slow circulation of blood to tissues; if blood vessels become blocked, the condition is known as an embolism. The term infarction describes the necrosis that occurs in tissues whose blood supply is blocked by an embolism.

METHODS OF EXAMINATION

Before an unhealthy animal is treated, an attempt is made to diagnose the disease. Both clinical findings, which include symptoms obvious to a nonspecialist and clinical signs appreciated only by a veterinarian, and laboratory tests may be necessary to establish the cause of a disease. A clinical examination should indicate if the animal is in good physical condition, is eating adequately, is bright and alert, and is functioning in an apparently normal way. Many disease processes are either inflammatory or result from tumours. Malignant tumours (*e.g.*, melanomas in horses, squamous cell carcinomas in small animals) rapidly spread and usually cause death. Other diseases cause the circulatory disturbances or the degenerative and infiltrative changes summarized in the preceding section. If a specific diagnosis is not possible, the symptoms of the animal are treated.

A case record of the information pertaining to an animal (or to a herd) suspected of having a disease is begun at the time the animal is taken to a veterinarian and is continued through treatment. It includes a description of the animal (age, species, sex, breed); the owner's report; the animal's history; a description of the preliminary examination; clinical findings resulting from an examination of body systems; results of specific laboratory tests; diagnosis regarding a specific cause for the disease (etiology); outlook (prognosis); treatment; case progress; termination; autopsy, if performed; and utilization of scientific references, if applicable.

The veterinarian must diagnose a disease on the basis of various examinations and tests, since he obviously cannot interrogate the animal. Methods used include inspection—a visual examination of the animal; palpation—the application of firm pressure with the fingers to tissues to determine characteristics such as abnormal shapes and possible tumours, the presence of pain, and tissue consistency; percussion—the application of a short, sharp blow to a tissue to provoke an audible response from body parts directly beneath; auscultation—the act of listening to sounds produced by the body during the performance of functions (*e.g.*, breathing, intestinal movements); smells—the recognition of characteristic odours associated with certain diseases; miscellaneous diagnostic procedures, such as eye examinations, the collection of urine, and heart, esophageal, and stomach studies.

General inspection. Deviation of various characteristics from the normal, observation of which constitutes the general inspection of an animal, is a useful aid in diagnosing disease. The general inspection includes examination of appearance; behaviour; body condition; respiratory movements; state of skin, coat, and abdomen; and actions.

The animal's appearance

The appearance of an animal may be of diagnostic significance; small size in a pig may result from retardation of growth, which is caused by hog-cholera virus (Table 2). Observation of the behaviour of an animal is of value in diagnosing neurological diseases; *e.g.*, muscle spasms occur in lockjaw (tetanus) in dogs, nervousness and convulsions in dogs with distemper (Table 4), dullness in horses with equine viral encephalitis (Table 2), and excitement in animals suffering from lead poisoning. Subtle behavioral changes may not be noticeable. The general condition of the body is of value in diagnosing diseases that cause excessive leanness (emaciation), including certain cancers, or other chronic diseases, such as a deficiency in the output of the adrenal glands or tuberculosis. Defective teeth also may point to malnutrition and result in emaciation.

The respiratory movements of an animal are important diagnostic criteria; breathing is rapid in young animals, in small animals, and in animals whose body temperature is higher than normal. Specific respiratory movements are characteristic of certain diseases—*e.g.*, certain movements in horses with heaves (emphysema) or the abdominal breathing of animals suffering from painful lung diseases. The appearance of the skin and hair may indicate dehydration (excessive loss of water) by lack of pliability and lustre; or the presence of parasites such as lice, mites, or fleas; or the presence of ringworm infections and allergic reactions by the skin changes they cause. The poisoning of sheep by molybdenum in their hay may be diagnosed by the loss of colour in the wool

Table 3: Examples of Noninfectious Diseases of Animals

animal(s) affected	name(s) of disease	nature of disease
Hereditary diseases		
Pigs, calves, foals	congenital absence of skin (epitheliogenesis imperfecti)	complete absence of skin over parts of body; fatal a few days after birth
Swine, cattle	congenital porphyria (pink tooth)	causes anemia, wine-coloured urine; results from a biochemical defect in the metabolism of a component (porphyrin) of the iron-containing pigment (hematin) of hemoglobin, the oxygen-carrying protein of blood
Cattle	prolonged gestation (prolonged pregnancy)	may cause a three-week to three-month delay in birth of calves, which when born are either large or deformed; a special type in Guernsey and Jersey breeds results in death of calves at birth
Metabolic diseases		
Cattle (rarely sheep and pigs)	milk fever (parturient paresis)	caused in lactating cattle by loss of calcium into the milk; low levels of calcium in blood cause muscular weakness, circulatory collapse, and loss of consciousness; treatment includes replacing calcium
Cattle, sheep	ketosis (acetonemia in cattle; pregnancy toxemia in sheep)	occurs in lactating cattle following calving and in sheep in terminal stage of pregnancy (both times of increased need for carbohydrates); causes paralysis and death; complex treatment includes replacing carbohydrates
Horses	azoturia (paralytic myoglobinuria)	paralytic disease of unknown cause; occurs during exercise following a period of inactivity; muscles degenerate; results in paralysis, dark-red urine
Functional diseases		
Cattle, sheep	kidney and bladder calculi (urolithiasis)	cattle eating range grasses with high silicon content may develop obstructions (solid masses containing silicon and protein) in urinary system; similar effects may occur with diets high in phosphate, in which case the solid mass contains magnesium ammonium phosphate and protein
Cattle, sheep	bloat (ruminal tympany)	distension (caused by gases) of first two stomachs of cows; conditions preventing eructation (belching) of gases are major causes; occurs primarily when cattle overeat on leguminous pastures (alfalfa and clovers); often fatal
Cattle, horses	pulmonary emphysema (heaves in horses)	acute form occurs in cattle, chronic form in horses; alveoli (small terminal air sacs) in lung rupture, reducing surface area for oxygen transport; causes not yet clear, but disease may follow pneumonia and allergic reactions
Nutritional deficiency diseases		
Most animals	vitamin A deficiency	caused by dietary insufficiency of vitamin A or substance from which vitamin A is formed (carotene); numerous manifestations in young and adult animals include night blindness
Pigs	iron deficiency	caused by insufficient iron in diet; manifestations include severe anemia and poor growth
Diseases caused by chemical agents		
Pigs, cattle, horses	bracken-fern poisoning	fern contains thiaminase, which destroys vitamin B_1 (thiamine), thereby producing a vitamin deficiency in horses and swine; in cattle the bone marrow is affected, and deficiency of blood cells and excessive hemorrhages occur
Cattle, sheep, horses	rye-grass staggers	manifestations include either liver degeneration and photosensitization or uncoordination, convulsions, and paralysis; cause not yet established, but fungus on the rye grass plays some role in the liver degeneration
Cattle (occasionally other animals)	sweet-clover poisoning	caused by eating moldy sweet-clover hay, which contains the anticoagulant compound dicoumarol; manifestations include extensive hemorrhages and severe blood loss after injury
Cattle, sheep	molybdenum poisoning	results if pasture soil contains toxic quantities (three parts per million) of molybdenum, which replaces copper in body; symptoms include diarrhea, loss of hair colour, anemia
Diseases caused by physical agents		
Cattle	hardware disease (traumatic reticuloperitonitis)	objects such as nails and small pieces of baling wire may be eaten with feed; perforation and inflammation of first stomach (reticulum) may occur; surgery sometimes necessary
Cattle	brisket disease (mountain sickness)	occurs in cattle at high altitudes where levels of atmospheric oxygen are too low to provide oxygen required; manifestations include enlargement of right heart, congestive heart failure

of black sheep. Distension of the abdomen may indicate bloat in cattle or colic in horses.

Abnormal activities may have special diagnostic meaning to the veterinarian. Straining during urination is associated with bladder stones; increased frequency of urination is associated with kidney disease (nephritis), bladder infections, and a disease of the pituitary gland (diabetes insipidus). Excessive salivation and grinding of teeth may be caused by an abnormality in the mouth. Coughing is associated with pneumonia. Some diseases cause postural changes: a horse with tetanus may stand in a stiff manner. An abnormal gait in an animal made to move may furnish evidence as to the cause of a disease, as louping ill in sheep.

Structural features involved

Clinical examination. Following the general inspection of an animal thought to have contracted a disease, a more thorough clinical examination is necessary, during which various features of the animal are studied. These include the visible mucous membranes (conjunctiva of the eye, nasal mucosa, inside surface of the mouth, and tongue); the eye itself; and such body surfaces as the ears, horns (if present), and limbs. In addition, the pulse rate and the temperature are measured.

The veterinarian examines the visible mucous membranes of the eye, nose, and mouth to determine if jaundice, hemorrhages, or anemia are present. The conjunctiva, or lining of the eye, may exhibit pus in pinkeye infections, have a yellow appearance in jaundice, or exhibit small hemorrhages in certain systemic diseases. Examination of the nose may reveal ulcers and vesicles (small sacs containing liquid), as in foot-and-mouth disease, a viral disease of cattle, or vesicular exanthema, a viral disease of swine. Ulceration of the tongue may be apparent in animals suffering from actinobacillosis, a disease of bacterial origin (see Table 2).

A detailed examination of the eye may show abnormalities of the cornea resulting from such diseases as infectious hepatitis in dogs (Table 4), bovine catarrhal fever, and equine influenza. Cataract, a condition in which the passage of light through the lens of the eye is obstructed, may result from a disorder of carbohydrate metabolism (diabetes mellitus), infections, or a hereditary defect.

An elevated temperature (fever) resulting from the multiplication of disease-causing organisms may be the earliest sign of disease. The increase in temperature

Table 4: Some Common Diseases of Dogs

name(s) of disease	causative agent	nature of disease
Distemper	virus	affects nonvaccinated (nonimmunized) puppies in contact with infected animals; symptoms include loss of appetite, fever; inflammation of the brain is usual cause of death; some dogs may recover, but others have spastic tremors; foxes, wolves, mink, skunks, raccoons, and ferrets also susceptible
Infectious hepatitis (Rubarth's disease)	virus	affects dogs by causing hemorrhages and severe liver damage; affects foxes by causing inflammation of the brain; clinical signs are variable because disease symptoms vary from severe to inapparent (*i.e.*, no manifest signs)
Salmon poisoning	rickettsia	occurs after consumption of raw salmon or trout carrying rickettsial-infected flatworm (fluke) larvae (*Nanophyetus salmincola*); affects dogs, foxes, and coyotes primarily in the Pacific northwestern United States; symptoms include high fever, swollen lymph nodes; usually fatal within five days
Prostatitis	varied	inflammation of a gland near the urinary bladder (prostate gland) in male dogs; usually controlled by antibiotic drugs; other prostate-gland disorders may result from tumours (carcinoma, sarcoma) or from abnormal increase in cell multiplication (hyperplasia)
Congenital heart disease	inherited tendency	may occur in 1 percent of all dogs; heart disorders may lead to secondary diseases such as pneumonia, accumulation of fluid in body cavities, laboured breathing, edema; heart failure occurs
Hip dysplasia	apparently inherited tendency	crippling disorder common in many breeds (especially German shepherds); a shallow hip socket (acetabulum) results in an unstable hip joint, particularly during motion of hindleg
Kidney stones (calculi, urolithiasis)	hereditary, functional disturbance	calculi develop in kidney, bladder, and male urethra (tube from bladder to outside of body); surgery usually necessary; inherited types include cystine calculi in certain dachshunds and uric acid calculi in male dalmatians
Hypothyroidism	functional disturbance	thyroid gland may function marginally or be absent; symptoms include awkward, slow movement, coarse, dry coat; treatment includes iodine, thyroid preparations
Dermatitis	varied	common symptoms include skin inflammation and loss of hair; causative agents include nutritional deficiencies, bacterial infections, hypothyroidism, allergies, hormone imbalances, and parasites (*e.g.*, fleas, lice, mites, fly larvae, and ticks)
Strychnine poisoning	chemical compound	accidental ingestion of 0.75 milligram of the poison (found in rat poisons) per kilogram (about 2¼ pounds) of body weight may cause death from convulsions and respiratory distress
Glaucoma	hereditary tendency in some breeds	a group of eye diseases in which the retina and optic nerve are damaged; certain breeds have a hereditary tendency for the disease; other breeds develop glaucoma as a result of other eye disorders
Granulomatous colitis	not yet characterized	usually found in boxer dogs; symptoms include bloody diarrhea; severely and chronically affected dogs become emaciated; an infectious agent observed microscopically in the thickened colon has not yet been isolated or characterized
Pancreatitis	unknown	in acute types the gland may be destroyed because of inflammation from unknown causes; an animal that lives may develop diabetes mellitus or be unable to secrete enzymes from pancreas, or both, thus preventing digestion, which increases the appetite and causes progressive weight loss; treatment difficult

activates the body mechanisms necessary to fight off foreign substances. The pulse rate is useful in determining the character of the heartbeat and of the circulatory system.

TESTS AS DIAGNOSTIC AIDS

In many cases, the final diagnosis of an animal disease is dependent upon a laboratory test. Some involve measuring the amount of certain chemical constituents of the blood or body fluids, determining the presence of toxins (poisons), or examining the urine and feces. Other tests are designed to identify the causative agents of the disease. The removal and examination of tissue or other material from the body (biopsy) is used to diagnose the nature of abnormalities such as tumours. Specific skin tests are used to confirm the diagnoses of various diseases —*e.g.*, tuberculosis and Johne's disease in cattle and glanders in horses (Table 2).

Confirmation of the presence in the blood of abnormal quantities of certain constituents aids in diagnosing certain diseases. Abnormal levels of protein in the blood are associated with some cancers of the bone, such as multiple myeloma in horses and dogs. Animals with diabetes mellitus have a high level of the carbohydrate glucose and the steroid cholesterol in the blood. The combination of an increase in the blood level of cholesterol and a decrease in the level of iodine bound to protein indicates hypothyroidism (underactive thyroid gland). A low level of calcium in the serum component of blood confirms milk fever in lactating dairy cattle. An increase in the activities of certain enzymes (biological catalysts) in the blood indicates liver damage. An increase in the blood level of the bile constituent bilirubin is used as a diagnostic test for hemolytic crisis, a disease in which red blood cells are rapidly destroyed by organisms such as *Babesia* species in dogs and in cattle and *Anaplasma* species in cattle.

Blood tests

The examination of the formed elements of blood, including the oxygen-carrying red blood cells (erythrocytes), the white blood cells (neutrophils, eosinophils, basophils, lymphocytes, and monocytes), and the platelets, which function in blood coagulation, is helpful in diagnosing certain diseases. Examination of the blood cells of cattle may reveal abnormal lymphocytic cells characteristic of leukemia. Low numbers of leucocytes indicate the presence of viral diseases, such as hog cholera and infectious hepatitis in dogs. Neutrophil levels increase in chronic bacterial diseases, such as canine pneumonia and uterine infections in female animals. Elevated monocyte levels occur in chronic granulomatous diseases; *e.g.*, histoplasmosis and tuberculosis. Canine parasitism and allergic skin disorders are characterized by elevated eosinophil levels. Prolonged clotting time may be associated with a deficiency of platelets.

Anemia, a disease characterized by a deficiency in red blood cells, has many causes. They include hemorrhages from blood loss after injuries; the destruction of red blood cells by the rickettsia *Haemobartonella felis* in cats; incompatible blood transfusions in dogs; the inadequate production of normal red blood cells, which occurs in iron or cobalt deficiency after exposure to radioactive substances; general malnutrition; and contact with substances that depress the activity of bone marrow.

Poisonings occur commonly in animals. Some species are more sensitive to certain poisons than others. Swine develop mercury poisoning if they eat too much grain

Table 5: Some Common Diseases of Domestic Cats

name(s) of disease	causative agent	nature of disease
Feline distemper (panleukopenia, infectious enteritis)	virus	the most important viral disease of cats; wildcats and raccoons also susceptible; number of white blood cells decreases; fluid losses cause dehydration; treatment includes replacing fluid and preventing bacterial infections; vaccination advisable in kittens
Feline rhinotracheitis (pneumonitis, coryza, and influenza)	viruses (*e.g., Miyagawanella*)	any upper-respiratory viral infection with eye and nose involvement; common among cats; seldom causes death; similar to severe head cold in man; treatment includes antibiotic drugs
Feline picornavirus pneumonia	virus	symptoms include watery eyes, nasal inflammation; pneumonia; in severe cases, death is preceded by laboured breathing
Lymphocytic leukemia	virus	cancerous lymphocytic cells enter blood from a malignant tumour (lymphosarcoma); lymph nodes may become enlarged; a progressive anemia, an increase in the level of circulating immature lymphocytes, symptoms may precede death; no curative treatment known thus far
Urinary-bladder infection with stones and obstruction (cystitis)	unknown	urethra is obstructed with crystals of magnesium ammonium phosphate and mucus associated with infection of the urinary bladder (cystitis); death occurs if obstruction not relieved; exact cause of syndrome (set of symptoms) unknown but may be related to diet or to a virus
Ringworm (dermatomycosis)	fungus (*Microsporum canis; Trichophyton* species)	kittens most often affected; disease (a scaly, spreading skin condition) may be transmitted to children; some cats may carry the disease and show no clinical signs
Toxoplasmosis	protozoan (*Toxoplasma gondii*)	toxoplasmosis probably occurs frequently in cats; chronic forms affect abdominal organs; organisms may invade nearly any body tissue
Ear mites, ear inflammation (otitis externa)	mite (*Otodectes cynotis*)	acquired through contact of kitten with an infected mother; constant scratching of ears eventually causes raw sores and scabs; treatment includes killing the mites and controlling secondary bacterial infections
Osteodystrophy fibrosa (osteogenesis imperfecta)	nutritional deficiencies	once considered congenital in kittens but probably is nutritional, resulting from a calcium deficiency; lameness, first noted at ten to 15 weeks, may lead to paralysis of rear legs, slow growth, bone fractures, and severe pain; treatment includes a diet with adequate calcium
"Fur ball" disease	physical agent	hair balls accumulate in stomach as a result of constant grooming; surgical removal sometimes necessary

that has been treated with mercury compounds to retard spoilage. Dogs may be poisoned by the arsenic found in pesticides or by strychnine, which is found in rat poison. Many plants are poisonous to animals if eaten, such as bracken fern, which poisons cattle and horses, and ragwort plant, which contains a substance poisonous to the liver of cattle.

Examination of an animal's urine may reveal evidence of kidney diseases or diseases of the entire urinary system or a generalized systemic disease. The presence of protein in the urine of dogs indicates acute kidney disease (nephritis). Although constituents of bile normally are found in the urine of dogs, the quantity increases in dogs with the presence of infectious hepatitis, a disease of the liver. The presence of abnormal amounts of the simple carbohydrate glucose and of ketone bodies (organic compounds involved in metabolism) in an animal's urine is used to diagnose diabetes mellitus, a disease in which the pancreas cannot form adequate quantities of a substance (insulin) important in regulating carbohydrate metabolism. The urine of horses with azoturia (excessive quantities of nitrogen-containing compounds in the urine, Table 3) or muscle breakdown may contain a dark-coloured molecule called myoglobin.

The presence of eggs or parts of worms in the excrement of animals suspected of suffering from intestinal parasites, such as roundworms, tapeworms, or flatworms, aids in diagnosis. Feces that are light in colour, have a rancid odour, contain fat, and are poorly formed may indicate the existence of a chronic disease of the pancreas. Clay-coloured fatty feces suggest obstruction of the bile duct, which conveys bile to the intestine during digestion.

Identification of causative agents

The identification of a disease-causing micro-organism within an animal enables the veterinarian to choose the best drug for therapy. Agglutination tests, which utilize serum samples of animals and micro-organisms suspected of causing a disease, many times confirm the presence of the following bacterial diseases (Table 2): brucellosis in cattle and swine, salmonellosis in swine, leptospirosis in cattle, and actinobacillosis in swine and cattle. Other tests measure the antibodies (specific proteins formed in response to a foreign substance in the body) formed against a disease-causing agent, such as those that cause brucellosis, foot-and-mouth disease, infectious hepatitis in dogs, and fowl pest.

The modern veterinary diagnostic laboratory performs, in addition to the tests mentioned, tests of cells in the bone marrow; specific-organ-function tests (liver, kidney, pancreas, thyroid, adrenal, and pituitary glands); radioisotope tests, tissue biopsies, and histochemical analyses; and tests concerning blood coagulation and body fluids.

III. Survey of animal diseases

INFECTIOUS AND NONINFECTIOUS DISEASES

Disease, the departure from the normal physiological state of an organism, may be either infectious or noninfectious. The term infection implies an interaction between two living organisms, called the host and the parasite. Infection is a type of parasitism, which may be defined as the state of existence of one organism (the parasite) at the expense of another (the host). Agents (*e.g.*, certain viruses, bacteria, fungi, protozoans, worms, and arthropods) capable of producing disease are called pathogens. The term pathogenicity refers to the ability of a parasite to enter a host and produce disease; the degree of pathogenicity—that is, the ability of an organism to cause infection—is known as virulence. The capacity of a virulent organism to cause infection is influenced both by the characteristics of the organism and by the ability of the host to repel the invasion and to prevent injury. A pathogen may be virulent for one host but not for another. Pneumococcal bacteria, for example, have a low virulence for mice and are not found in them in nature; if introduced experimentally into a mouse, however, the bacteria overwhelm its body defenses and cause death.

Many pathogens (*e.g.*, the bacterium that causes anthrax, Table 2) are able to live outside the animal's body until conditions occur that are favourable for entering and infecting it. Pathogens enter the body in various ways—by penetrating the skin or an eye, by being eaten with food, or by being breathed into the lungs. After their entry into a host, pathogens actively multiply and produce disease by interfering with the functions of specific organs or tissues of the host. Table 2 lists some infectious and parasitic diseases of animals and the causative agents.

Before a disease becomes established in a host, the barrier known as immunity must be overcome. Immunity is the ability of a host's body to prevent or to overcome an invasion by a virulent pathogen. Defense against infection is provided by a number of chemical and mechanical barriers, such as the skin, mucous membranes and secretions, and components of the blood and other body fluids. Antibodies, which are proteins formed in response to a specific substance (called an antigen) recog-

Table 6: Some Common Diseases of Poultry

name(s) of disease	causative agent	nature of disease
Coccidiosis	protozoans (*e.g., Emieria tenella*)	affected birds show low egg production, poor growth rates, and high mortality; cecal coccidiosis (the cecum is a pouch in the large intestine), a serious disease caused by tenella; many protozoa produce intestinal symptoms
Blackhead (histomoniasis)	protozoan (*Histomonas meleagridis*)	may kill up to 50 percent of a turkey flock; symptoms include droopy wings and damage to liver and cecum; *Heterakis gallinae*, a worm in the turkey cecum, probably transmits the protozoan
Psittacosis (ornithosis)	virus (*Bedsonia*)	affects parakeets, canaries, parrots, pigeons, and other pet birds; symptoms include lack of appetite, ruffled feathers; can be transmitted to man
Avian lymphomatosis	virus	an infectious disease; manifestations include formation of tumours; three forms occur, depending on location of tumours: visceral (internal organs); neural (nerve); and ocular (eye) lymphomatosis
Pullorum disease	bacterium (*Salmonella pullorum*)	affects most species of fowl (chickens); mortality rate high, also reduces egg productivity of mature females; transmitted from egg-producing organs of hen to chick; disease causes hemorrhages throughout body
Digestive diseases in caged birds	functional disturbance	symptoms include slowly emptying, loosely hanging crop (digestive organ); deficiency of grit (particles that aid in digestion) results in poor nutrition; symptoms include obstruction of crop, vomiting, intestinal inflammation, and various liver disorders
Fractured legs in caged birds	physical agent	in canaries, about 70 percent of fractures involve the metatarsus (hindfoot); in budgerigars, 70 percent involve the tibia (hindleg)

nized by the body as foreign, are another important factor in preventing infection. Immunity among animals varies with species, general health, heredity, environment, and previous contact with a specific pathogen (see also IMMUNITY).

As certain bacterial species multiply, they may produce and liberate poisons, called exotoxins, into the tissues; other bacterial pathogens contain toxins, called endotoxins, which produce disease only when liberated at the time of death of the bacterial cell. Some bacteria, such as certain species of *Clostridium* and *Bacillus*, have inactive forms called spores, which may remain viable (*i.e.*, capable of developing into active organisms) for many years; spores are highly resistant to environmental conditions such as heat, cold, and chemical compounds called disinfectants, which are able to kill many active bacteria.

The term infestation indicates that animals, including spiny-headed worms (Acanthocephala), roundworms (Nematoda), flatworms (Platyhelminthes), and arthropods such as lice, fleas, mites, and ticks, are present in or on the body of a host. An infestation is not necessarily parasitic. Table 2 includes various infestations.

Noninfectious diseases

Noninfectious diseases are not caused by virulent pathogens and are not communicable from one animal to another (see Table 3). They may be caused by hereditary factors or by the environment in which an animal lives. Many metabolic diseases of animals are caused by an unsuitable alteration, sometimes brought about by man, in an animal's genetic constitution or in its environment. Metabolic diseases usually result from a disturbance in the normal balance of the physiological mechanisms that maintain stability, or homeostasis. Examples of metabolic diseases include overproduction or underproduction of hormones, which control specific body processes; nutritional deficiencies; poisoning from such agents as insecticides, fungicides, herbicides, fluorine, and poisonous plants; and inherited deficiencies in the ability to synthesize active forms of specific enzymes, which are the proteins that control the rates of chemical reactions in the body.

Excessive inbreeding (*i.e.*, the mating of related animals) among all domesticated animal species has resulted in an increase in the number of metabolic diseases and an increase in the susceptibility of certain animals to infectious diseases.

ZOONOSES

As stated previously, zoonoses are human diseases acquired from or transmitted to any other vertebrate animal. Zoonotic diseases are common in currently developing countries throughout the world and constitute, with starvation, the major threat to human health. More than 150 such diseases are known; some examples are listed in Table 9.

Types of zoonoses

Zoonoses may be separated into four principal types, depending on the mechanisms of transmission and epidemiology. One type includes the direct zoonoses, such as rabies and brucellosis, which are maintained in nature by one vertebrate species. The transmission cycle of the cyclozoonoses, of which tapeworm infections are an example, requires at least two different vertebrate species.

Table 7: Some Common Diseases of Fish

name(s) of disease	causative agent	nature of disease
Mouth fungus (cotton-wool disease)	bacterium (*Chondrococcus columnaris*)	fungus-like disease inaccurately named, since causative agent is a bacterium; a contagious disease; produces swollen lips, loss of appetite, and a "cotton-wool-like" growth on mouth; treatment utilizes antibiotic drugs
Tailrot	bacteria (*Haemophilus* and *Aeromonas* species)	infection may spread from fin and tail to body and cause death; disease may be controlled by surgery or use of drugs
Dropsy (ascites)	possibly associated with a bacterium (*Aeromonas punctata*)	characterized by accumulation of liquid in internal organs and tissues, inflammation of intestines, and infection of liver; epidemics can occur; most treatment except the antibiotic chloramphenicol unreliable
Red pest of eels	bacteria (*Vibrio anguillarum; Aeromonas* species)	*Vibrio* multiplies readily in salty water (1.5 to 3.5 percent) and can cause extensive blood-coloured areas on skin; *Aeromonas* species produce ulcers in the skin
Fish tuberculosis	bacteria (*Mycobacterium* species)	symptoms include loss of appetite, emaciation (leanness), skin defects, blood spots, ulcers, and cysts (on internal organs)
Eye fungus	fungus	infection follows damage to cornea of eye; a typical symptom is a white cotton-wool growth hanging from eye; untreated eye is destroyed within days; untreated fish die
Fish lice	louse (*Argulus* species)	bloodsucking parasites on the surfaces of many fish species
Skin flatworms (flukes)	platyhelminth	small parasites; cause skin colour to fade, blood spots, increased respiratory rate, and debilitation
Bursting of the swim bladder	physical agent	occurs if fish in deep water rise to surface too rapidly; fatal
Air embolism	physical agent	occurs if oxygen content of water is higher than normal, as when water temperature is higher than normal; bubbles of nitrogen gas in blood cause the disease

Table 8: Some Important Diseases of Some Common Laboratory Animals

name or type of disease	causative organism
Mice	
Bacterial diseases	
Mouse pneumonitis	*Miyagawanella bronchopneumoniae*
Eperythrozoonosis	*Eperythrozoon coccoides*
Mycoplasma infection	*Mycoplasma* species
Pus-producing lesions	*Pseudomonas aeruginosa*
Salmonellosis	*Salmonella*
Pseudotuberculosis	*Pasteurella pseudotuberculosis*
Rat-bite fever	*Streptobacillus moniliformis*
Fungal disease	
Dermatophytoses (ringworm)	*Trichophyton* species
Viral diseases	
Mouse pox (ectromelia)	Ectromelia virus
Poliomyelitis of mice (Theiler's disease)	Theiler's encephalomyelitis virus
EDIM (epizootic diarrhea of infant mice)	EDIM virus
LIVIM (lethal intestinal virus of infant mice)	LIVIM virus
Protozoal diseases	
Coccidiosis	*Eimeria falciformis*
Toxoplasmosis	*Toxoplasma gondii*
Nosematosis (encephalitozoonosis)	*Nosema cuniculi*
Helminthic diseases	
Mouse tapeworm	*Hymenolepsis nana*
Rat tapeworm	*Hymenolepsis diminuta*
Cat tapeworm	*Taenia taeniaeformis*
Mouse pinworm	*Aspicularis tetraptera*
Pinworm	*Syphacia obvelata*
Liver threadworm	*Capillaria hepatica*
Arthropod diseases	
Sucking louse	*Polyplax serratus*
Mite	*Psorergates simplex*
Rats	
Bacterial and fungal diseases	
Mycoplasma infection (associated with chronic pneumonia)	*Mycoplasma* species
Salmonellosis	*Salmonella* species
Infectious anemia	*Haemobartonella muris*
Pus-forming lesions	*Pseudomonas aeruginosa* *Pasteurella pneumotropica*
Rat-bite fever	*Streptobacillus moniliformis* *Diplococcus pneumoniae*
Leptospirosis (Weil's disease)	*Leptospira* species *Corynebacterium kutscheri*
Dermatophytoses (ringworm; fungus)	*Trichophyton*
Viral diseases	
Pneumonitis	rat virus (Kilham)
Salivary-gland disease	rabula virus
Hemorrhagic encephalitis	hemorrhagic encephalopathy virus (HER)

name or type of disease	causative organism
Rats (cont.)	
Protozoal diseases	
Coccidiosis	*Eimeria nieschulzi* *Hepatozoon muris*
Toxoplasmosis	*Toxoplasma gondii*
Nosematosis (encephalitozoonosis)	*Nosema cuniculi*
Helminthic diseases	
Mouse tapeworm	*Hymenolepsis nana*
Rat tapeworm	*Hymenolepsis diminuta*
Cat tapeworm	*Taenia taeniaeformis*
Bladder threadworm	*Trichosomoides crassicauda*
Cecal worm	*Heterakis spumosa*
Liver threadworm	*Capillaria hepatica*
Arthropod diseases	
Sucking louse	*Polyplax spinulosa*
Ear mite	*Notoedres notoedres*
Hair mite	*Radifordia ensifera*
Guinea pigs	
Bacterial and fungal diseases	
Streptococcal infections (cervical abscesses and lymphadenitis)	*Streptococcus* species
Salmonellosis	*Salmonella* species
Bronchopneumonia	*Bordatella bronchiseptica*
Respiratory infections	*Klebsiella* species
Dermatophytoses (ringworm, fungus)	*Trichophyton* species
Viral diseases	
Cytomegalic inclusion disease	salivary-gland virus
Lymphocytic choriomeningitis	LCM virus
Protozoal diseases	
Coccidiosis	*Eimeria caviae*
Toxoplasmosis	*Toxoplasma gondii*
Hamsters	
Cytomegalic inclusion disease	salivary-gland virus
Encephalomyocarditis	picorna virus
Wet tail (regional ileitis)	virus, bacterium
Salmonellosis	*Salmonella* (bacterium)
Mouse tapeworm	*Hymenolepsis nana* (helminth)
Rabbits	
Pasteurellosis (snuffles)	*Pasteurella* (bacterium)
Spirochetosis (vent disease)	*Spirochaeta* (bacterium)
Mucoid enteritis	bacterium
Myxomatosis	myxoma virus
Rabbit pox	virus
Coccidiosis	*Eimeria* species (protozoan)
Ear mites	*Notoedres notoedres*
Nosematosis	*Nosema cuniculi*

Both vertebrate and invertebrate animals are required as intermediate hosts in the transmission to man of metazoonoses; arboviral and trypanosomal diseases are examples of metazoonoses. The cycles of saprozoonoses (*e.g.*, histoplasmosis) may require, in addition to vertebrate hosts, specific environmental sites or reservoirs.

Most animals that serve as reservoirs for zoonoses are domesticated and wild animals with which man commonly associates. People in occupations such as veterinary medicine and public health, therefore, have a greater exposure to zoonoses than do those in occupations less closely concerned with animals.

In addition to the numerous human diseases spread by contact with the parasitic worm helminth and by contact with arthropods (see Table 9), many diseases are transmitted by the bites and the venom of certain animals; poisonous or diseased food animals also transmit diseases. Dog bites may result in serious injury to tissues and also can transmit bacterial infections and rabies, a disease of viral origin. The bite of a diseased rat may transmit any of several diseases to man, including plague, salmonellosis, leptospirosis, and rat-bite fevers. Cat-scratch disease may be transmitted through cat bites, and the deadly herpes B virus can be spread by monkey bites. The bites of venomous snakes and fish account for considerable human discomfort and death. About 200 of the 2,500 known species of snakes are able to cause human disease. One estimate for snakebite deaths throughout the world is 30,000 to 40,000 per year, with the vast majority occurring in Asia. Poisonous wild animals inadvertently used for food include animals harbouring the anthrax bacillus, as well as those containing the causative agents of salmonellosis, trichinosis, and fish-tapeworm infection. The flesh of various types of fish is toxic to man. Japanese puffers, for example, contain the poisonous chemical compound tetrodotoxin; scombroid fish harbour *Proteus morganii*, which causes gastrointestinal diseases; and mullet and surmullet cause nervous disturbances.

Control of zoonoses

Approaches to the control of zoonoses differ according to the type under consideration. Because the majority of direct zoonoses and cyclozoonoses and some saprozoonoses are most effectively controlled by techniques involving the animal host, methods used to combat these diseases are almost entirely the responsibility of veterinary medicine. A good example is the elimination of stray dogs, for they are an important factor in the control of zoonoses such as rabies, hydatid disease, and visceral larva migrans. In addition, the control of diseases such as brucellosis and tuberculosis in cattle involves a combination of methods—mass immunization, diagnosis, slaughter of infected animals, environmental disinfection, and quarantine. Several supportive measures for the control of disease are useful in some cases. Air-sanitation measures are helpful in direct zoonoses in which human illness is spread by droplets or dust, and zoonotic infections that are spread through a fluid medium, such as water or milk, sometimes can be controlled. Heat, cold,

Table 9: A Partial List of Zoonoses

disease	causative organism	animals principally involved
Viral diseases		
Arbovirus infections		rodents, birds, equines, goats, sheep, monkeys, swine, marsupials
Febrile illnesses	various arboviruses	
Hemorrhagic fever		
Epidemic nephroso-nephritis		
Encephalitis (mosquito-borne and tick-borne)		
Encephalomyocarditis	encephalomyocarditis virus	rodents
Herpes B virus disease	herpes B virus	monkeys
Herpes T (= M) virus infection	herpes T (= M) virus	monkeys
Influenza	influenza virus type A	swine
Lymphocytic choriomeningitis	lymphocytic choriomeningitis virus	mice, dogs, monkeys
Newcastle disease	Newcastle disease virus	chickens
Poxvirus infections		
Buffalopox	buffalopox virus	buffalo
Camelpox	camelpox virus	camels
Cowpox	cowpox or vaccinia virus	cattle
Orf (contagious ecthyma)	contagious ecthyma virus	sheep and goats
Paravaccinia (milkers' nodules)	paravaccinia virus	cattle
Yaba disease	yaba virus	monkeys
Rabies	rabies virus	carnivores, bats, and other wild animals
Sendai virus disease	sendai virus	swine, rodents
Cat-scratch disease	cat-scratch virus	cats
Rickettsial diseases		
Flea-borne		
Murine (endemic typhus)	*Rickettsia mooseri*	rats, mice
Mite-borne		
Rickettsial pox	*R. akari*	mice
Scrub typhus (tsutsugamushi)	*R. tsutsugamushi*	rodents
Tick-borne		
(North) Queensland tick typhus	*R. australis*	bandicoots, rodents
Spotted fever (including Rocky Mountain, Brazilian, and Colombian spotted fevers)	*R. rickettsii*	dogs, rodents, and other animals
Fièvre boutonneuse	*R. conorii*	dogs, rodents
Kenya typhus		
South African tick typhus		
Indian tick typhus		
North Asian tick-borne rickettsiosis	*R. sibericus*	rodents
Rickettsial diseases (cont.)		
Q fever	*Coxiella burnetii*	cattle, sheep, goats, and other domesticated and wild mammals, birds
Bedsonia infection		
Psittacosis (ornithosis)	Psittacosis (P.L.T.) group (*Bedsonia*)	psittacines and other birds
Bacterial diseases		
Anthrax	*Bacillus anthracis*	ruminants, equines, swine
Brucellosis	*Brucella abortus, B. suis, B. melitensis*	cattle, swine, goats, sheep, horses
Enterobacterial infections		
Arizona infections	*Arizona* species	poultry, swine, dogs
Colibacillosis	*E. coli*	poultry, swine, dogs
Salmonellosis	*Salmonella* species	mammals and birds
Erysipeloid	*Erysipelothrix rhusiopathiae*	swine, poultry, fish
Glanders	*Actinobacillus mallei*	equines
Leptospirosis	*Leptospira interrogans*	rodents, dogs, swine, cattle, bandicoots
Listeriosis	*Listeria monocytogenes*	rodents, sheep, cattle, swine
Melioidosis	*Pseudomonas pseudomallei*	rodents, sheep, cattle, swine
Pasteurellosis	*Pasteurella multocida, P. haemolytica*	mammals, birds
Plague	*Pasteurella pestis*	rodents
Pseudotuberculosis	*Pasteurella pseudotuberculosis*	rodents, cats, fowls
Rat-bite fever	*Spirillum minus, Streptobacillus moniliformis*	rodents
Relapsing fever (tick-borne)	*Borrelia* species	rodents
Staphylococcosis	*Staphylococcus aureus*	cattle, dogs, occasionally other animals
Streptococcosis	*Streptococcus* species	mammals
Tuberculosis	*Mycobacterium tuberculosis var. hominis*	dogs, swine, monkeys
	M. bovis	cattle, goats, swine, cats
	M. avium	poultry, swine, cattle
Tularaemia	*Pasteurella tularensis*	rabbits, hares, sheep, wild rodents
Vibriosis	*Vibrio fetus; V. parahaemolyticus*	cattle, sheep fish
Fungal disease		
Dermatophytosis		
Ringworm, favus	*Microsporum* species	cats, dogs, horses
	Trichophyton species	horses, cattle, poultry, small mammals

and irradiation are effective in killing the immature forms of *Trichinella spiralis*, the causative agent of trichinosis, in meat; and certain antibiotic drugs help to prevent deterioration of food.

The control of metazoonoses may be directed at the infected vertebrate hosts, at the infected invertebrate host, or at both. Particularly effective in this instance has been the use of chemical insecticides to attack the invertebrate carriers of specific infections, even though several difficulties have been encountered—for example, the inaccessibility of the invertebrate to the chemicals, which occurs with organisms that breed in swiftly flowing waters or in dense vegetation, and the development of insecticide resistance by the organisms. Insecticides are used to destroy the mosquitoes that spread malaria (*Anopheles*). Mechanical filters placed across irrigation ditches help to prevent the dissemination of the snails that transmit *Schistosoma mansoni*, a parasitic flatworm.

DISEASE PREVENTION, CONTROL, AND ERADICATION

Prevention is the first line of defense against an infectious or a noninfectious disease. At least four preventive techniques are available for use in the prevention of disease in an animal population. One is the exclusion of causative agents of disease from specific geographical areas, or quarantine. A second preventive tool utilizes control methods such as immunization, environmental control, and chemical agents to protect specific animal populations from endemic diseases, diseases normally present in an area. The third preventive measure concerns the mass education of people about disease prevention. Finally, early diagnosis of illness among members of an animal population is important, so that disease manifestations do not become too severe, and so that affected animals can be more easily managed and treated.

Quarantine—or the restriction of movement of animals suffering from or exposed to infections such as bluetongue and scrapie (in sheep), foot-and-mouth disease (in cattle), and rabies (in dogs)—is one of the oldest tools known to preventive medicine. The principle of quarantine of domesticated animals was applied as early as Roman times. The establishment of international livestock quarantine in the United States in 1890 provided for the holding of all imported cattle, sheep, and swine at the port of entry for 90, 15, and 15 days, respectively. In this way, such diseases as Nairobi sheep disease, surra, and infections caused by *Brucella melitensis* have been eliminated or excluded from the United States, but international quarantine barriers did not prevent the entry of bluetongue, scrapie, and the tick *Rhipicephalus evertsi*, which acts as a carrier for several animal diseases. On the other hand, long-term quarantine of all dogs entering Great Britain has been effective since its initiation in 1919. It is possible that aircraft may pose new problems regarding livestock-disease quarantine since many disease carriers (*e.g.*, insects and viruses) may be accidentally carried by plane into a country.

Table 9: A Partial List of Zoonoses (continued)

disease	causative organism	animals principally involved
Protozoal diseases		
Amebiasis	*Entamoeba histolytica*	dogs, lower primates
Balantidiasis	*Balantidium coli*	swine
Coccidiosis	*Isospora* species	dogs
Leishmaniasis		
Kala Azar	*Leishmania donovani*	dogs
Oriental sore	*Leishmania tropica*	dogs, rodents
American	*Leishmania* species	dogs, wild mammals
Malaria	*Plasmodium knowlesi*	monkeys
	Plasmodium simium	monkeys
	Plasmodium cynomolgi	monkeys
Pneumocystis infection	*Pneumocystis carinii*	dogs
Toxoplasmosis	*Toxoplasma gondii*	mammals, birds
Trypanosomiasis	*Trypanosoma cruzi*	dogs, small mammals
	Trypanosoma rangeli	
	Trypanosoma rhodesiense	antelope, cattle
Platyhelminthic diseases		
Trematode (fluke) diseases		
Amphistomiasis	*Gastrodiscoides hominis*	swine
Cercarial dermatitis	*Schistosoma* species	birds, mammals
Clonorchiasis	*Clonorchis sinensis*	dogs, cats, swine, wild mammals, fish
Dicrocoeliasis	*Dicrocoelium* species	ruminants
Echinostomiasis	*Echinostoma ilocanum*	cats, dogs, rodents
	Echinostoma species	
Fascioliasis	*Fasciola hepatica, F. gigantica*	ruminants
Fasciolopsiasis	*Fasciolopsis buski*	swine, dogs
Heterophyiasis	*Heterophyes heterophyes* (and other heterophids)	cats, dogs, fish
Metagonimiasis	*Metagonimus yokogawai*	cats, dogs, fish
Opisthorchiasis	*Opisthorchis felineus*	cats, dogs
	Opisthorchis viverrini, other species	wildlife, fish
Paragonimiasis	*Paragonimus westermani*, other species	cats, dogs, wildlife
Schistosomiasis	*Schistosoma japonicum*	wild and domestic mammals
	S. mansoni	baboons, rodents
	S. mattheei, occasionally other species	cattle, sheep, antelopes
Cestode (tapeworm) diseases		
Bertiella infection	*Bertiella studeri*	primates
Diphyllobothriasis	*Diphyllobothrium latum*	fish, carnivores
Dipylidiasis	*Dipylidium caninum*	dogs, cats
Echinococcosis	*Echinococcus granulosus*	dogs, wild carnivores, domestic and wild ungulates
	E. multilocularis	foxes, dogs, rodents
Hymenolepiasis	*Hymenolepis diminuta, H. nana*	rats, mice
Inermicapsifer infection	*Inermicapsifer madagascarensis*	rodents
Platyhelminthic diseases (cont.)		
Sparganosis	Pseudophyllidea tapeworms	mice, carnivores including cats, and other vertebrates
Taeniasis, cysticercosis, and coenuriasis	*Taenia saginata*	cattle
	Taenia solium	swine
	Multiceps multiceps	sheep, dogs
Nematode diseases		
Ancylostomiasis	*Ancylostoma ceylanicum*, other species	dogs
Ascariasis	*Ascaris suum*	swine
Capillariasis	*Capillaria hepatica*	rodents
Dracuncoliasis	*Dracunculus medinensis*	dogs, other carnivores
Filariasis	*Brugia malayi*	primates, other mammals
	Dirofilaria species, occasionally other species	cats, dogs, other mammals
Larva migrans	*Ancylostoma braziliense*, other species	cats, dogs
	Angiostrongylus cantonensis	rats
	Anisakis species	fish
	Gnathostoma spinigerum	cats, dogs, other vertebrates
	Toxocara canis, other ascarid species	dogs, other vertebrates
Oesophagostomiasis	*Oesophagostomum apiostomum*	primates
Strongyloidiasis	*Strongyloides stercoralis*, occasionally other species	dogs, primates
Ternidens infection	*Ternidens deminutus*	primates
Trichinosis	*Trichinella spiralis*	swine, rodents, wild carnivores, marine mammals
Trichostrongylosis	*Trichostrongylus colubriformis*, occasionally other species	ruminants
Arthropod diseases		
Acariasis	*Sarcoptes* species	domesticated animals
Tunga infections	*Tunga penetrans*	domesticated and wild mammals
Myiasis	*Cochliomyia, Cordylobia, Dermatobia, Gastrophilus, Hypoderma, Oestrus*, and other genera	mammals
Pentastomid infections (including Halzoun)	*Linguatula* species, *Armillifer* species, *Porocephalus* species	dogs, snakes, and other vertebrates

Contact through travel

Mass immunization as a preventive technique has the advantage of allowing the resistant animal freedom of movement, unlike environmental control, in which the animal is confined to the controlled area; immunization may, however, provide only short-lived and partial protection. Mass-inoculation techniques against diseases such as Newcastle disease in chickens and distemper in mink and dogs have been successful. Animal diseases have been prevented by methods involving environmental control, including the maintenance of safe water supplies, the hygienic disposal of animal excrement, air sanitation, pest control, and the improvement of animal housing. One specific environmental program, called the portable-calf-pen system, involves routine movement of the pens to avoid a concentration of specific pathogens in them. Other programs involve the utilization of automatic and sanitary watering and feeding equipment and buildings with environmental controls. The use of chemical compounds to prevent illness (chemoprophylaxis) includes a variety of pesticides, which are used to kill insects that transmit diseases, and substances either used internally or applied to the animal's body to prevent the transmission or the development of a disease. An example is the use of sulfonamide drugs in the drinking water of poultry to prevent coccidiosis (see Table 6). Environmental-control methods in the poultry industry have resulted in the most efficient means of poultry production developed thus far.

The early detection of a disease in a population of animals—a herd of cattle, for example—is particularly useful in controlling certain chronic infectious diseases, such as mastitis, brucellosis, and tuberculosis, as well as certain noninfectious diseases such as bloat. Laboratory tests—the agglutination test in pullorum disease, the tuberculin skin test for tuberculosis, the examination of feces for eggs of specific parasites, the physical and chemical tests performed on milk to diagnose bovine mastitis—are used for the early detection of diseases in an animal population.

Methods of disease control and eradication have been successful in various countries throughout the world. In the United States, for example, the test and slaughter technique, in which simple tests are used to confirm the existence of diseased animals in a population, followed by slaughter of the affected animals, has been of great value in controlling infectious and hereditary diseases, including dourine, a venereal disease in horses, fowl plague, and foot-and-mouth disease in cattle and deer. Bovine tuberculosis has been eliminated from Denmark, Finland, and The Netherlands and reduced to a low level in various other countries, including Great Britain, Japan, the United States, and Canada, by the test and slaughter method. Many infectious diseases have been eradicated from Great Britain—sheep pox, rinderpest, pleuropneumonia, glanders, and rabies. Diseases eliminated from Australia by a combination of methods—control of agents that carry disease, the test and slaughter technique, the use of chemical agents, and, more recent-

Table 10: Animal Diseases Usually Confined to Certain Regions of the World

name(s) of disease	animal(s) affected	causative organism	distribution	nature of disease
African horse sickness (AHS), equine plague, pestis equorum, perdesiekte	primarily horses, donkeys, mules (occasionally zebras and dogs)	AHS virus	primarily Africa and Middle East; occasionally India, Pakistan	a seasonal disease occurring in late summer; acute form, sometimes fatal within five days, involves excessive fluid in lungs; symptoms of other forms include accumulation of fluid in body cavities
African swine fever (ASF), warthog disease, Montgomery's disease	swine	ASF virus	primarily Kenya and South Africa; occasionally Europe	highly contagious; usually fatal; resembles hog cholera in clinical manifestations (high fever, weakness in hindlegs, and hemorrhages throughout body) but can be distinguished by laboratory tests and isolation of the virus
Contagious pleuropneumonia, lung plague	cattle, buffalo, yaks, sheep, goats	*Mycoplasma mycoides*	Africa, Australia, Asia, Europe	transmitted by direct animal contact or by contaminated objects; an acute disease producing pneumonia and inflammation of the lung lining; vaccines ineffective because different strains of the organism occur throughout the world
East coast fever, theileriasis, Rhodesian red water or tick fever	cattle, African and Indian water buffalo	protozoan (*Theileria parva*)	Central Africa; East Africa	usually fatal; transmitted by three ticks containing pathogen; symptoms include high fever, swelling of lymph glands; not yet prevented effectively by vaccination
Foot-and-mouth disease (FMD), aphthous fever, aftosa	cattle, swine, sheep, goats	FMD virus	worldwide except North America, Central America, New Zealand	symptoms include high fevers, drool from mouth, where vesicles and ulcers form, and lameness; causes great economic losses throughout world; effective vaccines available
Fowl plague, fowl pest	birds, including chickens and turkeys	fowl-plague virus	Europe, Middle and Far East, Argentina, Japan	may cause no apparent symptoms; apparent symptoms include lack of appetite, swollen head, laboured breathing, and hemorrhaging
Heartwater, drunk bull sickness	cattle, sheep, goats	rickettsia (*Cowdria ruminantium*)	Africa (southern half); Madagascar	disease has acute and mild forms; symptoms include water in the membrane around heart and in the lung cavity, hemorrhages, and twitching
Louping ill (LI), infectious encephalomyelitis in sheep, trembling ill	primarily sheep (also cattle and man)	LI virus	British Isles, Czechoslovakia, U.S.S.R.	transmitted by bite of sheep tick; characterized by fever, dullness followed by excitement, muscular spasms, leaping gait, convulsions, and death
Nagana, tsetse disease, trypanosomiasis	most domesticated animals	protozoan (*Trypanosoma* species)	Africa	may be acute or inapparent; symptoms may include anemia resulting from red-blood-cell destruction; pathogen transmitted by tsetse fly (over 20 species of *Glossina*); prevents effective cattle production in nearly all of West Africa
Rift Valley fever (RVF), infectious enzootic hepatitis	cattle, sheep (occasionally man)	RVF virus	Central and South Africa	spread by bloodsucking insects associated with wild animals; symptoms include abdominal pain resulting from liver damage; young animals usually die; mature ones may recover
Rinderpest, cattle plague	cattle, sheep, goats, wild ruminants; yaks, caraboo, gazelles, deer	rinderpest virus	primarily Asia, Africa, Philippines; rarely Europe	rapidly fatal; symptoms include fluid losses (dehydration) from diarrhea caused by massive pathological changes (*e.g.*, hemorrhages, ulcers) in intestinal tract
Surra	primarily in camels and horses; many animals susceptible	protozoan (*Trypanosoma evansi*)	primarily Far East (*e.g.*, China), India, Near East (*e.g.*, Iran); North Africa	transmitted by bloodsucking flies and mosquitoes; symptoms include anemia, loss of weight, large swellings in limbs, abdomen, and sex organs
Teschen disease, swine encephalomyelitis, porcine poliomyelitis	swine	Teschen virus	primarily Europe	symptoms include prostration, immobilization, nervous tremors, convulsions, paralysis of legs

ly, biological control—include hog cholera, rinderpest, scrapie, glanders, surra, rabies, and foot-and-mouth disease.

In biological control, enemies of the agents that transmit the disease, enemies of the reservoir host, or a specific parasite are introduced into the environment. If a natural enemy of the tsetse fly could be found, for example, African sleeping sickness in man and trypanosomiasis in cattle could be controlled in West Africa. Successful biological control of the European-rabbit population in Australia has been accomplished through the use of the myxomatosis virus, which is transmitted by mosquitoes and causes the formation of malignant tumours. Although the Brazilian white rabbit is relatively unaffected by the virus, it causes rapid death in the European rabbit. The elimination of the European rabbit in France by the virus was accompanied by a decrease in tick-borne typhus in people, suggesting that the rabbit may be a significant intermediate host for the causative agent, *Rickettsia conorii*. Screwworms, an immature form of the fly *Cochliomyia hominivorax*, have been eradicated in the United States by the release of more than 3,000,000,000 sterilized males.

Disease control and elimination programs require many sophisticated techniques, in addition to diagnosis and the slaughter of affected animals. They include: the control of insects known to transmit diseases; the cooperation of animal owners with disease-control programs; the development through research of new diagnostic tests for use on large populations; the eradication of animal species from areas in which they are known to transmit disease; sterilization of strains of animals known to carry inheritable metabolic diseases; and effective meat inspection.

BIBLIOGRAPHY. M.R. CLARKSON, *Livestock Health and Human Food Needs*, proceedings of a National Research Council, National Academy of Sciences symposium on the use of drugs in animal feeds (1969); FOOD AND AGRICULTURAL ORGANIZATION (FAO), *Third World Food Survey* (1963); and W.R. PRITCHARD, "Increasing Protein Foods Through Improving Animal Health," *Proc. Nat. Acad. Sci. USA*, 56:360–369 (1966), three general references on the role of veterinary medicine in combatting world hunger; UNITED STATES DEPARTMENT OF AGRICULTURE, "Losses in Agriculture," *Agriculture Handbook 291* (1965), a recent review of the economic losses to agriculture from animal diseases and parasites; C.W. SCHWABE, *Veterinary Medicine and Human Health*, 2nd ed. (1969), a comprehensive reference on medical public health; JOINT FAO/WHO EXPERT COMMITTEE ON ZOONOSES, *Third Report of World Health Organization, Technical Report Series No. 378* (1967), report on the many animal diseases transmissible to man; J.F. SMITHCORS, *Evolution of the Veterinary*

Art (1957), the only complete book on veterinary medical history and the history of the existence of animal diseases; H.E. BIESTER and L.H. SCHWARTE (eds.), *Diseases of Poultry*, 5th ed. (1965); D.C. BLOOD and J.A. HENDERSON, *Veterinary Medicine*, 3rd ed. (1968); E.J. CATCOTT (ed.), *Canine Medicine* (1968); C. VAN DUIJN, *Diseases of Fishes*, 2nd ed. (1967); H. KIRK, *Index of Treatment in Small-Animal Practice*, 3rd ed. (1954); E.C. MELBY and R.A. SQUIRE, *Proceedings of the American Animal Hospital Association, 36th Annual Meeting, Washington, D.C.* (1969); I.A. MERCHANT and R.D. BARNER, *An Outline of the Infectious Diseases of Domestic Animals*, 3rd ed. (1964); and the UNITED STATES DEPARTMENT OF AGRICULTURE, *The Year Book of Agriculture: Animal Diseases* (1956), all major textbooks about animal diseases; G.F. BODDIE, *Diagnostic Methods in Veterinary Medicine*, 6th ed. (1969); G.O. DAVIES, *Gaiger and Davies' Veterinary Pathology and Bacteriology*, 4th ed. (1955); and J.J. KANEKO and C.E. CORNELIUS (eds.), *Clinical Biochemistry of Domestic Animals*, 2nd ed., vol. 1 (1970), general references to physical and laboratory diagnostic techniques; C.E. CORNELIUS, "Biomedical Models in Veterinary Medicine," *Amer. J. Med.*, 40: 165–169 (1966), and "Animal Models: A Neglected Medical Resource," *New Eng. J. Med.*, 281:934–944 (1969); R.W. LEADER, "Lower Animals, Spontaneous Disease and Man," *Arch. Path.*, 78:390–404 (1964); and W.R. PRITCHARD, "Animal Research in the New Biology," *Lab. Anim. Care*, 18: 230–233 (1968), general articles indicating the use of animals as models for studying counterpart diseases in man.

(C.E.Co.)

Diseases of Plants

All species of plants, wild and cultivated alike, are subject to disease. Although each species is susceptible to characteristic diseases, these are, in each case, relatively few in number. The occurrence and prevalence of plant diseases vary from season to season, depending on the presence of the pathogen (disease-causing organism), environmental conditions, and the crops and varieties grown. Some plant varieties are particularly subject to outbreaks of diseases; others are more resistant to them.

This article is divided into the following sections:

I. General features
Nature and importance of plant diseases
Disease development and spread
Diagnosis of plant diseases
II. Classification and control
Noninfectious disease-causing agents
Infectious disease-causing agents
Principles of disease control

I. General features

NATURE AND IMPORTANCE OF PLANT DISEASES

Economic impact of plant diseases

Some 80,000 diseases of plants have been recorded throughout the world; more than 50,000 occur in the United States, where the average annual loss from plant diseases and attack by nematodes (small roundworms), despite extensive control measures, is estimated at about $4,000,000,000, or equivalent to the production of over 40,000,000 acres (16,000,000 hectares). This is a loss of about 12 percent, but it is small compared with the total loss that would occur if adequate disease-control measures were unavailable for most crops. In many developing areas of the world, where controls are not used or are unavailable, losses of up to 30 to 50 percent may occur on major crops each year.

Plant diseases are known from times preceding the first writings of mankind. Fossil evidence indicates that plants were affected by disease 250,000,000 years ago. The Bible and other early writings mention diseases, such as rusts, mildews, blights, and blast, that have caused famine and drastic changes in the economy of nations since the dawn of recorded history. Other plant-disease outbreaks with similar far-reaching effects in recent times include potato late blight (*Phytophthora infestans*) in Ireland (1845–60); Sigatoka leaf spot (*Mycosphaerella musicola*) and Panama disease (*Fusarium oxysporum* form species *cubense*) in Central America (1900–65); powdery and downy mildews (*Uncinula necator* and *Plasmopara viticola*) of grape in France (1851 and 1878); coffee rust (*Hemileia vastatrix*) in Ceylon (starting in the 1870s); *Fusarium* wilts of cotton and flax (*Fusarium oxysporum* form species *vasinfectum* and *F. oxysporum* form species *lini*), southern bacterial wilt of tobacco (*Pseudomonas solanacearum*—early 1900s); black stem rust of wheat (*Puccinia graminis tritici*—1916, 1935, 1953–54); and southern corn-leaf blight (*Helminthosporium maydis*—1970) in the United States.

Diseases—a normal part of nature. Plant diseases are a normal part of nature and one of many ecological factors that help keep the hundreds of thousands of living plants and animals in balance with one another. Man has carefully selected and cultivated plants for food, clothing, shelter, fibre, and beauty for thousands of years. Disease is just one of many hazards that must be considered when plants are taken out of their natural environment and grown in pure stands under what are often abnormal conditions.

Many valuable crop and ornamental plants are very susceptible to disease and would have difficulty surviving in nature without man's help. Cultivated plants are often more susceptible to disease than are their wild relatives. This is because large numbers of the same species or variety, having a uniform genetic background, are grown close together, sometimes over many thousands of square miles. A pathogen, or disease-causing organism, may spread rapidly when established under these conditions.

Definitions of plant disease. There is no sharp distinction between the healthy and diseased condition (see DISEASE). In general, a plant becomes diseased when it is continuously disturbed by some causal agent that results in an abnormal physiological process. The primary causal factor, which may be a living disease-causing organism (pathogen) or an unfavourable environmental condition, disrupts the plant's normal structure, growth, function, or activities; the result is an interference with one or more essential physiological or biochemical systems. The resulting disease is expressed in characteristic pathological conditions called symptoms. Knowledge of normal growth habits, variety characteristics, and normal variability of plants within a species—as these relate to the conditions under which the plants are growing—is required for recognition of disease.

Injury and disease contrasted

Injury, in contrast, usually results from a momentary or discontinuous effect, such as lightning, hail, extreme variations in temperature or availability of water, toxic chemicals in the air and soil, or wounding by an insect or rodent.

The study of plant diseases is called plant pathology. Pathology is derived from the two Greek words *pathos* (suffering, disease) and *logos* (discourse, study). Plant pathology thus means a study of the diseased plant.

Etiology is the study of the nature, classification, and life histories of the pathogens (*e.g.*, bacteria, fungi, viruses, mycoplasmas, nematodes, and parasitic flowering plants) causing infectious diseases. Many pathogens, especially among the bacteria and fungi, spend part of their life cycles as pathogens and the remainder as saprophytes —organisms that live on dead organic matter.

DISEASE DEVELOPMENT AND SPREAD

Pathogenesis and saprogenesis. Pathogenesis is the stage of disease in which the pathogen is in intimate association with living host tissue. Three fairly distinct stages are involved: (1) Inoculation—transfer of the pathogen to the infection court, or area in which invasion of the plant occurs. The infection court may be the unbroken plant surface, a variety of wounds, or natural openings—*e.g.*, stomates (microscopic pores in leaf surfaces), hydathodes (stomate-like openings that secrete water), or lenticels (small openings in tree bark). (2) Incubation—the period of time between the arrival of the pathogen in the infection court and the appearance of symptoms. (3) Infection—the appearance of disease symptoms accompanied by the establishment and spread of the pathogen.

Saprogenesis is the part of the pathogen's life cycle when it is not in vital association with living host tissue and either continues to grow in dead host tissue or becomes dormant. During this stage, some fungi produce their sexual fruiting bodies; the apple scab (*Venturia inaequalis*), for example, produces perithecia, flask-shaped spore-

producing structures, in fallen apple leaves. Other fungi produce compact resting bodies, such as the sclerotia formed by certain root- and stem-rotting fungi (*Rhizoctonia solani* and *Sclerotinia sclerotiorum*) or the ergot fungus (*Claviceps purpurea*). These resting bodies, which are resistant to extremes in temperature and moisture, enable the pathogen to survive for months or years in soil and plant debris in the absence of a living host.

Epiphytotics. When a disease affects large numbers of individual plants, it is called an epidemic (meaning "on or among people"). A more precise term, however, is epiphytotic ("on plants"). In contrast, endemic (enphytotic) diseases occur to some extent in the same area each year and generally cause little concern.

Epiphytotics affect a high percentage of the host plant population, sometimes over a wide area. They may be mild or destructive and local or regional in occurrence. Epiphytotics result from various combinations of factors, including the right combination of climatic conditions. An epiphytotic may occur, for example, when a pathogen is introduced into some area in which it had not previously existed. Examples of this condition include the downy mildews (*Sclerospora* species) and rusts (*Puccinia* species) of corn in Africa during the 1950s; the introduction of the coffee-rust fungus into Brazil in the 1960s, and the entrance of the chestnut blight (*Endothia parasitica*) into the United States shortly after 1900. In addition, when new plant varieties are produced by plant breeders without regard for all enphytotic diseases that occur in the same area to some extent each year (but which are normally of minor importance), some of these varieties may prove very susceptible to previously unimportant pathogens. Examples of this situation include the development of oat varieties with Victoria parentage, which, although highly resistant to rusts (*Puccinia graminis avenae* and *P. coronata avenae*) and smuts (*Ustilago avenae, U. kolleri*), proved very susceptible to *Helminthosporium* blight (*H. victoriae*), formerly a minor disease of grasses. The destructiveness of this disease resulted in a major shift of oat varieties on 50,000,000 acres in the U.S. in the mid-1940s. Maize (corn) with male-sterile cytoplasm (*i.e.*, plants with tassels that do not extrude anthers or pollen), grown on 60,000,000 acres in the U.S., was attacked in 1970 by a virulent new race of the southern corn-leaf blight fungus (*Helminthosporium maydis* race T), resulting in a loss of about 700,000,000 bushels of corn. In 1971 the new *Helminthosporium* race was widely disseminated and was reported from most continents. Finally, epiphytotics may occur when host plants are cultivated in large acreages where previously little or no land was devoted to that crop.

Factors leading to epiphytotics

Epiphytotics may occur in cycles. When a plant disease first appears in a new area, it may grow rapidly to epiphytotic proportions. In time, the disease wanes, and, unless the host species has been completely wiped out, the disease subsides to a low level of incidence and becomes enphytotic. This balance may change dramatically by conditions that favour a renewed epiphytotic. Among such conditions are weather (primarily temperature and moisture), which may be very favourable for multiplication, spread, and infection by the pathogen; introduction of a new and more susceptible host; development of a very aggressive race of the pathogen; and changes in cultural practices that create a more favourable environment for the pathogen.

Environmental factors affecting disease development. Important environmental factors that may affect development of plant diseases and determine whether they become epiphytotic include temperature, relative humidity, soil moisture, soil acidity (pH), soil type, and soil fertility.

Temperature. Each pathogen has an optimum temperature for growth. In addition, different growth stages of the fungus, such as the production of spores (reproductive units), their germination, and the growth of the mycelium (the filamentous main fungus body), may have slightly different optimum temperatures. Storage temperatures for certain fruits, vegetables, and nursery stock are manipulated to control fungi and bacteria that cause storage decay, provided the temperature does not change the quality of the products. Little, except limited frost protection, can be done to control air temperature in fields, but greenhouse temperatures can be regulated to check disease development.

Knowledge of optimum temperatures, usually combined with optimum moisture conditions, permits forecasting, with a high degree of accuracy, the development of such diseases as blue mold of tobacco (*Peronospora tabacina*), downy mildews of vine crops (*Pseudoperonospora cubensis*) and lima beans (*Phytophthora phaseoli*), late blight of potato and tomato (*Phytophthora infestans*), leaf spot of sugar beets (*Cercospora beticola*), and leaf rust of wheat (*Puccinia recondita tritici*). Effects of temperature may mask symptoms of certain virus and mycoplasma diseases, however, making them more difficult to detect.

Disease forecasting

Relative humidity. Relative humidity is very critical in fungal spore germination and the development of storage rots. *Rhizopus* soft rot of sweet potato (*Rhizopus stolonifer*) is an example of a storage disease that does not develop if relative humidity is maintained at 85 to 90 percent, even if the storage temperature is optimum for growth of the pathogen. Under these conditions, the sweet-potato root produces suberized (corky) tissues that wall off the *Rhizopus* fungus.

High humidity favours development of the great majority of leaf and fruit diseases caused by fungi and bacteria. Moisture is generally needed for fungal spore germination, the multiplication and penetration of bacteria, and the initiation of infection. Germination of powdery mildew spores, however, is inhibited by water; germination occurs best at 90 to 95 percent relative humidity. Diseases in greenhouse crops—such as leaf mold of tomato (*Cladosporium fulvum*) and decay of flowers, leaves, stems, and seedlings of flowering plants, caused by *Botrytis* species—are controlled by lowering air humidity or by avoiding spraying plants with water. The combination of heating and circulating greenhouse air usually lowers the humidity enough to prevent these fungi from causing infection.

Low humidity also may be destructive to plants. Leaf scorch, a nonparasitic disease, is common on trees in exposed locations following hot, dry, windy weather when water is lost from leaves faster than it is absorbed by roots. Leaf scorch and sudden flower drop are common indoor plant problems because the humidity in a home, apartment, or office is usually below about 30 percent.

Soil moisture. High or low soil moisture may be a limiting factor in the development of certain root-rot diseases. High soil-moisture levels favour development of destructive water mold fungi, such as species of *Aphanomyces*, *Pythium*, and *Phytophthora*. Excessive watering of house plants is a common problem. Overwatering, by decreasing oxygen and raising carbon dioxide levels in the soil, makes roots more susceptible to root-rotting organisms.

Diseases, such as take-all of cereals (*Ophiobolus graminis*); charcoal rot of corn, sorghum, and soybean (*Macrophomina phaseoli*); common scab of potato (*Streptomyces scabies*); and onion white rot (*Sclerotium cepivorum*) are most severe under low soil-moisture levels.

Soil pH. Soil pH, a measure of acidity or alkalinity, markedly influences a few diseases, such as common scab of potato and clubroot of crucifers (*Plasmodiophora brassicae*). Growth of the potato-scab organism is suppressed at a pH of 5.2 or slightly below (pH 7 is neutral; numbers below 7 indicate acidity and those above 7 indicate alkalinity). Scab is not normally a problem when the natural soil pH is around 5.2. Some farmers add sulfur to their potato soil to keep the pH around 5.0. Clubroot of crucifers (members of the mustard family), on the other hand, can usually be controlled by thoroughly mixing lime into soil until the pH becomes 7.2 or higher.

Controlling acidity of soils

Soil type. Certain pathogens are favoured by loam soils and others by clay soils. *Phymatotrichum* root rot attacks cotton and some 2,000 other plants in the southwestern United States. This fungus is serious only in black alkaline soils—pH 7.3 or above—that are low in

organic matter. *Fusarium* wilt disease, which attacks a wide range of cultivated plants, causes more damage in lighter and higher (topographically) soils. Nematodes are also most damaging in lighter soils that warm up quickly.

Soil fertility. Greenhouse and field experiments have shown that raising or lowering the levels of certain nutrient elements required by plants frequently influences the development of some infectious diseases—for example, fire blight of apples and pears; stalk rots of corn and sorghum; *Botrytis* blights; *Septoria* diseases; powdery mildew of wheat; and northern leaf blight of corn. These diseases and many others are more destructive after application of excessive amounts of nitrogen fertilizer. This condition can often be counteracted by adding adequate amounts of potash, a fertilizer containing potassium.

A superabundance of nitrogen may cause deficiency symptoms of potassium, zinc, or other nutrient elements; a lack of or delay in flower and fruit development; and a predisposition to winter injury. If potassium is high, calcium and magnesium deficiencies may occur. Deficiencies or excesses of phosphorus, iron, zinc, manganese, boron, molybdenum, sulfur, copper, and several other elements are known to cause noninfectious diseases.

Ingredients needed for disease development. Infectious disease cannot develop if any one of the following three basic ingredients is lacking: (1) the proper environment, the most important environmental factors being the amount and frequency of rains or heavy dews, the relative humidity, and the air and soil temperature; (2) the presence of a virulent pathogen; and (3) a susceptible host. Effective disease control measures are aimed at breaking this environment–pathogen–host triangle. Loss resulting from disease is reduced, for example, if the host can be made more resistant or immune through such techniques as plant breeding. In addition, the environment can be made less favourable for invasion by the pathogen and more favourable for the growth of the host plant. Finally, the pathogen, or disease agent, can be killed or prevented from reaching the host. These basic methods of control can be divided into a number of cultural and chemical practices to help keep disease in check.

The environment–pathogen–host triangle

DIAGNOSIS OF PLANT DISEASES

Rapid and accurate diagnosis of disease is necessary before proper control measures can be suggested. It is the first step in the study of any disease. Diagnosis is largely based on characteristic symptoms expressed by the diseased plant. Identification of the pathogen (by "signs," see Table 2) is also essential to diagnosis.

Three steps involved in diagnosis include careful observation and classification of the facts, evaluation of the facts, and a logical decision as to the cause.

Variable factors affecting diagnosis. A skilled diagnostician must know what is normal for an affected plant species; its local air and soil environment; the cultural conditions under which it is growing; and the pathogens described for the area; and the disease-developing potential of the pathogen. Diagnosis is best done in the presence of the growing plant. Disease is suspected when, for example, part or all of a plant starts to die. Disease is also indicated when blossoms, leaves, stems, roots, or other plant parts are abnormal—*i.e.*, misshapen, curled, discoloured, overdeveloped, or underdeveloped. Diseased plants also often fail to respond normally to fertilizing, watering, pruning, insect and mite control, or other recommended practices. Finally, if the plant or variety declines in vigour and productivity after performing well for years, this may be an indication of disease.

Other factors or agencies, however, may produce similar or identical symptoms. Some of these have been described; numerous others that exist must be considered when plants are adversely affected.

For example, an affected plant may not be adapted to the area in which it is growing. It may not be able to withstand the extremes in soil moisture, temperature, wind, light, or humidity of the local situation. Damage to plants may be caused by insects, mites, rodents, pets, or humans. The soil may be poorly drained, gravelly, or overly sandy; it may be covering buried debris—boards, cement blocks, bricks, and mortar; or it may be too dry or otherwise unfavourable for good plant growth. Problems are also caused by high winds, hail, lightning, blowing sand, a heavy load of snow or ice, flooding, fire, ice-removal chemicals, mechanical injury by garden tools or machinery, and fumes from weed-killing chemicals, trash burners, nearby industrial plants, or car exhaust. The affected plant may have received different treatment from nearby healthy ones—watering, fertilizing, chemical pest control, pruning, or depth of planting are examples. If different species or kinds of plants in the same area have similar symptoms, the chances are that plant disease is not involved. Most infectious diseases are highly specific for individual or closely related plant species and similar symptoms on unrelated plants are usually an indication of some environmental factor rather than a disease-causing organism.

Conditions that produce symptoms similar to disease

Examination of leaves is usually considered to be the best starting point in diagnosis. Leaves are often the first part to show symptoms and are normally the easiest to examine. Localized spotting or blighting, for example, which may be caused by a wide range of fungi and some bacteria, commonly appears one to two weeks after periods of rainy, cloudy weather. The colour, size, shape, and margins of spots and blights (lesions) are often associated with a particular fungus or bacterium. Many fungi produce "signs" of disease, such as mold growth or fruiting bodies that appear as dark specks in the dead area. Early stages of bacterial infections that develop on leaves or fruits during humid weather often appear as dark and water-soaked spots with a distinct margin and sometimes a halo; a lighter coloured ring around the spot.

Low winter temperatures and late spring or early fall freezes cause blasting (sudden death) of leaf and flower buds or sudden blighting (discoloration and death) of tender foliage.

Insect-injured leaves usually show evidence of feeding, such as holes, discoloration, stippling, blotching, downward curling, or other deformations.

Scorching of leaf margins and between the veins is common following hot, dry, windy weather. Similar symptoms are produced by an excess of water, imbalance of essential plant elements, an excess of soluble salts, changes in the soil water table or soil grade, gas or fume injury, and root injury or disease.

Virus diseases, such as mosaics and yellows, are sometimes confused with injury by a hormone-type weed-killer, unbalanced nutrition, and soil that is excessively alkaline or acid. Nearby plant species are often examined to see if similar symptoms are evident on several different types of plants.

Examination of stems, shoots, branches, and trunk follows a thorough leaf examination. Sunken, swollen, or discoloured areas in the fleshy stem or bark may indicate canker infection by a fungus or bacterium or injury caused by excessively high or low temperatures, hail, tools, equipment, vehicles, or girdling wires.

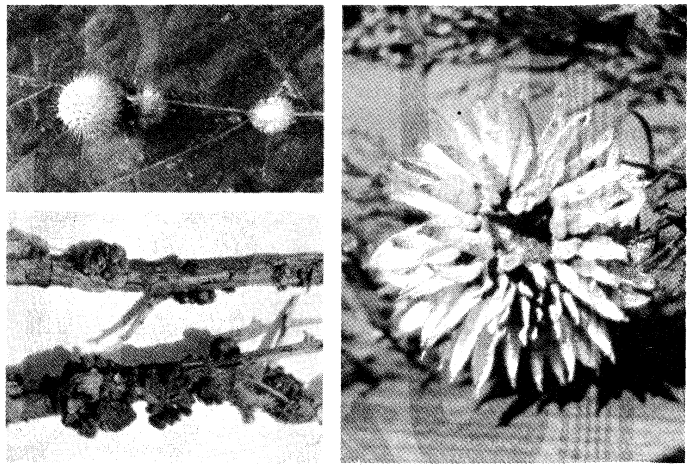
Diagnosis of stem system

Fungal mold and fruiting bodies in or on such areas often indicate secondary infection. Accurate identification of signs as belonging to a pathogenic organism or a secondary or saprophytic one is difficult. Tissues directly infected by pathogenic fungi or bacteria normally show a gradual change in colour or consistence. Injuries, in comparison, are usually well defined with an abrupt change from healthy to affected tissue.

Holes and sawdust-like debris are evidence of boring insects that usually invade woody plants in a low state of vigour. Other borer signs include wilting and dieback. These symptoms are also produced by fungi and bacteria that invade water- and food-conducting vascular tissue.

Symptoms of wilt-inducing micro-organisms include dark streaks in sapwood of wilted branches when the wood is cut through at an angle.

Abnormal suckers or water sprouts on trees can indicate careless pruning, extremes in temperature or water supply, structural injury, or disease.



Three causes of galls.
Galls, localized swellings usually affecting only part of a plant, are symptoms of various types of diseases. (Top left) Hedgehog galls on oak leaf are caused by insect damage to vein of leaf. (Right) Cedar apple gall, natural-looking and somewhat attractive, is actually result of fungal infection. (Bottom left) Cane galls on raspberry stems are produced by bacterial infection.
By courtesy of (top left, right) the Morton Arboretum, Lisle, Illinois, (bottom left) Malcolm C. Shurtleff, University of Illinois, Urbana

Galls, which are unsightly overgrowths on stem, branch, or trunk, may indicate crown gall, insect injury, water imbalance between plant and soil, or other factors. Crown gall is infectious and develops as rough, roundish galls at wounds resulting from grafting, pruning, or cultivating.

Wood-decay fungi also enter unprotected wounds resulting in discoloured, water-soaked, spongy, stringy, crumbly, or hard rots of living and dead wood. External evidence of wood decay fungi are clusters of mushrooms (or toadstools) and hoof- or shelf-shaped fungal fruiting structures, called conks, punks, or brackets.

Aboveground symptoms of many root problems look alike. They include stunting of leaf and twig growth, poor foliage colour, gradual or sudden decline in vigour and productivity, shoot wilting and dieback, and even rapid death of the plant. The causes include infectious root and crown rot; nematode, insect or rodent feeding; low temperature or lightning injury; household gas injury; poor soil type or drainage; change in soil grade; or massive removal of roots in digging utility trenches and in street and sidewalk construction.

Abnormal root growth is revealed by comparison with healthy roots. Nematodes, including root knot, produce small to large galls in roots; sometimes affected roots become discoloured, stubby, excessively branched, and decayed. Bacterial and fungal root rots commonly follow feeding by nematodes, insects, and rodents. Such roots are often discoloured and decayed (water-soaked, mushy, spongy, or firm) and may be covered with mold growth.

Diagnosis of a disease complex, one with two or more causes, is usually difficult and requires separation and identification of the individual causes.

Symptoms. The variety of symptoms, the internal and external expressions of disease, that result from any disease form the symptom complex, which, together with the accompanying signs, makes up the syndrome of the disease.

Classification of symptoms

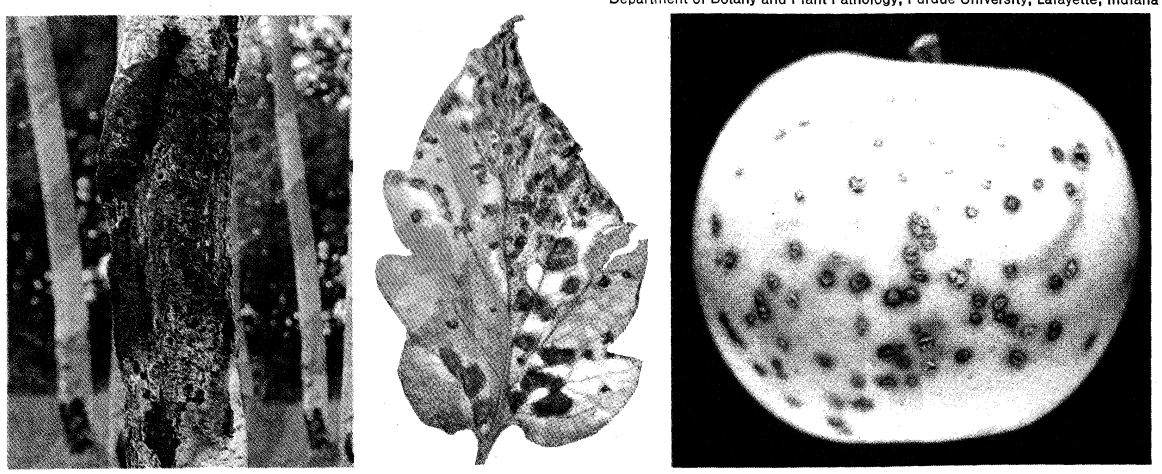
Generalized symptoms may be classified as local or systemic, primary or secondary, and microscopic or macroscopic. Local symptoms are physiological or structural changes within a limited area of host tissue, such as leaf spots, galls, and cankers. Systemic symptoms are those involving the reaction of a greater part or all of the plant, such as wilting, yellowing, and dwarfing. Primary symptoms are the direct result of pathogen activity on invaded tissues (*e.g.*, swollen "clubs" in clubroot of cabbage and "galls" formed by feeding of the root-knot nematode). Secondary symptoms result from the physiological effects of disease on distant tissues and uninvaded organs (*e.g.*, wilting and drooping of cabbage leaves in hot weather resulting from clubroot or root knot). Microscopic disease symptoms are expressions of disease in cell structure or cell arrangement seen under a microscope. Macroscopic symptoms are expressions of disease that can be seen with the unaided eye. Specific macroscopic symptoms are classified under one of four major categories: prenecrotic, necrotic, hypoplastic, and hyperplastic, or hypertrophic. These categories reflect abnormal effects on host cells, tissues, and organs that can be seen without a hand lens or microscope. See Table 1 for examples of the main disease symptoms that are classified in these four categories.

Signs. Besides symptoms, the diagnostician recognizes signs characteristic of specific diseases. Signs are either structures formed by the pathogen or the result of interaction between pathogen and host—*e.g.*, ooze of fire blight bacteria; slime flux from wetwood of elm; odour of tissues affected with bacterial soft rot. See Table 2 for the most frequently encountered signs of pathogen presence and examples of organisms producing them.

II. Classification and control

Plant diseases are often classified by their physiological effects or symptoms. Many, however, produce practically identical symptoms and signs yet are caused by very different micro-organisms or agents, thus requiring com-

By courtesy of (left) Malcolm C. Shurtleff, University of Illinois, Urbana, (centre, right) Department of Botany and Plant Pathology, Purdue University, Lafayette, Indiana



Symptoms of plant diseases.
(Left) Hypoxylon canker, extensive destruction of tissue seen on quaking aspen tree, is a symptom of fungal infection. (Centre) Variation from normal colour is a symptom of disease caused by many different agents, as in this early blight fungal infection of tomato leaf. (Right) Blister spots, shown on apple, are symptomatic of bacterial infection.

Table 1: Plant Disease Symptoms

symptom name	description and causes	examples
Pre-necrotic symptoms	symptom expression that precedes death of cells or their disintegration of tissues	
Water-soaking	a water-soaked, translucent condition of tissues caused by water moving from host cells into intercellular spaces	late-blight lesions on potato and tomato leaves; bacterial soft rot of fleshy vegetables
Wilting	temporary or permanent drooping of leaves, shoots, or entire plants from lack of water	bacterial wilt of cucumber; *Fusarium* wilt of tomato
Abnormal coloration	yellowing, reddening, bronzing, or purpling in localized areas of leaves where chlorophyll has been destroyed; may be due to a variety of causes	cabbage and aster yellows; halo blight of beans; potassium or phosphorus deficiency
	the presence of two or more colours in leaves and flowers due to a genetic abnormality is called variegation; viral infection results in "flower breaking"	tulip mosaic
Necrotic symptoms	localized or general death and disintegration of cells and/or tissues	
Blast	sudden blighting or death of young buds, flowers, or young fruit; also, a failure to produce fruit or seeds	*Botrytis* blight of peony buds; oat blast
Bleeding	flow of sap, often discoloured, from a split crotch, branch stub, or other wound; usually accompanied by an odour of fermentation	bleeding canker of beech, dogwood, and maple
Blight	sudden or total discoloration and killing of large numbers of blossoms, leaves, shoots, limbs, or the entire plant; usually young tissues are attacked; the disease name is often coupled with the name of host and part attacked—blossom blight, twig blight, tip blight	fire blight of pome fruits; *Diplodia* or *Sphaeropsis* tip blight of conifers
Canker	a definite, dead, often sunken or swollen and cracked area on a stem, limb, trunk, tuber, or root surrounded by living tissues	anthracnose of sycamore and brambles; *Nectria* canker of hardwoods; fire blight of pome fruits
Damping-off	decay of seed in soil, rapid death of germinating seedlings before emergence, or when emerged seedlings suddenly wilt, topple over, and die from rot at or near the soil line	pre-emergence damping-off and post-emergence damping-off; both are common in seedbeds
Dieback	progressive browning and death of shoots, branches, and roots starting at the tips	winter injury; wet soil; excess soil nutrients; girdling cankers; stem or root rots; nematodes
Firing	drying and dying of leaves	nitrogen or potassium deficiency in corn; *Verticillium* wilt of eggplant
Fleck	a small, white to translucent spot or lesion visible through a leaf	ozone injury to many plants; necrotic fleck of lily
Mummification	final stage in certain fruit rots, in which the dried, shrivelled, and wrinkled fruit is called a "mummy"	brown rot of stone fruits; black rot of apple
Net necrosis	an irregular criss-crossing of dark brown to black lines giving a netted appearance	in potato tubers of plants with virus leaf roll
Pitting	small dead areas within fleshy or woody tissues that appear healthy externally; definite sunken grooves or pits are formed	virus stem-pitting in apple and peach trunks; stony pit of pear fruit
Rot	decomposition and putrefaction of cells; later of tissues and organs; the rot may be dry, firm, watery, or mushy and characterized by such names as hard rot, soft rot, dry rot, black rot, and white rot	bacterial soft rot; berry rot; bud rot; bulb rot
Scald	blanching of young fruit, foliage, and shoot tissue; generally superficial	sunscald; apple and pear scald
Scorch	sudden death and "burning" of large, indefinite areas in leaves and fruit	toxicity from pesticides and air pollutants; drought; wind; lack or excess of some nutrient
Shot hole	dead spotting of leaves with diseased tissue dropping out leaving small holes	bacterial spot; *Coryneum* blight of peach
Spot	a definite, localized, round to regular lesion often with a border of a different colour; characterized as to location (leaf spot, fruit spot) and colour (brown spot, black spot); if numerous or if spots enlarge and merge a large irregular blotch or blight may develop	gray leaf spot of tomato; black spot of rose; tar spot of maple
Staghead	an advanced form of dieback applied to a tree in which large branches in the upper crown are killed	oak wilt on bur oak; dwarf mistletoe on Douglas fir; *Armillaria* root rot of oak
Streak	narrow, elongated, somewhat superficial necrotic lesions, with irregular margins, on stems or leaf veins	virus streak of pea, raspberry, and tomato; Stewart's wilt of sweet corn
Stripe	narrow, elongated, parallel, necrotic lesions especially in leaf diseases of cereals and grasses	*Helminthosporium* stripe of barley; *Scolecotrichum* brown stripe of forage grasses
Hypoplastic	the underdevelopment of plant cells, tissues, or organs	
Abortion	halting development of an organ after partial differentiation	ergot of rye and other grasses
Chlorosis	yellowing or whitening of normal green tissue due to partial or complete failure of chlorophyll to develop	strawberry and aster yellows; genetic variegation in corn; iron deficiency of azalea
Stunting or dwarfing	the underdevelopment of the plant or some of its organs	dahlia stunt or mosaic; curly top of beans; little-leaf disease of pines
Rosetting	shortening of internodes of shoots and branches, producing a bunchy growth habit	peach and lily rosette
Hyperplastic or hypertrophic	an overdevelopment or overgrowth of plant cells, tissues, or organs; hyperplastic has come to mean an increase in number of cells; hypertrophic, an increase in cell size	
Abscission or cast	early dropping of leaves, flowers, or small fruits; usually associated with premature formation of an abscission (separation) cell layer	black spot of rose; early blight of tomato; apple scab
Callus	overgrowth of tissues, often at margins of a canker or wound	*Nectria* canker of hardwoods; stem pitting of peach
Curl	distortion and crinkling of leaves or shoots resulting from unequal cell growth of opposite sides or in certain tissues	tobacco and tomato mosaic; leaf roll of potato; peach leaf curl
Epinasty	downward or outward curling and bending of a leaf or petiole	2,4-D injury to broadleaf plants; *Fusarium* wilt of tomato
Fasciation, or witches'-broom	a distortion that results in a dense, bushy overgrowth of thin, flattened, and sometimes curved shoots, flowers, fruit, and roots at a common point; usually due to adventitious (abnormally located) development of organs	witches'-broom of hackberry; hairy root of apple; leaf gall or fasciation of geranium (see also *Rosetting* under Hypoplastic symptoms in this table)
Metamorphosis or transformation	development of more or less normal tissues or organs in an abnormal location	crazy-top of corn and sorghum; formation of aerial potato tubers
Proliferation	continued development of an organ after it would normally stop growing	adventitious shoots in China aster and chrysanthemum from aster yellows mycoplasma
Russeting	usually a brownish, superficial roughening or corking of the epidermis of leaves, fruit, tubers, or other organs; often due to suberization (cork development) of cells following injury	spray or weather injury to apples; sweet potato scurf
Scab	roughened to crustlike, more or less circular, slightly raised or sunken lesions on the surface of leaves, stems, fruit, or tubers	apple, peach, and cucumber scab; common scab of potato
Gall, knot, or tumefaction	formation of local, fleshy to woody outgrowths or swellings; the outgrowth is often composed of unorganized cells	crown gall; black knot of plum; *Fusiform* gall rust of pine; nematode galls

Table 2: Signs of Pathogen Presence in Diseased Plants*

signs	description	examples
Acervulus	a shallow, saucer-shaped fungus structure that bears asexual spores (conidia); it is usually formed below the cuticle or epidermis of leaves, stems, and fruits later rupturing the surface and exposing its spore-bearing surface	anthracnose of muskmelon and tomato; *Marssonina* leaf spot and twig blight of poplar
Apothecium	a disk-, saucer-, or cup-shaped fungus structure that produces sexual spores (ascospores); it is often stalked and fleshy	brown rot of stone fruits; *Sclerotinia* white mold of fleshy vegetables
Cleistothecium	a speck-sized, black fruiting body completely enclosing sexual spores	many powdery mildew fungi
Conidiophores	asexual fungus structures of various colours that bear conidia and appear powdery, velvety, or downy en masse; often cover lesions on leaf, stem, or fruit	*Botrytis* blight or gray mold of many flowers; *Penicillium* mold of citrus fruit; downy mildew of grape
Conk or punk	fruiting bodies (sporophores) of wood-rotting fungi that produce tremendous numbers of spores (up to 100 billion per day); conks are usually large and woody and are found on tree stumps, branches, or trunks	*Fomes* and *Polyporus* wood rots of hardwoods and conifers
Mushroom (toadstool)	fleshy, umbrella-shaped fruiting bodies of wood-decay fungi	*Armillaria* and *Clitocybe* root rots
Mycelium	the vegetative body of a fungus, which is composed of a mass of branched filaments (hyphae) often interwoven into a feltlike or woolly mass	*Rhizopus* soft rot of sweet potato and leak of strawberry; *Sclerotinia* white mold of beans
Nematode cysts	round to lemon-shaped, speck-sized bodies, white to brown in colour, are diagnostic for cyst nematodes; they are often evident on the root surface	sugar beet, soybean, and clover cyst nematodes
Odours	many pathogens or the process of host colonization give off characteristic odours	bacterial soft rot; stinking smut or bunt of wheat; slime flux of elm
Ooze or exudate	droplets of bacteria or fungus spores, usually mixed with host-cell decomposition products, found on surfaces of lesions	ooze from fire blight; scab on cucumber fruit; cut stem of cucumber affected with bacterial wilt
Perithecium	speck-sized fungus fruiting body that produces large numbers of sexual spores; perithecia are dark coloured, round to flask-shaped, usually partially buried in diseased tissue; they resemble pycnidia	apple and pear scab; *Gibberella* stalk and ear rot of corn
Powdery mildew	white, powdery to mealy, superficial growths of mycelia and conidiophores on surfaces of leaves, stems, flowers, and fruit	powdery mildew diseases of bluegrass, phlox, zinnia, and rose (see also *Cleistothecium*, this table)
Pycnidium	speck-sized fungus fruiting body that produces large numbers of asexual spores (conidia); pycnidia are dark coloured, round to flask-shaped, usually partially buried in diseased tissue; they resemble perithecia	*Septoria* leaf spots; *Diplodia* stalk rot of corn
Rhizomorph	cordlike or rootlike strands, composed of a bundle of closely intertwined hyphae, by which certain fungi make their way through soil and over or under bark of woody plants	*Armillaria* and *Clitocybe* root rots; *Sclerotium rolfsii* stem rot of peanuts
Sclerotium	brown to black, compact, hard resting bodies of certain fungi with a rindlike covering; their size varies from a fly speck to a large sweet potato depending on the fungus forming it	ergot of rye; onion white rot; *Verticillium albo-atrum*
Seeds	dodder seed is a sign of this parasitic flowering plant when found in clover or alfalfa seed	dodder (*Cuscuta*, about 170 species)
Sorus (pustule)	a compact mass of spores, or a cluster of sporangia (spore bearing structures), produced in or on the host by fungi causing such diseases as white rust, smut, and true rust; before rupturing, the sorus is normally covered by host epidermis	white rust of crucifers; corn and bluegrass smuts; black stem rust of cereals
Spores	microscopic, usually single- or few-celled reproductive bodies of fungi corresponding in function to seeds of higher plants; spores vary greatly in size, shape, and colour; they are asexually produced, or result from sexual processes; asexual spores may be formed directly from vegetative hyphae but are often produced in special fruiting structures, (*e.g.*, acervulus, coremium, pycnidium, and sporodochium)	
Sporodochium	a cushion-shaped stroma covered with conidiophores bearing asexual spores; found scattered in leaf, stem, and fruit lesions	*Cercospora* leaf spot of celery and sugar beet; brown rot of stone fruits; *Fusarium* blight of bluegrass
Stroma	a crustlike or cushion-like mass of fungus hyphae often intermingled with host tissue on or in which spores are produced—usually in reproductive bodies	tar spot of maple and sycamore
Synnema or coremium	a tight cluster of erect conidiophores forming an elongated column on which asexual spores are borne	Dutch elm disease; oak wilt; black rot of sweet potato

*The structures listed are formed by the pathogen.

pletely different control methods. Classification according to symptoms is also inadequate because a causal agent may induce several different symptoms, even on the same plant organ, which often intergrade. Classification also may be according to the species of plant affected. Host indices (lists of diseases known to occur on certain hosts in regions, countries, or continents) are valuable in diagnosis. When an apparently new disease is found on a known host, a check into the index for the specific host often leads to identification of the causal agent. It is also possible to classify diseases according to the essential process or function that is adversely affected. The best and most widely used classification of plant diseases is based on the causal agent, such as a noninfectious agent, virus, mycoplasma, bacterium, fungus, nematode, or parasitic flowering plant.

Use of the host index

Agents or factors inducing disease in plants may be divided into two broad groups: noninfectious (also called abiotic, physiogenic, or nontransmissible) and infectious (biotic or transmissible).

NONINFECTIOUS DISEASE-CAUSING AGENTS

Plants in poor health because of some factor in the environment probably outnumber plants attacked by pathogens. Noninfectious diseases, which sometimes arise very suddenly, are caused by the excess, deficiency, nonavailability, or improper balance of light, air circulation, relative humidity, water, or essential soil elements; unfavourable soil moisture–oxygen relations; extremes in soil acidity or alkalinity; high or low temperatures; pesticide injury, other poisonous chemicals in air or soil; soil-grade changes; girdling of roots; mechanical and electrical agents; and soil compaction. In addition, unfavourable preharvest and storage conditions for fruits, vegetables, and nursery stock often result in losses. The effects of noninfectious diseases can often be seen on a variety of plant species growing in a given locality or environment. Many diseases and injuries caused by noninfectious agents result in heavy loss but are difficult to check or eliminate because they often reflect ecological factors beyond man's control. Symptoms may appear several weeks or months after an environmental disturbance.

Injuries from accidents, poisons, or adverse environmental disturbances often result in injured tissues that weaken a plant, enabling bacteria, fungi, or viruses to enter and add further damage. The cause may be obvious (lightning or hail); often it is obscure. Symptoms alone are often unreliable in identifying the causal factor.

Adverse environment. High temperatures may scald corn, cotton, and bean leaves and may induce formation of cankers at the soil surface of tender flax, cotton, and peanut plants. Frost injury is relatively common, but temperatures just above freezing may also cause damage, such as net necrosis in potato tubers and "silvering" of corn leaves. Isolated, thin-barked trees growing in northern climates, subjected to frequent thawing by day and freezing by night, may develop dead bark cankers or vertical frost cracks on the south or southwest sides of the trunk. Alternate freezing and thawing, heaving, low air moisture, and smothering due to an ice-sheet cover are damaging to alfalfa, clovers, strawberries, and grass on golf greens. Legume (members of the pea family) crowns commonly split under these conditions and are invaded by decay-forming fungi.

Low-temperature injury

The drought and dry winds that often accompany high temperatures cause stunting, wilting, blasting, marginal scorching of leaves, and dieback of shoots. Similar symptoms are caused by a change in soil grade; an altered water-table level; a compacted and shallow soil; paved surface over tree roots; temporary flooding or a waterlogged (oxygen-deficient) soil; girdling tree roots; salt spray near the ocean; and an injured or diseased root system. Injured plants are often very susceptible to air and soil pathogens and secondary invaders.

Blossom-end rot of tomato and pepper is prevalent when soil moisture and temperature levels fluctuate widely and calcium is low.

Poor aeration may cause blackheart in stored potatoes. Accumulation of certain gases from respiration of apples in storage may produce apple scald and other disorders.

A deficiency, excess, or imbalance of essential elements in soil often results in symptoms that vary depending on stage of plant growth, soil moisture, and other factors. Symptoms include stunting of plants; scorching or malformation of leaves; abnormal coloration; premature leaf, bud, and flower drop; delayed maturity or failure of flower and fruit buds to develop; and dieback of shoots. Nutrient deficiency or excess symptoms are rare if temperature and soil moisture is adequate, the soil contains an ample and balanced supply of available mineral elements, and soil reaction is favourable. A pH of 5.5 to 7 is best for most crop, garden, and house plants.

Symptoms resulting from nutrient element imbalance

An excess of water-soluble salts is a common problem with house plants. Salt concentrations may build up as a whitish crust on soil and container surfaces of potted plants following normal evaporation of water over a period of time. Symptoms include leaf scorching, bronzing, yellowing and stunting, wilting, plus root and shoot dieback. Damage from soluble salts is also common in arid regions and in regions in which ice-control chemicals are applied in large amounts.

Several nonparasitic diseases (*e.g.*, oat blast, weakneck of sorghum, straighthead of rice, and crazy-top of cotton) are caused by combinations of environmental factors—*e.g.*, high temperatures, moisture stress and/or poor irrigation practices, imbalance of mineral nutrients, and reduced light.

Environmental disturbances alter the normal physiology of the plant, activity of pathogens, and host–pathogen interactions.

Toxic chemicals. Many complex chemicals are routinely applied to plants alone or in combination to prevent attack by insects, mites, and pathogens; to kill weeds; or to control growth. Serious damage may result when fertilizers, herbicides, fumigants, growth regulators, antidesiccants, insecticides, miticides, fungicides, nematicides, and surfactants (substances with enhanced wetting, dispersing, or cleansing properties, such as detergents) are applied at excessive rates or under hot, cold, or slow-drying conditions.

Some 200,000,000 tons of air pollutants are released over the U.S. each year; other industrialized and densely populated countries have similar problems. In fact, plant injury is increasing all over the world near urban centres, refuse dumps, major highway systems, factories, and power plants. The major plant-toxic pollutants are sulfur dioxide, fluorine, ozone, and peroxyacetyl nitrate.

Sulfur dioxide results largely from the burning of large amounts of soft coal and high-sulfur oil. It is toxic to a wide range of plants at concentrations as low as 0.25 parts per million (ppm) of air (*i.e.*, on a volume basis, a part per million represents one volume of pure gaseous toxic substance mixed in 1,000,000 volumes of air) for eight to 24 hours. Gaseous and particulate fluorides are more toxic to sensitive plants than is sulfur dioxide because they are accumulated by leaves. They are also toxic to animals that feed on such foliage. Fluorine injury is common near metal-ore smelters, refineries, and industries making fertilizers, ceramics, aluminum, glass, and bricks.

Ozone and peroxyacetyl nitrate injury (also called oxidant injury) have become more prevalent in and near cities with heavy traffic problems. Exhaust gases from internal combustion engines contain large amounts of hydrocarbons (substances containing principally carbon and hydrogen molecules—gasoline, for example). Smaller amounts of unconsumed hydrocarbons are formed by combustion of fossil fuels (coal, oil, natural gas) and refuse burning. Ozone, peroxyacetyl nitrate, and other oxidizing chemicals (smog) are formed when sunlight reacts with nitrogen oxides and hydrocarbons. This pollutant complex is damaging to susceptible plants many miles from its source. Ozone and peroxyacetyl nitrate are capable of causing injury if present at levels of 0.01 to 0.05 parts per million for several hours.

Photochemical air pollutants

Physical injury. Lightning, hail, high winds, ice and snow loads, machinery, insect and animal feeding, and various cultural practices may seriously injure plants or plant products. With the exception of lightning, which may cause death of trees and succulent crop plants in limited areas, such injury does not usually kill plants. Wounds are created, however, through which pathogens may enter.

INFECTIOUS DISEASE-CAUSING AGENTS

Some parasitic micro-organisms secure their nutrition from living plants. If these organisms cause injury or disease, they are called pathogens. An obligate parasite requires living tissue as a food source; a facultative parasite may grow on living or nonliving material.

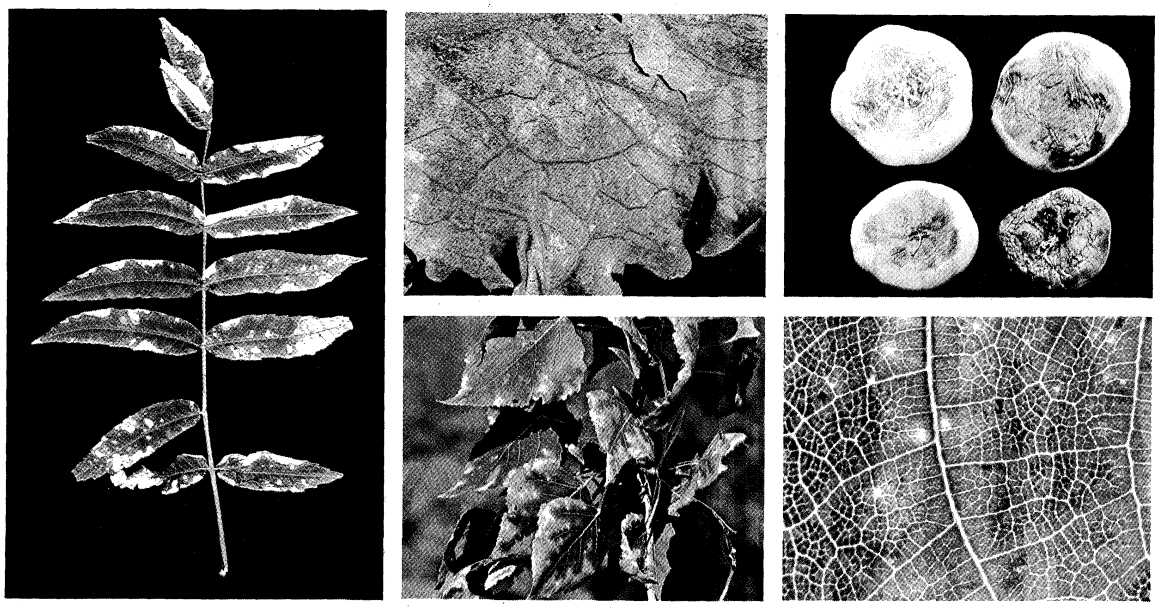
Viruses. Plant viruses are extremely small infectious agents that pass through fine filters and are visible only with the electron microscope. Viruses may occur naturally as complexes or strains differing in mode of transmission, virulence, host plants attacked, symptoms produced, and physical and chemical characteristics (see VIRUS). One plant species may be susceptible to 20 to 40 different viruses.

Plant viruses cause many major diseases of important food crops such as potato, tomato, wheat, oats, rice, corn, peach, orange, sugar beet, sugarcane, and palms. Virus diseases are generally most serious in plants that are vegetatively propagated (*i.e.*, grown from cuttings, cut divisions, sprouts, etc., rather than from seeds; examples include potato, sweet potato, carnation, chrysanthemum, dahlia, geranium, gladiolus, lilies, orchids, tree and cane or bush fruits, and strawberries).

Virus-induced disease symptoms. Viruses induce a wide range of plant symptoms. Rarely they may kill the host in a short time (*e.g.*, spotted wilt and curly top of tomato). More commonly, they cause reduced yield and lower quality of product. Virus symptoms fall into four groups: (1) change in colour—yellowing, green and yellow mottling, and vein clearing; (2) malformations—distortion of leaves and flowers, rosetting, proliferation and witches'-brooms, and little or no leaf development between the veins; (3) necrosis—leaf spots, ring spots, streaks, wilting or drooping, and internal death, especially of phloem (food-conducting) tissue; and (4) stunting or dwarfing of leaves, stems, or entire plants.

Virus-infected plants are often more susceptible to root rots, stem or stalk rots, seedling blights, and possibly other types of diseases.

Some plants may carry one or more viruses and show no symptoms; thus, they are latent carriers and a source of infection for other plants. Symptoms of certain virus-



Non-pathogenic injuries.
(Left) Sumac damaged by sulfur dioxide. (Top centre) Lettuce leaf damaged by peroxyacetyl nitrate, or PAN. (Top right) Blossom-end rot of tomatoes produced by unsuitable weather conditions that result in flooding of tissues with liquid. (Bottom centre) Fluoride injury to poplar leaves. (Bottom right) English walnut leaf damaged by ozone. Except for blossom-end rot, all of these injuries are the result of air pollution.
(Left, top right, bottom centre) Malcolm C. Shurtleff, University of Illinois, Urbana, (top centre, bottom right) F.K. Anderson—EB Inc.

infected plants, such as geraniums, may be masked at high temperatures. Virus symptoms reappear when the weather cools.

For convenience, virus diseases are often grouped together generally by symptoms, regardless of true virus relationships. Viruses can also be grouped into strains, each differing greatly in virulence (ability to cause disease) and other properties. For example, two virus strains, chemically distinct, may produce indistinguishable symptoms in one orchid plant but strikingly different symptoms in another. Diseases caused by unrelated viruses may resemble one another more closely than diseases caused by strains of the same virus. Certain variegated plants, such as *Abutilon* and Rembrandt tulips, owe their horticultural uniqueness and desirability to being inherently virus infected.

Virus strains

Virus transmission. Some viruses are quite infectious and spread easily from diseased plants by contact. With the exception of tobacco-mosaic virus, however, relatively few viruses are spread extensively in the field by contact between diseased and healthy leaves.

All viruses that spread within their host tissues (systemically) can be transmitted by grafting branches or buds from diseased plants on healthy plants. Natural grafting and transmission are possible by root grafts and with dodder (*Cuscuta* species). Vegetative propagation often spreads plant viruses. Fifty to 60 viruses are transmitted in seed, and a few seed-borne viruses, such as sour-cherry yellows, are carried in pollen and transmitted by insects.

Most disease-causing viruses are carried and transmitted naturally by insects and mites, which are called vectors of the virus. The principal virus-carrying insects are about 200 species of aphids, which transmit mostly mosaic viruses, and over 100 species of leafhoppers, which carry yellows-type viruses. Whiteflies, thrips, mealybugs, plant hoppers, grasshoppers, scales, and a few beetles also serve as vectors for certain viruses. Some viruses may persist for weeks or months and even duplicate themselves in their insect vectors; others are carried for less than an hour. Slugs, snails, birds, rabbits, and dogs also transmit a few viruses, but this is not common.

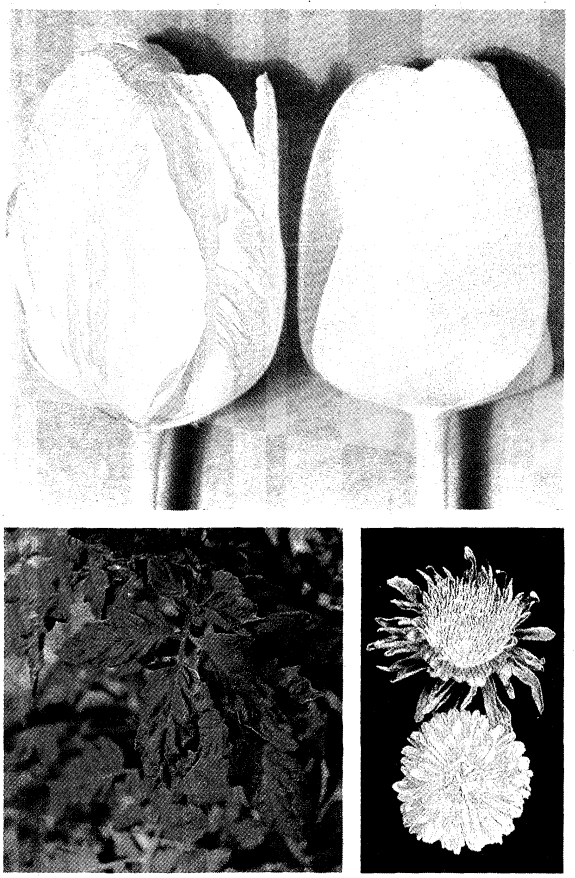
A small number of plant viruses are soil borne. Viruses causing grape fanleaf, tobacco rattle, tobacco and tomato ring spots, as well as several strawberry viruses, are spread by nematodes feeding externally (*i.e.*, ectoparasitic) on plant roots. A few soil-borne viruses may be spread by the swimming spores of primitive, soil-inhabiting pathogenic fungi, such as those causing big vein of lettuce, soil-borne wheat mosaic, and tobacco necrosis.

Viruses often overwinter in biennial and perennial (plants that overwinter by means of roots and produce seed in their second year or during several years, respectively) crops and weeds, in plant debris, and in insect vectors. Plants, once infected, normally remain so for life. Most, if not all, plant viruses can persist in a number of different host plants.

Control of virus diseases. After a plant is infected with a virus, little can be done to restore its health. Control is accomplished by several methods, such as growing resistant species and varieties of plants or obtaining virus-free seed, cuttings, or plants as a result of indexing and certification programs. Indexing is a procedure to determine the presence or absence of viruses not readily transmitted mechanically. Material from a "test" plant is grafted to an "indicator" plant that develops characteristic symptoms if affected by the virus disease in question. In addition, more drastic measures are sometimes followed, including destroying (roguing) infected crop and weed host plants and enforcing state and national quarantines or embargoes. Further control measures include controlling insect vectors by spraying plants with contact insecticides or fumigating soil to kill insects, nematodes, and other possible vectors. Growing valuable plants under fine cheesecloth or wire screening that excludes insect vectors is also done. Separation of new from virus-infected plantings of the same or closely related species is sometimes effective, and the simple practice of not propagating from plants suspected or known to harbour a virus also reduces loss.

Temperature treatments for virus control

Infected peach, apple, and rose budwood stock and carnations have been grown for weeks or months at temperatures around 37° to 38° C (99° to 100° F) to free new growth from viruses. Soaking some woody plant parts or virus-infected sugarcane shoots in hot water at about 50° C (120° F) for short periods is also effective. Both dry- and wet-heat treatments are based on the sensitivity of certain viruses to high temperatures. Rapidly growing dahlia and chrysanthemum sprouts outgrow viruses so that stem tips can be used to propagate healthy plants. With certain carnations, chrysanthemums, and potatoes, a few cells from the growing tip have been grown under sterile conditions in tissue culture; from these, whole plants have been developed free from viruses.



Diseases caused by viruses.
(Top) Clara Butt tulip showing normal coloration and the commercially sought-after streaked colour produced by virus. (Bottom left) Tomato leaves puckered and blistered by tobacco mosaic virus. (Bottom right) Normal China aster and aster showing symptoms of aster yellows caused by mycoplasmalike organism.
(Top) Frank P. McWhorter, (bottom left) Malcolm C. Shurtleff, University of Illinois, Urbana, (bottom right) James E. Kuntz, University of Wisconsin, Madison

Examples of virus-caused diseases. Among the viruses affecting plants, tobacco mosaic virus (TMV) is best known and has great economic importance. Besides affecting many members of the nightshade family (Solanaceae), which includes tobacco, tomato, potato, pepper, eggplant, and petunia, TMV infects more than 70 species of plants in 14 other families. Symptoms characteristically appear as dark-green leaf blotches, irregularly elevated and scattered. The remainder of the leaf becomes more and more pale green and sometimes yellow. The name mosaic derives from these small irregular blotches of different shades of green and yellow.

TMV, unlike most viruses, is resistant to high temperatures (up to 90° C, or 190° F), alcohol, and various germicides; it has been established that TMV retains infectivity in dried tobacco leaves for more than 50 years.

The virus disease curly top caused almost complete abandonment of the sugar-beet industry in parts of the world from 1916 to 1932, when it was controlled by the development of resistant varieties. Intermittently, it still causes severe damage to tomatoes, beans, and other crops. The virus affects a wide range of cultivated and weed hosts in over 70 different plant genera. Young susceptible beet plants turn yellow and usually die quickly when infected by the virus. Symptoms on older sugar beets include a stunting or dwarfing of plants and a thickening, curling, and yellowing of affected leaves; cross sections of roots show dark necrotic areas in the phloem, or food-conducting vascular tissue, of the plant. The virus survives between crops in weed hosts and in the beet leafhopper, the main carrier. Distribution of curly top in sugar-beet fields reflects migration patterns of the beet leafhopper as influenced by the prevalence of other hosts in the vicinity, direction of prevailing winds, and temperature. Judicious use of herbicides, early planting, controlled grazing, and other management practices help reduce the magnitude of the spring leafhopper migration from weedy to cultivated areas, which may sometimes be many miles apart. The key to maintaining high production in sugar beets, however, has been the use of resistant varieties.

Spotted wilt, transmitted by onion and flower thrips, has become widespread and now is responsible for losses to tomato, potato, lettuce, pea, other vegetables, and ornamentals in many parts of the world. Control measures involve sanitation and weed control, spraying to control the thrips vector, and obtaining virus-free plants.

Virus diseases of citrus are a major concern wherever these crops are grown; for example, between 1936 and 1946, tristeza (quick decline) caused the loss of 7,000,000 orange trees in the State of São Paulo, Brazil. Tristeza has attacked or now threatens many millions of trees in various tropical and subtropical areas. The principal controls for citrus virus diseases include costly programs of roguing (destroying) infected trees plus the development of virus-free rootstocks and topworking (grafting) them with budwood from trees known to be free of virus diseases, such as tristeza, exocortis, psorosis, xyloporosis, stubborn disease, tatterleaf, and vein enation.

Control of citrus virus

The worldwide increase in destructiveness of virus diseases and in the number of known plant viruses largely results from the expansion of agricultural enterprises and increased movement of plants and plant products in recent years and from the introduction of single-crop planting in large regions.

Mycoplasmas. Since 1968, around 60 of the "yellows" or "witches-broom-type" diseases previously thought to be virus in origin have been shown to be caused by tiny, unicellular organisms known as mycoplasmas. These are the simplest and smallest known organisms capable of growing on laboratory culture media.

Plant mycoplasmas are characteristically delimited by a distinct double membrane that is flexible. Being soft and plastic, they assume a great variety of sizes and shapes. When viewed in the electron microscope, mycoplasmas are seen to vary from small, round bodies to branched, wormlike forms smaller than bacteria but within the range of large virus particles.

Antibiotic drugs and several sulfa drugs used to treat one type of pneumonia in man caused by a mycoplasma are effective in combatting aster yellows and other diseases caused by mycoplasmas. Plants showing severe symptoms, when treated with an effective antibiotic, produce new, healthy appearing foliage. If test plants are no longer given drugs, symptoms may reappear in several weeks.

Aster yellows infects about 300 species of crop and weed plants in 48 families. The disease is common and destructive in most of the world where air temperatures do not persist much above 32° C (90° F). High temperatures inactivate the mycoplasma in plants and in the leafhopper vectors. Aster yellows can cause serious losses on China aster, daisies, chrysanthemum, lettuce, endive, escarole, carrot, parsnip, celery, parsley, and New Zealand spinach. Diseased asters are stiff, upright, and stunted with numerous yellowish secondary shoots. Flowers are green (virescent) and distorted, dwarfed or lacking. Seventeen species of leafhoppers, depending on the area and strain of the mycoplasma, serve as transmitting agents.

The mycoplasma multiplies both in the leafhopper and in the host plant. Nine to 16 days must elapse after the leafhopper feeds on a diseased plant before it can infect a healthy plant. Transmission does not occur through insect eggs or aster seeds. Control is against leafhopper vectors and involves frequent spraying or dusting with contact insecticides; prompt destruction of infected plants and overwintering weeds that may harbour leafhopper eggs; and growth of asters and other plants commercially under cheesecloth or fine wire screening to exclude insects. Antibiotic sprays or dips appear promising for providing mycoplasma-free cuttings, grafting materials, and seed.

Control of leafhopper transmitted mycoplasmas

Bacteria. Most of the 4,000 described species of bacteria are harmless or even beneficial to man. Over 100 species are known to cause human and animal diseases; however, the almost 200 species that cause plant disease infect over 150 different genera of higher plants in more than 50 families. Bacterial diseases can be grouped into three categories: wilting, necrosis, and overgrowth, or hypertrophy. Wilting results from invasion of the vascular system of the plant by bacteria; examples include bacterial wilts of sweet corn, alfalfa, tobacco, tomato, and cucurbits (squash, pumpkins, cucumbers); and black rot of crucifers. Necrosis is a condition in which plant cells are killed, forming leaf spots, stem blights or cankers, and soft rots; examples are delphinium black spot, fire blight of pome fruits, bacterial blights of beans and peas, and soft rot of iris. Examples of overgrowth, or hypertrophy, include crown gall of many plants, cane gall of brambles, hairy root of apple, and olive knot.

Most bacterial pathogens produce one major symptom; others manifest a range or combination of symptoms. In general, it is difficult to determine the species designation of a bacterium, but it is not particularly difficult to tell that a plant is affected by a bacterial pathogen.

All true bacteria (order Eubacteriales) that cause plant disease are one-celled, rod-shaped organisms that do not form spores. Only five genera of true bacteria cause plant disease. About one half of the species are in the genus *Pseudomonas*. Another important genus, *Xanthomonas*, is characterized by the water-insoluble yellow pigment it forms when grown in the laboratory. In contrast to animal and human pathogens, those that infect plants are usually favoured by a nearly neutral (pH 6.5 to 7.5) medium and low temperatures (20° to 30° C, or 68° to 86° F).

Infection processes. Bacterial pathogens enter plants through wounds, principally produced by adverse weather conditions, man, tools and machinery, insects, and nematodes; or natural openings such as stomates (pores in leaves), lenticels (pores in bark), hydathodes (structures that discharge water from leaf interiors to the surface), nectar-producing glands, and leaf scars.

Most foliage invaders are spread from plant to plant by wind-blown rain or dust. Man disseminates bacteria through cultivation, grafting, pruning, and transporting diseased plant material. Animals, including insects and mites, are other common transmission agents. Some bacteria, such as the causal agent of Stewart's, or bacterial, wilt of corn (*Erwinia stewarti*), are not only spread by a flea beetle but also survive over winter in this insect.

Bacterial dormancy

When conditions are unfavourable for growth and multiplication, bacteria remain dormant on or inside plant tissue. A few, such as the crown gall bacterium, may survive for months or years in the soil.

Bacterial diseases are influenced greatly by temperature and moisture. Often, a difference of only a few degrees in temperature determines whether a bacterial disease will develop. Moisture as a water film on plant surfaces is often essential for entry.

Most pathogenic bacteria of plants are quickly killed by exposure to high temperatures (*e.g.*, ten minutes at 52° C [126° F]), very dry conditions, and strong sunlight. Many pathogenic bacteria in soil are eaten by protozoans, lysed (destroyed) by phages, or killed by extracellular toxic products of other soil-inhabiting organisms.

Control of bacterial diseases. In general, the diseases caused by bacteria are relatively difficult to control. This is partly attributable to the speed of invasion since bacteria enter natural openings or wounds directly. Direct introduction also enables them to escape the toxic effects of chemical protectants. Losses from bacterial diseases are reduced by the use of pathogen-free seed grown in arid regions. Examples of diseases controlled by this method include bacterial blights of beans and peas, black rot of crucifers, and bacterial spot and canker of tomato. Seed treatment with hot water at about 50° C (120° F) is also effective for crucifers, cucurbits, carrot, eggplant, pepper, and tomato. Bactericidal seed compounds control some bacterial diseases, such as angular leaf spot of cotton, gladiolus scab, and soft rot of ornamentals. Rotation with nonhost crops reduces losses caused by wilt of alfalfa, blights of beans and peas, black rot of crucifers, crown gall, and bacterial spot and canker of tomato. Eradication and exclusion of host plants has been useful against citrus canker, angular leaf spot of cotton, fire blight, and crown gall. Resistant varieties of crop plants have been developed to reduce losses from wilts of alfalfa, corn, and tobacco; angular leaf spot of cotton and tobacco; and bacterial pustule of soybeans, among others. Protective insecticidal sprays help control bacterial diseases, such as wilts of sweet corn and cucurbits and soft rot of iris. Protective bactericidal sprays, paints, or drenches containing copper or antibiotics are used against bacterial blights of beans and celery, fire blight, crown gall, blackleg of delphinium, and filbert and walnut blights. Finally, sanitary measures—*i.e.*, clean plow down of crop refuse, destruction of volunteer plants and weeds, sterilization of pruning and grafting tools—as well as refraining from cultivating when foliage is wet, overhead watering and spraying of indoor plants, and late cutting or grazing of alfalfa and other crops, are useful in reducing the incidence of bacterial diseases.

Fire blight. The first plant disease proved to be caused by a bacterium was fire blight of apple, pear, quince, hawthorn, *Pyracantha*, and related plants. Fire blight, caused by *Erwinia amylovora*, has destroyed all high-quality pear orchards in parts of the United States. The bacteria invade flower nectaries and leaves, spread to young fruits, and pass rapidly down susceptible shoots in the phloem, darkening and killing the tissue. Extensive, slightly sunken cankers, with a definite margin, form, and during warm, moist weather, bacterial ooze (exudate) appears on the surface. The bacteria overwinter at the edges of cankers on the branches or the trunk. Plant parts and other trees become infected when rains splash the bacteria-laden exudate that appears at the edges of cankers.

Invasion and spread of fire blight bacteria

Careful removal of diseased parts and weekly application of antibiotic sprays during the blossom and post-bloom period greatly reduce disease severity. Resistance to fire blight is available in species of *Pyracantha*, *Cotoneaster*, and *Crataegus* (hawthorn) as well as in varieties of crabapples, apples, and pears.

Crown gall. Perhaps the best known bacterial plant disease is crown gall, which derives its name from the rounded, rough growth that appears at the stem base or root crown. The disease occurs widely over the world. The causative bacterium, *Agrobacterium tumefaciens*, which is distributed with nursery stock, has a wide host range, affecting plants in more than 40 families, including peach, apple, pecan, grape, fig, raspberry, rose, willow, euonymus, and sugar beet.

The bacteria enter only through wounds, and grow between cells that are then stimulated to increase in size and number. Bacteria may be washed from the gall surface into the soil or transported by chewing insects. The disease has been controlled by nursery inspection, disinfection of parts used for vegetative propagation, and planting in well-drained soil free of root-chewing insects.

Southern bacterial wilt. One of the most important bacterial diseases is the wilt caused by *Pseudomonas solanacearum*, which attacks tobacco, potato, tomato, eggplant, beans, bananas, peanut, and numerous other vegetables and ornamentals. The disease occurs in warm temperate, subtropical, and tropical areas throughout the world.

Typical external symptoms include stunting, yellowing, and wilting. The vascular tissue usually becomes dark brown. The bacteria may live in fallow (plowed but unplanted) soil for six years or more and then usually enter roots through wounds and spread upward in water-conducting tissue. There the bacteria multiply, forming an extracellular slime and other products that eventually prevent water movement to the leaves and cause wilting.

Insects are usually not involved in plant-to-plant spread, with the exception of banana, in which case insects may introduce the bacteria into the flower parts. Spread into new areas occurs by infected transplants, potato tubers, or by cultural practices—such as irrigation and use of



Diseases caused by bacteria and virus.
(Left) Decayed leaves of lettuce infected with bacterial soft rot. (Centre) Viral ring spot of peonies. (Right) Apple twig infected with fire blight.
By courtesy of (left, right) Malcolm C. Shurtleff, University of Illinois, Urbana, (centre) Timothy H. Bowyer, University of Illinois, Urbana

tools and machinery coated by soil containing the organism. Development of disease is favoured by high soil moisture levels and high temperatures (optimum about 30–32° C, or 86–90° F).

Control of bacterial wilt

Bacterial wilt of tobacco and peanut has been controlled by development of resistant varieties. High-quality varieties of tomato, potato, and banana with adequate wilt resistance are not yet available. Normal crop rotations, soil treatments, and other practices have not provided adequate control on most crops.

Fungi. Of the 100,000 described species of fungi, approximately 20,000 produce disease in plants. These include 4,600 rusts, 700 smuts, and 1,000 powdery and downy mildews—a total of 6,300 obligate parasites.

Fungi cause the great majority of infectious plant diseases (about 75 percent). They include all white and true rusts, smuts, needle casts, leaf curls, mildews, sooty molds, and athracnoses; most leaf, fruit, and flower spots; cankers; blights; scabs; root, stem, fruit, and wood rots; wilts; leaf, shoot, and bud galls; and many others. All economically important plants apparently are attacked by one or more fungi; often ten to 50 or more different fungi may cause disease on one plant species.

Fungal hyphae (filaments of the main growing body of a fungus) may penetrate a plant by growing into a wound, by passing through a natural opening, or by forcing their way directly through a plant's protective epidermis. Fungi, like bacteria, are more prevalent and damaging in damp areas or wet seasons than in dry ones. Moisture is usually essential for their rapid spread, penetration, and infection. Fungi overwinter in soil, plant refuse, living plants, seeds, storage organs, and sometimes insects.

Descriptions of the various fungus groups and their activity appear elsewhere (see MYCOTA). The examples below represent major types of diseases incited by fungi.

Late blight. The causal fungus of late blight (*Phytophthora infestans*) occurs in cool, moist regions wherever potatoes are grown. A devastating epiphytotic of this disease on potatoes began in Europe in 1845 and brought about the Irish famine that caused starvation, death, and mass migration of the population. In a population of 8,000,000, about 1,000,000 (about 12.5 percent) died of starvation and 1,500,000 (almost 19 percent) emigrated, mostly to the United States, as refugees from the destructive blight. This fungus thus had a tremendous influence on the economic, political, and cultural development in Europe and the U.S. During World War I, late blight damage to the potato crop in Germany may have helped end the war.

Symptoms of late blight

Symptoms of late blight may appear on any part of the potato plant, including potatoes in storage. Water-soaked, dark green to black or purplish lesions with pale green margins appear first on the lower leaves. The lesions often rapidly increase in size. A downy white mildew (composed of sporangiophores and sporangia—spore-producing structures) is usually present near the lesion margins on the underside of leaves. Tubers may also be infected in field or storage. A brown or rusty, sometimes purplish skin discoloration appears, and a reddish-brown dry rot may extend into the tuber for five to ten millimetres. Later, a slimy, foul-smelling rot may destroy the tuber. Progress of late blight through a potato field is rapid in cool, moist weather. Hot, dry weather checks the spread of the disease.

The infective cycle begins with the planting of infected tubers, in which the fungus overwinters. Infection may also result from diseased cull (waste) piles or volunteer field plants. Spores from these sources infect nearby plants and spread rapidly in successive cycles of increasing intensity so long as cool, moist weather prevails. External symptoms appear on the potatoes five days or less from the time of infection; the *Phytophthora* fungus quickly forms mildew again on the undersides of leaves.

Controls include the frequent and thorough application of protectant sprays, use of certified seed potatoes, destruction of potato dump and cull piles, delay of digging until two weeks after tops die or are purposely killed, and growing of resistant varieties.

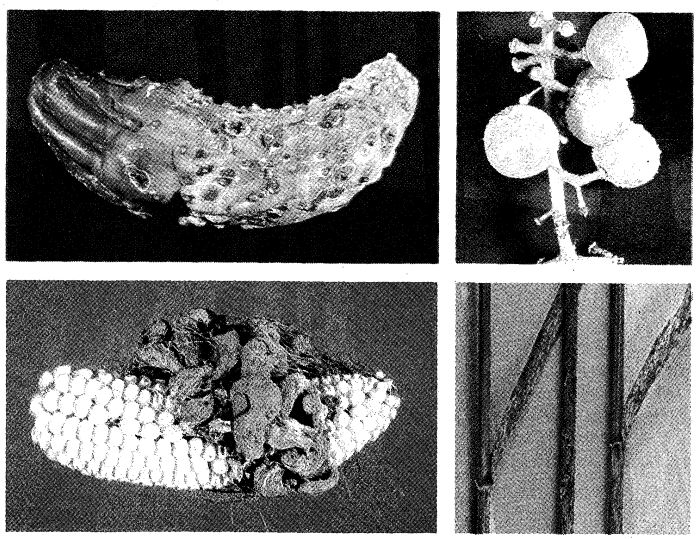
Blight-forecasting services have been started in several countries; they are based on weekly graphs prepared by plotting the cumulative rainfall, periods of time leaf surfaces are wet, and mean temperatures. The forecasting of weather favourable and unfavourable for a blight epiphytotic has saved potato farmers time and money by eliminating unneeded sprays.

Chestnut blight. Native American chestnuts have been practically eliminated from their natural range by a fungus (*Endothia parasitica*), accidentally imported from the Orient, where blight occurs as a minor disease on native chestnuts. When blight was first observed in New York in 1904, its seriousness was not realized. By 1908, however, chestnut blight had killed thousands of trees in areas around New York, and the fungus had spread many miles in all directions. By 1925, the disease had eliminated chestnuts in Illinois; four years later, it reached the Pacific Northwest. Costly efforts failed to bring the epiphytotic under control, and it continued to spread, killing millions of trees and finally eliminating this valuable native hardwood species as a major forest component in eastern North America.

Young infections on two- to six-year-old chestnut shoots appear as yellowish to reddish-brown patches against the olive green of healthy bark. The lesions spread until the twig or limb is girdled and killed by slightly sunken or swollen and cracked cankers. After one to several years, large trees assume a staghead appearance; blight quickly spreads down into the trunk, killing the entire top of the tree. Roots and old stumps have a surprising capacity for sending up sprouts, which often escape the disease for several years. During moist weather, spores (conidia), embedded in a mucilaginous matrix are extruded from reddish reproductive structures (pycnidia) in back crevices, forming long, irregularly twisted, yellow to buff coloured tendrils or spore horns.

Spread of chestnut blight

The extruded conidia are spread short distances by splashing rain and wind or insects and over long distances by migratory birds. Presumably, birds carry spores on their feet to healthy trees, sometimes many miles away. The number of spores produced is enormous. An average spore horn contains over 100,000,000 spores. In addition to conidia in spore horns, a second sexual stage is produced—vast numbers of ascospores are discharged from flask-shaped fruiting bodies (perithecia) and are carried many miles by wind currents.



Diseases caused by fungi.
(Top left) Cucumber showing small, circular, water soaked lesions and brown necrotic tissue of fungal scab. (Top right) Grapes covered with white fungus spores of downy mildew. (Bottom left) Ear of corn deformed by large smut gall. (Bottom right) Stem, sheath, and leaves of wheat covered with red rust, or summer stage of black stem rust fungus (*Puccinia graminis*).
By courtesy of (top left) University of Illinois, Urbana, (top right) the Morton Arboretum, Lisle, Illinois, (bottom left) University of Wisconsin, Madison, (bottom right) Malcolm C. Shurtleff, University of Illinois, Urbana

Although no control measures were available to save the native American chestnuts, Chinese and Japanese species show good to excellent resistance. Crosses between American and Asiatic chestnuts have produced varieties with excellent nuts; timber quality of these trees, however, usually is closely linked with susceptibility to blight.

Dutch elm disease. A serious new disease of elms was described in 1921 in The Netherlands, thus the name Dutch elm disease. The causal fungus (*Ceratocystis ulmi*) is believed to have moved to Europe from Asia during World War I. Following description of the disease, reports of its occurrence soon came from many regions in Europe and Asia. The fungus was apparently introduced into the United States in 1930 on elm burl logs imported for furniture veneer. Overland spread from original introduction points occurred by road and rail traffic and especially by movement of elm bark beetles (European: *Scolytus multistriatus;* American: *Hylurgopinus rufipes*—both are carriers). These beetles may fly up to several miles in search of breeding sites in weakened or dead elm wood. Dutch elm disease has killed millions of elms in the eastern two-thirds of the U.S. and Canada, and the disease has continued to spread westward and southward.

Leaves on diseased elm twigs suddenly and progressively wilt, turn dull green to yellow or brown, curl, and usually drop off. Progress of the disease is rapid in young, rapidly growing trees, which may die in one to two months. Older or more slowly growing elms sometimes die slowly over several years, the upper dead branches appearing stagheaded. A cut across an infected branch shows a brown to almost black arc or ring where tyloses (growths or plugs that fill the interiors of vessel cells) and brownish gum have accumulated in water-conducting vessels. Symptoms are easily confused with other elm wilts caused by species of *Verticillium* or *Dothiorella*. Growth in the laboratory of the causal fungus is thus necessary for positive diagnosis.

The fungus spreads rapidly, especially in the spring, through the large xylem vessels. European and American elm bark beetles tunnel into dying or dead elm branches and trunks where the females lay eggs under the bark. Conidia of the fungus develop in great abundance on coremia (fruiting structures) in the beetles' galleries. When the adult beetles emerge from the galleries, sticky spores are carried on their bodies. Spores are deposited in vascular tissue of nearby healthy elms when the beetles feed in the crotches of young twigs or leaf axils (the angles between leafstalks and twigs). Healthy elms also become infected by natural-root grafts within 50 feet (15 metres) of diseased trees.

Smoothleaf elm (*Ulmus carpinifolia*) seedlings have shown resistance in The Netherlands. Some species, such as Chinese elm (*U. parvifolia*) and Siberian elm (*U. pumila*), as well as hybrids between American and Asiatic elms, are resistant. The best control measures include the removal and destruction of beetle breeding sites and all dead and weakened elm wood and the use of a dormant spray with a suitable insecticide which prevents the beetles from feeding in small elm twig crotches. Disease spread by root grafts can be stopped by a fumigant.

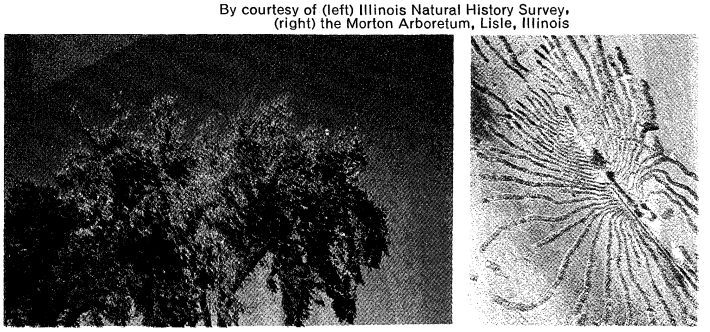
Control of Dutch elm disease

Black stem rust of wheat. One of the oldest and still important diseases is black stem rust, which is caused by the fungus *Puccinia graminis.* All cereals and many grasses are infected by one of the six varieties of this rust. The discussion below is limited to wheat. Black stem rust occurs essentially wherever wheat is grown. In 1935 this rust destroyed almost 60 percent of the total hard red spring-wheat crop in Minnesota and the Dakotas. In 1953 and 1954 outbreaks of black stem rust in the same area reduced yields as much as 80 percent on durum wheats and 30 percent on spring wheats. Losses were measured in tens of millions of bushels.

During its complex life cycle, *Puccinia graminis* produces five different kinds of spores, three on wheat and two on barberry. The uredial, or red-rust, stage occurs on wheat culms (stems), leaves, and glumes (bracts enclosing flowers) as elongated, reddish-brown pustules (uredia), each containing up to 350,000 rust-coloured spores (uredospores). Developing pustules rupture the epidermis (skin) of the plant, giving it a ragged, powdery appearance. Uredospores, blown to other wheat plants, germinate and initiate infections. The cycle is repeated at intervals of seven to 14 days as long as there is susceptible living wheat and warm, moist weather prevails. Symptoms vary from scattered yellow flecks to numerous large pustules, depending on resistance of the wheat variety and weather conditions.

When the wheat ripens, the rust-coloured pustules gradually turn black as another spore type, the blackish teliospore, develops (hence the name black stem rust). The thick-walled, two-celled teliospores survive the winter, germinating in the spring to form special spores, called sporidia, or basidiospores. Basidiospores cannot infect wheat; they affect only the common barberry, in which the fungus develops yellow-orange spots on the upper leaf surface that contain pycnia—spore-producing structures. The dark, flask-shaped pycnia exude a sticky liquid containing pycniospores. Following cross-fertilization, which is aided by insects, the fungus develops an orange cluster of tubular cups (aecia) on the lower leaf surface. In these cups are formed aeciospores, which can infect wheat but not barberry. When wind-borne aeciospores infect wheat, the red-rust stage develops again. Wind may carry uredospores many hundreds of miles, even across continents.

By courtesy of (left) Illinois Natural History Survey, (right) the Morton Arboretum, Lisle, Illinois



Dutch elm disease.
(Left) Elm trees affected by Dutch elm disease. (Right) Galleries of the elm bark beetle in wood immediately beneath bark, where fungus spores multiply.

The fungus overwinters in the red-rust stage in warm climates. In the spring, waves of uredospores blow northward where they start new infections on wheat—bypassing the barberry.

Eradication of barberry

Since 1918, intensive efforts to eradicate the susceptible common barberry (*Berberis vulgaris*) have continued in North America. Many millions of bushes, each capable of producing several billion spores each spring, have been removed by hand or killed by chemicals. The Japanese barberry (*B. thunbergi*), a common ornamental shrub, is fortunately immune to black stem rust.

The rust fungus and the wheat plant are highly variable. Through distinctive rust symptoms produced on 12 differential wheat varieties, over 325 physiological races of black stem rust have been described. Only about 12 races are important in any year. Certain wheat varieties are highly resistant or immune to some races but susceptible to others. No wheat is highly resistant to all races.

Eradication of rust-spreading barberries is important for two reasons. First, new rust races are produced on barberry from hybridization or cross-fertilization of existing races; second, rust spreads from barberry to nearby wheat two or three weeks before uredospores are normally wind-borne from the south. With an early start, rust can severely damage young grain plants.

Wheat-breeding programs emphasize the development of varieties resistant to regionally important races of black stem rust, other rusts, smuts, root rots, other diseases, insects, and various characters that add to the quality and yield of grain. To keep ahead of the variability of the fungi, wheat breeding must be a continuing process.

Blister rust disease of white pines, caused by *Cronartium ribicola*, has a comparable cycle to black stem rust, but with gooseberries and currants (*Ribes* species) as the alternate hosts. In this case, the white pines carry the fungus spore stages equivalent to those found in the barberry of the wheat rust disease; and the stages infecting the *Ribes* alternate host are like those found in the wheat plants.

Wood rots. Certain higher fungi actively decompose and stain forest timber. Losses from dead trees and effects of decay on growth may approach 45 percent of the timber losses from all causes. Such fungi may decay lumber and wood products already in use, making expensive replacements necessary.

Nematodes. Nematodes parasitic on plants are active, slender, unsegmented roundworms (also called nemas, or eelworms). The great majority cannot be seen with the unaided eye because they are tiny and translucent. Practically all adult forms fall within the range of 0.25 to two millimetres (0.01 to 0.08 inch) in length. About 1,200 species cause disease in plants. Probably every form of plant and animal life is fed upon by at least one species of nematode. They usually live in soil and attack small roots, but some species inhabit and feed in bulbs, buds, stems, leaves, or flowers (see ASCHELMINTHES).

Mode of nematode attack. Nematodes parasitic on plants obtain food by sucking juices from them. Feeding is accomplished by a hollow, needlelike mouthpart called a spear or stylet. The nematode pushes the stylet into plant cells and injects a liquid containing enzymes (biological catalysts), which digest plant cell contents. The liquefied contents are then sucked back into the nematode's digestive tract through the stylet. Nematode feeding lowers natural resistance, reduces vigour and yield of plants, and affords easy entrance for wilt-producing or root-rot-producing fungi or bacteria and other nematodes. Nematode-infested plants are weak and often appear to suffer from drought, excessive soil moisture, sunburn or frost, a mineral deficiency or imbalance, insect injury to roots or stem, or disease.

Symptoms of nematode injury

Common symptoms of nematode injury include stunting, loss of green colour and yellowing; dieback of twigs and shoots; slow general decline; wilting on hot, bright days; and lack of response to water and fertilizer. Feeder root systems are reduced; they may be stubby or excessively branched, often discoloured, and decayed. Winterkill of orchard trees, raspberries, strawberries, ornamentals, and other perennials is commonly associated with nematode infestations.

Root injury develops partly from the nematodes feeding on cells and partly from toxic salivary excretions of the parasite. Tissues often respond by producing either an enlargement (hypertrophy) or degeneration of cells; sometimes both occur.

Many nematodes are native and attack cultivated plants when their natural hosts are removed. Others have been introduced with seedling plants, bulbs, tubers, and particularly in soil balled around roots of infested nursery stock.

Nematodes may live part of the time free in soil around roots or in fallow gardens and fields. They tunnel inside plant tissues (endoparasites) or feed externally from the surface (ectoparasites) and may enter a plant through wounds, natural openings, or by penetrating roots and pushing in between cells. All nematodes parasitic on plants require living plant tissues for reproduction. Nematodes are attracted to host roots by sensing either the heat given off by roots or the chemicals secreted by roots.

Most species require 20 to 60 days to complete a generation from egg through four larval stages to adult and back to egg. Some nematodes have only one generation a year but still produce several hundred offspring.

Soil populations and developmental rate of nematodes are affected by the length of the growing season; temperature; availability of water and nutrients; and moisture, type, texture, and structure of soil. Also important are populations of nematode-parasitic bacteria, viruses, some 50 different nematode-trapping fungi, protozoans, mites, flatworms, or other pests, and other nematodes. Toxic chemicals in the soil or secreted by plant roots; crop rotations and past cropping history; species, variety, age and nutrition of growing plants; and other factors are additional conditions that affect nematode populations.

Certain species live strictly in light, sandy soils; some built up high populations in muck soils; and a few seem to thrive in heavy soils. High populations and greater crop damage are much more common in light sandy soils than in heavy clay soils.

Many plant-infecting nematodes become inactive at temperatures between 5° and 15° C (41° and 59° F) and 30° and 40° C (86° and 104° F). The optimum for most is 20° to 30° C (68° to 86° F), but this varies greatly with the species, stage of development, activity, growth of the host, and other factors.

Numbers of nematodes in diseased plants

Nematodes may be found in plant tissues in large numbers. A large *Narcissus* bulb may contain 250,000, an infested root system of a peppermint or tomato plant 100,000, and one gram of young strawberry roots 3,500 at the height of the growing season.

After a plant-infecting nematode has been accidentally introduced into a garden or field, several years pass before the population builds up sufficiently (*i.e.*, up to several billion or more active nematodes per acre) to cause conspicuous symptoms in a large number of plants. This is because nematodes move very slowly through soil—rarely more than 30 inches (75 centimetres) a year. Nematodes, are easily spread, however, by moving infested soil, plant parts, or contaminated objects—*e.g.*, tools and machinery, bags and other containers, running water, wind, clothing, shoes, animals, and birds, and infested planting stock.

Nematode diseases. Root-knot nematodes (*Meloidogyne* species) are well known because of conspicuous "knots," or gall-like swellings, they induce on roots. Over 2,000 kinds of higher plants are subject to their attack. Losses are often heavy, especially in warm regions with long growing seasons. Certain species, however, such as the northern root-knot nematode (*M. hapla*), are found where soil may freeze to depths of three feet. Vegetables, cotton, strawberries, and orchard trees are commonly attacked. Garden plants and ornamentals frequently become infested through nursery stock.

Root-lesion nematodes (*Pratylenchus* species), cosmopolitan in distribution, are endoparasites that cause severe losses to hundreds of different crop and ornamental plants by penetrating roots and making their way through

the tissues, breaking down the cells as they feed. They deposit eggs from which new colonies develop. After a root begins to decline in vigour, they move into the soil in search of healthy roots. Lesions form in the root as fungi and bacteria enter damaged tissues, and root rot often occurs. Annual crops may succumb early in the season, but perennials and orchard trees may not decline (show reduced vigour and growth) for several years.

The golden nematode of potatoes (*Heterodera rostochiensis*) is a menace of the European potato industry. Great efforts have been made to control it. The speck-sized golden cysts that dot infested plant roots are the remains of female bodies. Each cyst may contain up to 500 eggs, which hatch in the soil over a period of up to 17 years. A chemical given off by potato and tomato roots stimulates hatching of the eggs.

A related, cyst-forming species, the sugar-beet nematode (*H. schachtii*), is a pest that has restricted acreage of sugar beets in Europe, Asia, and America.

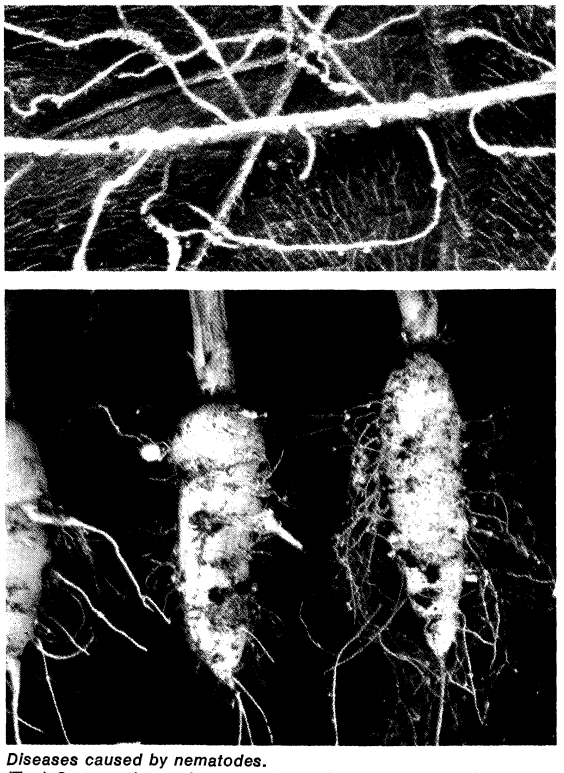
The citrus nematode (*Tylenchulus semipenetrans*) occurs wherever citrus is grown, exacting a heavy toll in fruit quality and production. Typical symptoms are a slow decline, yellowing and dying of leaves, and dieback of twigs and branches in many groves 15 years or older. Infested nursery stock has widely distributed the nematode. The burrowing nematode (*Radopholus similis*) is a serious endoparasite in tropical and subtropical areas, where it attacks citrus (causing spreading decline), banana, avocado, tomato, black pepper, abaca, and over 200 important crops, trees, and ornamentals, causing severe losses.

Root-feeding nematodes

Many important ectoparasites feed on plant roots—dagger nematodes (*Xiphinema*), stubby-root nematodes (*Trichodorus*), spiral nematodes (*Rotylenchus* and *Helicotylenchus*), sting nematodes (*Belonolaimus*), and pin nematodes (*Paratylenchus*). Leaf, or foliar, nematodes (*Aphelenchoides* species) and bulb and stem nematodes (*Ditylenchus dipsaci*) cause severe losses in vegetable and ornamental bulb crops, clovers, alfalfa, strawberry, sweet potato, orchids, chrysanthemums, begonias, and ferns.

Control measures. Control measures for nematodes often include rotation with nonhost plants, growing of resistant varieties and species, use of certified, nematode-free nursery stock, and use of soil fumigants (nematicides) as preplant or postplant treatments. Steam or dry heat is applied to soil in confined areas, such as greenhouse benches and ground beds. Exposure to moist heat, such as steam or hot water, at 50°C (120° F) for 30 minutes, is sufficient to kill most nematodes and nematode eggs. Shorter periods are needed at higher temperatures. Hot water soak treatments have been developed for freeing a wide range of seeds, bulbs, corms, and tubers, as well as potted or dormant plants, of nematodes. State and federal quarantines prohibiting movement of infested soil, plants, or plant parts, machinery, and other likely carriers also exist. Cultural practices to promote vigorous plant growth (*i.e.*, watering during droughts, proper application of fertilizers, clean cultivation, fall and summer fallowing, use of heavy organic mulches or cover crops, and plowing out roots of susceptible plants after harvest) are useful for specific nematodes. Asparagus, marigolds (*Tagetes* species), and *Crotalaria* species are toxic to many plant-infecting nematodes. They contain a number of chemical substances that are receiving attention as possible means of biological control in reducing soil nematode populations.

By courtesy of (top) University of Illinois, Urbana, (bottom) Malcolm C. Shurtleff, University of Illinois, Urbana



Diseases caused by nematodes.
(Top) Cysts or tiny nodes on soybean plant roots, containing eggs of nematodes. (Bottom) Small, gall-like swellings on carrot roots caused by root-knot nematodes that distort tissue.

Parasitic seed plants. A number of flowering plants are important as parasites of other plants. Among the more important ones are mistletoe, dodder, and witchweed.

Mistletoe. Mistletoes are semiparasitic seed plants that feed on trees and obtain water and mineral salts by sending rootlike structures (haustoria) into vascular tissue of the inner bark (see SANTALALES). There are three important types: American (*Phorodendron* species), European (*Viscum album*), and dwarf (*Arceuthobium* species). All produce sticky seeds spread by birds. American mistletoe, restricted to the Americas, is best known for its ornamental and sentimental uses at Christmas time. The leafy, bushy evergreen masses, up to three feet or more in diameter, appear on tree branches. They are most conspicuous after deciduous leaves have fallen. The European mistletoe is similar in habit and appearance to its American relative. Tree branches infected by mistletoes may become stunted or even die.

Dwarf mistletoe

Dwarf mistletoe is common on and very destructive to conifers in forests. Seedlings and young trees may be stunted, deformed, or killed. Conspicuous witches'-brooms form in the crown or spindle-shaped swellings (later cankers) in limbs and trunk. Canker and wood-rotting fungi often enter through mistletoe wounds. Dwarf mistletoes often escape detection because the scaly-leaved plants may be less than one inch (2½ centimetres) long; they do range to 12 or 18 inches (30 to 45 centimetres), however. Dwarf mistletoes occur scattered along conifer limbs and small branches. After the mistletoe has grown internally for about a year, the branch may start to form a witches'-broom. Four to five years elapse before the yellow to brown to olive-green shoots form fruits. The sticky seeds are shot with explosive force from the fruit for horizontal distances ranging from 15 (five metres) to more than 60 feet (18 metres); this is one of the most remarkable methods for seed discharge among plants. Once seeds adhere to a branch, they germinate on young bark and penetrate into the host tree's vascular system. Control for mistletoes in individual trees involves removal of infected branches a foot or more beyond any evidence of the parasite before the fruits ripen.

Dodder. The 170 species of dodder (*Cuscuta*) are widely distributed and called such names as strangleweed, devil's-hair, pull down, hell-bind, love vine, and goldthread. The leafless, yellow-orange, threadlike stems twine around a number of field and garden host plants. By extending to nearby plants, it may draw them together and downward until a tangled yellowish-orange patch is formed. The infested area is usually less than three metres across the first year; it spreads more rapidly in succeeding years. Dodder is widely distributed as a contaminant with field seed; hence the losses in clover, alfalfa, and flax fields. Dodder is controlled by planting

certified, properly cleaned seed, and by mowing patches of dodder in the field well before the seeds form. The dried patches are sprinkled with fuel oil and burned. Careful application of selective herbicides or a soil fumigant, and sowing heavily infested areas with resistant plants (*e.g.*, garden beans, soybean, corn, cowpea, pea, grasses, or small grains) are also control methods.

Witchweed. Witchweed, a small parasitic weed (*Striga asiatica*), is widely distributed in Asia and south Africa. It has been known in the coastal sandy soils of North and South Carolina since the mid 1950s but through intensive efforts is slowly being eradicated. Witchweed parasitizes the roots of many hosts, including maize (corn), sorghum, sugarcane, rice, small grains, and more than 50 species in the grass and sedge families. A serious infestation may cause corn plants to be severely stunted, wilt, and turn yellow or brown thus reducing the acre yield. *Striga* plants rarely grow over eight to ten inches (20 to 25 centimetres) tall with small, red, yellowish-red, yellow, or white flowers. One plant may produce 500,000 tiny brown seeds that can remain alive in soil for years until stimulated to germinate by a secretion from a nearby host root. Witchweed robs the host of water and food, causing it to grow more slowly than normal and often to die before maturing. Control is difficult; useful measures include application of selective herbicides before seeds are produced; rotation with a resistant crop and keeping it free of weed grasses that may serve as hosts; and prevention of seed set by growing trap crops and then destroying them with herbicides.

Control of witchweed

PRINCIPLES OF DISEASE CONTROL

Successful disease control is a broad, highly technical, and rapidly developing scientific field based on accurate diagnosis; thorough knowledge of the causal agent and the disease cycle; host–pathogen interactions in relation to environmental factors; and cost. Disease control starts with the best variety, seed, or planting stock available and continues through the growing season in garden or field. Relatively few diseases are controlled by one method; the majority require several approaches. These often need to be integrated into a broad program of biological or cultural and chemical methods to control as many different pests—including diseases, insects, mites, rodents, and weeds—on a given crop as possible (see PEST CONTROL; WEED CONTROL).

Most control measures are directed against inoculum of the pathogen or disease agent and involve the principles of exclusion and avoidance, eradication, protection, host resistance and selection, and therapy.

Exclusion and avoidance. The principle of exclusion and avoidance is to keep the pathogen away from the growing host plant. This practice commonly excludes pathogens by disinfection of plants, seeds, or other parts, using chemicals or heat. Inspection and certification of seed and other planting stock help insure freedom from disease. For gardeners this involves sorting bulbs or corms before planting and refusing to buy obviously diseased plants. Federal and state plant quarantines, or embargoes, have been established to prevent introduction of potentially destructive pathogens into areas currently free of the disease. More than 150 countries now have established quarantine regulations.

Eradication. Eradication is concerned with elimination of the disease agent after it has become established in the area of the growing host or has penetrated the host. Such measures include crop rotation, destruction of the diseased plants or alternate host plants, pruning, disinfection, and heat treatments.

Crop rotation with nonsusceptible crops "starves out" bacteria and fungi with a restricted host range. Some pathogens can survive only as long as the host residue persists, usually no more than a year or two. Many pathogens, however, are relatively unaffected by rotation because they become established as saprophytes or soil inhabitants (*e.g.*, *Fusarium* and *Pythium* species; *Rhizoctonia solani;* and the potato scab actinomycete, *Streptomyces scabies*) or their propagative structures remain dormant but viable for many years (*e.g.*, cysts of cyst nematodes, sporangia of the cabbage clubroot fungus, and onion smut spores).

Burning, deep plowing of plant debris, and fall spraying are used against such diseases as leaf blights of tomato, Dutch elm disease, and apple scab. Destruction of weed hosts also helps control such virus diseases as cucumber mosaic and curly top. Destruction of alternate hosts is used against black stem rust of cereals and white pine blister rust. Destruction of diseased plants helps control Dutch elm disease, oak wilt, and peach virus diseases—mosaic, phony peach, and rosette. Elimination of citrus canker in the southeastern United States has been one of the few successful eradication programs in history. Infected trees were sprayed with oil and burned.

Burning and destruction

Pruning and excision of a diseased portion of the plant, followed by a protective wound dressing, have aided in reducing inoculum sources for canker and wood-rot diseases of shade trees and fire blight of pome fruits. Disinfection of contaminated tools, as well as packing and shipping containers, controls a wide range of diseases. Direct application of dry or wet heat is used to obtain seeds, bulbs, other propagative materials, and even entire plants free of viruses, nematodes, and other pathogens.

Protection. The principle of protection involves placing a protective barrier between the pathogen and the susceptible part of the host. This can be accomplished by regulation of the environment, cultural and handling practices, control of insect carriers, and application of chemical toxicants (poisons).

Regulation of the environment. Selection of outdoor growing areas where weather is unfavourable for disease is a method of controlling disease by regulating the environment. Control of virus diseases of potato, for example, can be accomplished by growing the seed crop in northern regions where low temperatures are unfavourable for the aphid carriers. Another environmental factor that can be brought under control is the storage and in-transit environment. A variety of post-harvest diseases of potatoes, sweet potatoes, onions, cabbage, apples, pears, and other crops are controlled in storage and shipment by keeping humidity and temperature low and by reducing the quantity of ethylene and other natural gases in storage houses.

Cultural practices. Selection of the best time and depth of seeding and planting is an effective cultural practice that reduces disease impact. Shallow planting of potatoes may help to prevent *Rhizoctonia* canker. Early fall seeding of winter wheat may be unfavourable for seedling infection by wheat-bunt teliospores. Cool-temperature crops can be grown in soils infested with root-knot nematode and harvested before soil temperatures become favourable for nematode activity. Adjustment of soil moisture is another cultural practice of widespread usefulness. For example, seed decay, damping-off, and other seedling diseases are favoured by excessively wet soils. The presence of drain tiles in poorly drained fields and the use of ridges or beds for plants are often beneficial. Adjustment of soil pH also leads to control of some diseases. Common potato scab can be controlled by adjusting the pH to 5.2 or below; other acid-tolerant plants then must be used in crop rotation, however.

Temperature, moisture, and soil reaction

Regulation of fertility level and nutrient balance. Potash and nitrogen, and the balance between the two, may affect the incidence of certain bacterial, fungal, and viral diseases of corn, cotton, tobacco, and sugar beet. A number of microelements (*i.e.*, nutrient elements required by plants in minute quantities), including boron, iron, zinc, manganese, magnesium, copper, sulfur, and molybdenum, may cause noninfectious diseases of many crop and ornamental plants. Adjusting the soil pH, adding chelated (bound or enclosed in large organic molecules) or soluble salts to the soil, or spraying the foliage with these or similar salts is a corrective measure.

Handling practices. Late blight on potato tubers can be controlled by delaying harvest until the foliage has been killed by frost, chemicals, or mechanical beaters. Avoidance of bruises and cuts while digging, grading, and packing potatoes, sweet potatoes, and bulb crops also reduces disease incidence.

Control of insect vectors. There are many examples in which losses by virus and mycoplasma diseases can be reduced by controlling aphids, leafhoppers, thrips, and other carriers of these agents.

Toxicants. The term toxicants includes chemicals that inhibit or kill bacteria (bactericides), fungi (fungicides), and nematodes (nematicides). Toxicants can be applied to seed, foliage, fruit, or soil and prevent infection by placing a toxic barrier between plant and inoculum. Germinating fungus spores are very sensitive to fungitoxicants (see PEST CONTROL).

Seed treatments are used to protect many kinds of seeds, bulbs, corms, and tubers. These are routinely treated with chemicals to eradicate pathogenic bacteria, fungi, and nematodes and to protect the seed against decay and damping-off organisms (mainly fungi) in the soil. Major materials used in eradicative seed treatments are hot water, formaldehyde, benomyl, and DMOC (5,6-dihydro-2-methyl-1,4-oxathiin-3-carboxanilide). Benomyl and DMOC are also protectants. Some protective seed treatments include thiram, captan, chloranil, maneb, HCB (hexachlorobenzene), and Dexon (*para*-dimethylaminobenzenediazosodium sulfonate). Products used for systemic (internal) treatments are taken up by seeds or seedlings and distributed within the plants. The better materials are DMOC, benomyl, TBZ (2-[4-thiazolyl]-benzimidazole), and chloroneb.

Chemicals used in soil treatments

Soil treatments are designed to kill soil-inhabiting nematodes, fungi, and bacteria. Such soil-treatment types as steam, methyl bromide, chloropicrin, MIT (methyl isothiocyanate), DMTT (3,5-dimethyltetrahydro-1,3,5,2*H*-thiadiazine-2-thione), SMDC (sodium *N*-methyldithiocarbamate dihydrate), D–D (a mixture of 1,3-dichloropropene and 1,2-dichloropropane), EDB (ethylene dibromide), and formaldehyde are used after the crop is harvested or before planting. Some fungitoxicants (*i.e.*, PCNB [pentachloronitrobenzene], captan, chlorothalonil, folpet, thiram, zineb, ziram, chloroneb, Dexon, ferbam, TBZ, and benomyl) can be sprayed on young seedlings in the seedbed or mixed with soil at planting time.

Foliage and fruit treatments include a wide range of organic chemicals that are used for spraying and dusting a diversity of field, plantation, fruit, and vegetable crops as well as ornamentals—*e.g.*, captan, zineb, maneb, chlorothalonil, dodine, ferbam, folpet, thiram, benomyl, fixed coppers, and sulfur compounds. Antibiotics, such as streptomycin and cycloheximide (Acti-dione), have been developed for use on plants. Most of these organic materials have relatively little toxicity to plants, animals, or humans and are rapidly degraded on plants and in soil. Without these materials to control fungal diseases, production on a commercial scale of such crops as bananas, potatoes, tomatoes, grapes, stone fruits, apples, and pears would be impossible.

Many toxicants are combined by the grower or are sold as multipurpose sprays and dusts so that one application acts against insects, mites, and plant pathogens.

Host resistance and selection. Ideal control is normally provided by plant varieties with good genetic resistance to most common diseases. Resistant or immune varieties are critically important for low-value crops in which other controls are unavailable or their expense makes them impractical. Much has been accomplished in developing disease-resistant varieties of field crops, vegetables, fruits, turf grasses, and ornamentals. Although great flexibility and potential for genetic change exist in most economically important plants, pathogens are also flexible. Sometimes, a new plant variety is developed that is highly susceptible to a previously unimportant pathogen.

Variable resistance. The characters associated with resistance or susceptibility are subject to several types of variation. These include nongenetic variations, which result from a variation in ecological conditions, and genetic variations. Genetic variations are new gene combinations resulting from normal sexual reproduction.

Obtaining disease resistant plants. Several means of obtaining disease-resistant plants are commonly employed alone or in combination. These include introduction from an outside source, selection, and induced variation. All three may be used at different stages in a continuous process; for example, varieties free from injurious insects or plant diseases may be introduced for comparison with local varieties. The more promising lines or strains are then selected for further propagation, and they are further improved by promoting as much variation as possible through hybridization or special treatment. Finally, selection of the plants showing greatest promise takes place. Developing disease-resistant plants is a continuing process.

Special treatments for inducing gene changes include the application of mutation-inducing chemicals and irradiation with ultraviolet light and X-rays. These treatments commonly induce deleterious genetic changes, but, occasionally, beneficial ones may also occur.

Methods used in breeding plants for disease resistance are similar to those used in breeding for other characters except that two organisms are involved—the host plant and the pathogen. Thus, it is necessary to know as much as possible about the nature of inheritance of the resistant characters in the host plant and the existence of physiological races or strains of the pathogen.

Therapy. Therapeutic measures are much less used in plant pathology than in human or animal medicine. Recent advances in the development of systemic fungitoxicants that can be applied directly to seeds, soil, or foliage and fruit indicate that revolutionary changes in disease control lie ahead. These new chemicals may restrict spread and development of pathogens by direct or indirect toxic effects or increase the ability of the host to resist infection.

BIBLIOGRAPHY

General works: COMMONWEALTH MYCOLOGICAL INSTITUTE, *Plant Pathologist's Pocket-Book* (1968), a compilation of useful information on the diagnosis, isolation, and culture of various types of plant pathogens; G.L. CAREFOOT and E.R. SPROTT, *Famine on the Wind: Man's Battle Against Plant Disease* (1967), a popular account, vividly told, of how some plant diseases have profoundly influenced the history of many countries of the world; J.L. FORSBERG, "Diseases of Ornamental Plants," *University of Illinois, College of Agriculture, Spec. Publ. No. 3* (1963), a well-illustrated popular book for home owners, professional and amateur gardeners, and commercial florists; N.W. FRAZIER (ed.), *Virus Diseases of Small Fruits and Grapevines* (1970), an important handbook on virus and viruslike diseases of non-tree fruits; K.F. BAKER and W.C. SNYDER (eds.), *Ecology of Soil-Borne Plant Pathogens* (1965), an important collection of 41 review articles on plant pathogens in soil and their biological control; C.S. HOLTON *et al.* (eds.), *Plant Pathology: Problems and Progress, 1908–1958* (1959); J.G. HORSFALL and A.E. DIMOND (eds.), *Plant Pathology: An Advanced Treatise*, 3 vol. (1959–60), an important work that presents an integrated synthesis of the parts of plant pathology; E.C. LARGE, *The Advance of the Fungi* (1940), a popular account of how plant disease epidemics have affected man's economic and political history; P.P. PIRONE, *Tree Maintenance*, 3rd ed. (1959), a popular and practical book written for arborists, city foresters, park superintendents, nurserymen, and the average tree owner that covers all phases of shade and ornamental tree maintenance; M.C. SHURTLEFF, *How to Control Plant Diseases in Home and Garden*, 2nd ed. (1966), a profusely illustrated, popular, encyclopaedia-type book covering more than 5,200 diseases of over 3,200 cultivated plants; UNITED STATES DEPARTMENT OF AGRICULTURE, "Plant Diseases," *Yearbook of Agriculture, 1953* (1953), a series of popular articles written primarily for farmers and home gardeners covering practical discussions on the causes and controls of plant diseases; CYNTHIA WESTCOTT, *Plant Disease Handbook*, 3rd ed. (1971), an excellent reference book for professional and amateur gardeners on over 2,000 diseases of plants grown in gardens, under glass, or in the home within the continental United States.

College-level texts: G.N. AGRIOS, *Plant Pathology* (1969), general considerations of disease, parasitism and pathogenicity, and the biochemistry of host-parasite relationships; F.C. BAWDEN, *Plant Viruses and Virus Diseases*, 4th ed. (1964); A.F. BIRD, *The Structure of Nematodes* (1971); W. CARTER, *Insects in Relation to Plant Disease* (1962); C. CHUPP and A.F. SHERF, *Vegetable Diseases and Their Control* (1960), an exhaustive treatment with numerous fine illustrations; H.B. COUCH, *Diseases of Turfgrasses* (1962); E. EVANS, *Plant Diseases and Their Chemical Control* (1968), a summary of basic concepts, principles, and modern methods of chemical

disease control; S.D. GARRETT, *Pathogenic Root-Infecting Fungi* (1970); R.N. GOODMAN, Z. KIRALY, and M. ZAITLIN, *The Biochemistry and Physiology of Infectious Plant Disease* (1967), a discussion of key biochemical and physiological processes that occur in the healthy plant and how these processes are disturbed; G.H. HEPTING, *Diseases of Forest and Shade Trees of the United States* (1971); W.R. JENKINS and D.P. TAYLOR, *Plant Nematology* (1967), covers plant-parasitic nematodes; K. MARAMOROSCH (ed.), *Viruses, Vectors, and Vegetation* (1969), review articles on the many fascinating interactions between viruses, their numerous carriers, and host plants; T.R. PEACE, *Pathology of Trees and Shrubs* (1962), covers all important infectious and noninfectious diseases of temperate forest and ornamental trees and shrubs; P.P. PIRONE, *Diseases and Pests of Ornamental Plants*, 4th ed. (1970), reference for professional and amateur gardeners that covers the diseases, pests, and other troubles that affect nearly 500 genera of trees, shrubs, flowers, vines, houseplants, and turfgrasses; K.M. SMITH, *Plant Viruses*, 4th ed. (1968), discussion of the properties and diseases caused by each plant-infecting virus and its important strains; C. STAPP, *Pflanzenpathogene Bakterien* (1958; Eng. trans., *Bacterial Plant Pathogens*, 1961), text stressing description and classification, symptoms, host-parasite relations, geographical distribution and control of bacterial plant pathogens in temperate countries; R.B. STREETS, *The Diagnosis of Plant Diseases* (1969), useful field and laboratory manual emphasizing the most practical methods for rapid identification of plant diseases and the pathogens or agents that cause them; G.A. STROEBEL and D.E. MATHRE, *Outlines of Plant Pathology* (1970); J.C. WALKER, *Plant Pathology*, 3rd ed. (1969); B.E.J. WHEELER, *An Introduction to Plant Diseases* (1969); R.K.S. WOOD, *Physiological Plant Pathology* (1967), a description of the factors affecting infection, how bacteria and fungi enter higher plants and cause them to be diseased, and the properties of plants that make them resistant to disease; B.M. ZUCKERMAN, W.F. MAI, and R.A. ROHDE (eds.), *Plant Parasitic Nematodes*, 2 vol. (1971).

(M.C.S./Ar.Kn.)

Disney, Walt

The American motion-picture and television producer Walt Disney pioneered the art of animated cartoon films; he created such universally beloved cartoon characters as Mickey Mouse and Donald Duck and cinematically adapted such storybook favourites as Snow White and the Seven Dwarfs as full-length cartoon motion pictures. He also designed and promoted a world-famous amusement park, Disneyland, in California. A second Disneyland (Walt Disney World) opened in Florida after his death. His imagination and energy, his whimsical humour, and his gift for being attuned to the vagaries of popular taste inspired him to develop well-loved amusements for "children of all ages" throughout the world. His achievement as a creator of entertainment for an almost unlimited public and as a highly ingenious merchandiser of his wares may be compared to that of a successful industrialist. Yet, in spite of his tremendous success, he remained in fact a rather homespun individual, a lover of gadgets who made no pretensions to a knowledge of art and, in the early years of his work, would hardly have dreamed of emerging a tycoon.

Early life. Walter Elias Disney was born in Chicago on December 5, 1901, the fourth son of Elias Disney, a peripatetic carpenter, farmer, and building contractor, and his wife, Flora Call, who had been a public school teacher. When Walt was little more than an infant, the family moved to a farm near Marceline, Missouri, a typical small Midwestern town, which is said to have furnished the inspiration and model for the Main Street U.S.A. of Disneyland. Here Walt began his schooling and first showed a taste and aptitude for drawing and painting with crayons and watercolours.

Inspiration for Disneyland's Main Street U.S.A.

His restless father soon abandoned his efforts at farming and moved the family to Kansas City, Missouri, where he bought a morning newspaper route and compelled his young sons to assist him in delivering papers to home subscribers in rain or shine. Walt later said that many of the habits and compulsions of his adult life stemmed from the disciplines and discomforts of helping his father with the paper route. In Kansas City, the young Walt began to study cartooning with a correspondence school and later took classes at the Kansas City Art Institute and School of Design.



Disney.
EB Inc.

In 1917 the Disneys moved back to Chicago, and Walt entered McKinley High, where he took photographs, made drawings for the school paper, and studied cartooning on the side, for he was hopeful of eventually achieving a job as a newspaper cartoonist. But his progress was interrupted by World War I, in which he participated as a truck driver for the American Red Cross in France and Germany.

Returning to Kansas City in 1919, he found occasional employment as a draftsman and inker in commercial art studios, where he met Ub Iwerks, a young artist who was to prove perhaps the most fortunate associate of his career after his brother Roy, who was his partner and the strongest counsellor throughout life.

First animated cartoons. Dissatisfied with their progress, Disney and Iwerks started a small studio of their own and acquired a second-hand motion-picture camera with which they made one- and two-minute animated advertising films shown on local movie-theatre programs, much as commercials are shown on television today. They also did a series of animated cartoon sketches called "Laugh-O-Grams" and a series of seven-minute animated fairy tales, which they called "Alice in Cartoonland." A New York film distributor cheated the young producers, and, destitute and disheartened, Disney left for Los Angeles to join his brother Roy.

With Roy as business manager, Disney resumed the "Alice" series, persuading Iwerks to join him and assist with the drawing of the cartoons. They invented a character called Oswald the Rabbit, contracted for distribution of the films at $1,500 each, and propitiously launched their small enterprise. Just before the transition to sound in motion pictures in 1927, Disney and Iwerks experimented with a new character—a cheerful, energetic, and mischievous mouse called Mickey. They planned two shorts, called *Plane Crazy* and *Gallopin' Gaucho*, that were to introduce Mickey Mouse when *The Jazz Singer*, a motion picture with the popular singer Al Jolson, brought the novelty of sound to the movies. Fully recognizing the possibilities for sound in animated-cartoon films, Disney quickly produced a third Mickey Mouse cartoon equipped with voices and music, entitled *Steamboat Willie*, casting the other two soundless cartoon films aside. When it appeared in 1928, *Steamboat Willie* was a sensation.

Introduction of Mickey Mouse

The following year Disney started a new series called "Silly Symphonies" with a picture entitled *The Skeleton Dance*, in which a skeleton rose from the graveyard and did a grotesque, clattering dance to the music of Saint-Saëns's *Danse macabre*. Original and briskly syncopated, the film launched the series most successfully, but with costs mounting because of the more complicated drawing and technical work, Disney's operation was continually in peril.

The growing popularity of Mickey Mouse and his girl friend, Minnie, however, attested to the public's taste for

the fantasy of little creatures with the speech, skills, and personality traits of human beings. (Disney himself provided the voice for Mickey.) This popularity led to the invention of other animal characters for the "Silly Symphonies." Donald Duck and the dogs Pluto and Goofy were introduced in 1931 and 1932, and in 1933 Disney produced a short, *The Three Little Pigs*, which arrived in the midst of the Great Depression and took the country by storm. Its treatment of the fairy tale of the little pig who works hard and builds his house of brick against the huffing and puffing of a threatening wolf suited the need for fortitude in the face of economic disaster; and its song "Who's Afraid of the Big Bad Wolf" was a happy taunting of adversity. It was in this period of economic hard times in the early 1930s that Disney fully endeared himself and his cartoons to audiences all over the world, and his operation began making money in spite of the Depression.

Through successive additions and advances in the animated-cartoon field, Disney continued to progress all through the 1930s. He had now gathered a staff of creative young people, who were headed by Iwerks. Colour was introduced in the "Silly Symphony" film, *Flowers and Trees* (1932), while other animal characters came and went in films such as *The Grasshopper and the Ants* (1934) and *The Tortoise and the Hare* (1935). Roy franchised tie-in sales with the cartoons of Mickey Mouse and Donald Duck—watches, dolls, shirts, and tops—and reaped more wealth for the company.

First feature-length cartoon

Walt Disney was never one to rest or stand still. He had long thought of producing feature-length animated films in addition to the shorts. In 1935 he began work on a version of the classic fairy tale that he called *Snow White and the Seven Dwarfs*, a project that required great organization and coordination of the creative and technical talents in his studio. Disney possessed a unique taste and capacity for such a task. While he actively engaged in all phases of creation in his films, he functioned chiefly as coordinator and final decision maker rather than as designer and artist. *Snow White* was widely acclaimed by critics and audiences alike as an amusing and sentimental romance. By animating substantially human figures in the characters of Snow White, the Prince, and the Wicked Queen and by forming caricatures of human figures in the seven dwarfs, Disney departed from the scope and techniques of the shorts and thus made a momentous transition in the nature of his type of film. While he continued for a while to do short films presenting the anthropomorphic characters of his little animals, he was henceforth to develop a wide variety of full-length entertainment films.

Snow White was followed three years later by other feature-length classics for children, *Pinocchio* (1940) and *Dumbo* (1941), the story of an elephant that could fly; and then Disney produced another totally unusual and exciting film—his multisegmented and stylized *Fantasia* (1940), in which cartoon figures and colour patterns were made to move to the music of Stravinsky, Paul Dukas, Tchaikovsky, and others.

But music critics and other intellectuals began to question Disney's taste and artistic talents in his more ambitious projects. They challenged him as a commercial opportunist and a mixer of artistic metaphors, an accusation that did not greatly disturb him, though Disney had by now acquired some of the smugness and intractability that many individuals do who achieve sudden success. In 1940 Disney moved his company into a new studio in Burbank, California, abandoning the old plant it had occupied in the early days of growth. The following year his staff went on strike. Although he and Roy held out against the artists, it was a blow to the image of Disney as a generous-hearted, amiable befriender of his own people.

Major film production. During World War II, the Disney studio did a great deal of work for the military and the federal government in the course of which it perfected the methods of combined live action and cartoon. After the war, Disney made many films with these hybrid techniques: *The Reluctant Dragon* (1941), *Saludos Amigos* (1942), *The Three Caballeros* (1944), *Make Mine Music* (1946), and *Song of the South* (1946).

The Disney studios were now established as a big business enterprise, and they began to produce a variety of entertainment films. One immensely popular series, called "True-Life Adventures," featured actual motion pictures of nature, but they were usually so trickily edited that, rather than being true-to-life documentaries, they betrayed the Disney tendency to fantasize. Among these were *Seal Island* (1948), *Beaver Valley* (1950), and *The Living Desert* (1953). They also turned to production of live-action fictional feature films; and more full-length animation romances, such as *Cinderella* (1950), *Alice in Wonderland* (1951), and *Peter Pan* (1953) and a flow of low-budget, live-action films, including *The Parent Trap* and *The Absent-Minded Professor* in 1961.

Films for television

The Disney studio was among the first to foresee the potentialities of television as a popular entertainment medium and to produce films directly for it. The *Zorro* and *Davy Crockett* series were fantastically popular with children and led to further profits for the company through tie-in sales of coonskin hats, powder horns, and Zorro capes. *Walt Disney's Wonderful World of Colour* became a continuing television fixture. But the climax of his career as a theatrical film producer came with his release of *Mary Poppins* in 1964. This adaptation of the popular children's story by Pamela L. Travers won worldwide acclaim and popularity.

Disneyland. Meanwhile, back in the early 1950s, Disney had initiated plans for a huge amusement park to be built near Los Angeles. When Disneyland opened in 1955, much of Disney's disposition toward nostalgic sentiment and fantasy was evident in its design and construction. It soon became a mecca for tourists from all over the world. A second Disney park in Florida was under construction at the time of his death, on December 15, 1966.

Reputation. In his later years, critical estimations of Disney and his works changed considerably, and his taste, as well as his political conservatism, were criticized by some. Social scientists and educators challenged the violence, cruelty, and sadism they found in many of his films, as well as the aesthetic vulgarities, and Disneyland was often referred to as "an amusement supermarket." But Disney himself, who in his last years contributed heavily with his family to the establishment of the California Institute of the Arts (called Cal-Art) in Valencia, California, was never visibly fazed by critics. "I've never called this art," he once commented with respect to all his activities. "It's show business, and I'm a showman."

MAJOR WORKS

FILM SHORTS: Included are 121 Mickey Mouse, 77 Silly Symphony, 126 Donald Duck, 51 Goofy, 48 Pluto, and 3 Figaro. The following in this category are all Academy Award winners: *Flowers and Trees* (1932); *The Three Little Pigs* (1933); *The Tortoise and the Hare* and *Three Orphan Kittens* (both 1935); *The Country Cousin* (1936); *The Old Mill* (1937); *Ferdinand the Bull* (1938); *The Ugly Duckling* (1939); *Lend a Paw* (1941); and *Der Fuehrer's Face* (1942).

FEATURE FILMS: *Snow White and the Seven Dwarfs* (1937); *Pinocchio* and *Fantasia* (both 1940); *The Reluctant Dragon* (1941); *Bambi* and *Saludos Amigos* (both 1942); *Victory Through Air Power* (1943); *The Three Caballeros* (1944); *Make Mine Music* and *Song of the South* (both 1946); *Fun and Fancy Free* (1947); *Melody Time* (1948); *So Dear to My Heart* and *The Adventures of Ichabod and Mr. Toad* (both 1949); *Cinderella* and *Treasure Island* (both 1950); *Alice in Wonderland* (1951); *The Story of Robin Hood and His Merrie Men* (1952); *Peter Pan, The Sword and the Rose*, and *Rob Roy: The Highland Rogue* (all 1953); *Stormy, the Thoroughbred with an Inferiority Complex* and *20,000 Leagues Under the Sea* (both 1954); *Davy Crockett, King of the Wild Frontier, Lady and the Tramp*, and *The Littlest Outlaw* (all 1955); *Westward Ho! The Wagons, The Great Locomotive Chase*, and *Davy Crockett and the River Pirates* (all 1956); *Johnny Tremain* and *Old Yeller* (both 1957); *The Light in the Forest* and *The Sign of Zorro* (both 1958); *Darby O'Gill and the Little People, The Shaggy Dog, Sleeping Beauty, The Third Man on the Mountain, Zorro the Avenger*, and *The Peter Tchaikovsky Story* (all 1959); *Toby Tyler or Ten Weeks with a Circus, Kidnapped, Pollyanna, The Hound That Thought He Was a Raccoon, Ten Who Dared Swiss*

Family Robinson, and *One Hundred and One Dalmatians* (all 1960); *The Absent-Minded Professor, Greyfriars Bobby, The Parent Trap,* and *Babes in Toyland* (all 1961); *Moon Pilot, The Prince and the Pauper, Bon Voyage!, Big Red,* and *In Search of the Castaways* (all 1962); *Son of Flubber, Savage Sam, The Miracle of the White Stallions, Yellowstone Cubs, Summer Magic, The Incredible Journey, The Three Lives of Thomasina,* and *Sword in the Stone* (all 1963); *The Misadventures of Merlin Jones, The Waltz King, The Moon-Spinners, Mary Poppins,* and *Emil and the Detectives* (all 1964); *Those Calloways, The Legend of Young Dick Turpin, The Monkey's Uncle,* and *That Darn Cat* (all 1965); *Jungle Book* (1966).

BIBLIOGRAPHY. RICHARD SCHICKEL, *The Disney Version* (1968), is the sole objective and definitive biography. DIANE DISNEY MILLER (with PETE MARTIN), *The Story of Walt Disney* (1957), is an affectionate daughter's reminiscences. R.D. FEILD, *The Art of Walt Disney* (1942), provides excellent technical analysis of stories and animation styles in the films up to the time of its publication; and JOHN HALAS and ROGER MANVELL, *Design in Motion* (1962), includes a well-informed discussion of Disney's work in a brief but commendable survey of film animation. DEEMS TAYLOR, *Walt Disney's Fantasia* (1940), is a pleasant and informative participant's account of the making and the contents of that film, with many colour illustrations.

(B.Cr.)

Disraeli, Benjamin

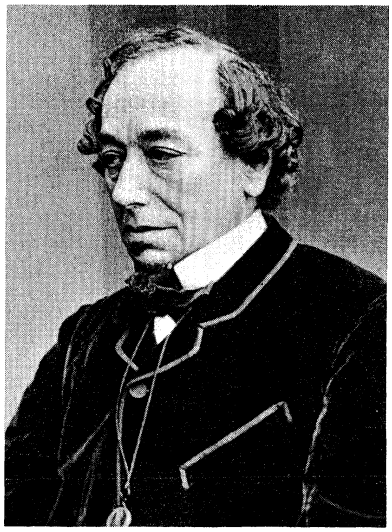
British statesman and novelist, twice prime minister, who provided the Conservative Party with a twofold policy of democracy and imperialism as well as an efficient organization, Benjamin Disraeli, later earl of Beaconsfield, was one of the most extraordinary figures ever to reach the summit of British politics. That a man of Jewish origin, loaded with debts, and widely regarded as a combination of flashy litterateur and opportunist adventurer should have become leader of the Tory (Conservative) Party may well be evidence of the unexpected fluidity of the Victorian social structure, but it is also evidence of a determination, a courage, and a parliamentary genius seldom surpassed in British history. If the "myth" of popular Toryism by which he interpreted past history can be easily destroyed, this does not alter the fact that he made it a reality in his own day. More than almost any other contemporary statesman he discerned intuitively what the newly enfranchised masses wanted both at home and abroad. His policy of social reform, his vision of the British Empire as a great confederation of self-governing states tied to the mother country by an imperial tariff, and his determination to stand up for what he deemed to be Great Britain's vital interests might well have ensured a long lease of power for the Conservative Party, had his health been stronger and his energy greater. The sincerity of his beliefs and the extent to which he was influenced by opportunism and love of power rather than high principle must always remain a matter of argument. He was an enigma to his contemporaries, and he remains one even today.

Early life. Disraeli was born in London on December 21, 1804, of Italian-Jewish descent, the eldest son and second child of Isaac D'Israeli and Maria Basevi. His grandfather Benjamin D'Israeli had emigrated to England from Cento, near Ferrara in Italy, in 1748 and, after a moderately successful business career, had become a member of the London Stock Exchange. His second wife, Disraeli's grandmother, was descended from the great Portuguese Jewish family of Vila Real. The most important event in Disraeli's boyhood was Isaac D'Israeli's quarrel in 1813 with the Sephardic synagogue of Bevis Marks, which led him in 1817 to decide to have his children baptized as Christians. Since, before 1858, Jews (by religion) were excluded from Parliament, it can safely be asserted that but for this curious accident Disraeli's subsequent political career could never have taken the form it did.

Disraeli was educated at small private schools. At the age of 17 he was articled to a firm of solicitors but longed to make his mark on the world in a more sensational manner than was open to a mere lawyer's clerk. His first efforts were disastrous.

In the autumn of 1824 he began to engage in reckless speculation in South American mining shares. The bubble burst a year later, and Disraeli was subsequently burdened with a debt that was to encumber him until he was well past middle age. Before this calamity he had persuaded his father's friend, the great publisher John Murray, to launch a daily newspaper, the *Representative,* which proved a complete failure. Disraeli was unable to pay his promised share of the capital, and the resulting recriminations involved him in quarrels with Murray and various other people of importance in the Tory literary-political world. As if this were not enough for a young man of barely 21, he then produced an anonymous novel, *Vivian Grey,* in which he lampooned Murray and related the whole story of the *Representative* under the thinnest of disguises. The author's identity was soon discovered, and the critics did not spare him.

By courtesy of the Gernsheim Collection, the University of Texas at Austin



Disraeli, albumen print by W. & D. Downey.

The strain of all these events induced in Disraeli something that would now be called a nervous breakdown. He did little of interest during the next four years apart from writing another equally extravagant novel, *The Young Duke,* and in 1830 he set off on 16 months of travel in the Mediterranean and the Middle East. Disraeli's experience of the colour and glamour of the East not only produced the Oriental descriptions with which his novels abound but also affected his attitude toward foreign affairs, and it is by no means fanciful to see in those experiences the origin of his policy toward India, Egypt, and Turkey in the 1870s. Levantine tour

On his return to England he at once plunged into London social and literary life. His dandified dress, his conceit and affectation, his exotic good looks made him a striking though not always popular figure wherever he went. He soon found himself beset with invitations to fashionable parties and met most of the celebrities of the day. During the next few years he published a number of books; the only one of any literary value was *Contarini Fleming,* which also has considerable autobiographical interest.

Political beginnings and marriage. Disraeli by 1831 had made up his mind to enter politics, and since his family was then settled at Bradenham, near Wycombe in Buckinghamshire, it was natural that he should seek a seat in that locality. As an independent radical, he fought and lost High Wycombe twice in 1832 and once in 1835. He then realized that he must attach himself to one of the great political parties, and, since his radicalism had many features not inconsistent with a somewhat eccentric interpretation of Toryism, he gravitated toward the Tory-Conservative Party and in 1835 unsuccessfully fought Taunton as the official Conservative candidate. Disraeli's extravagant behaviour, load of debts, and open liaison

with Henrietta, wife of Sir Francis Sykes (and the original of his novel *Henrietta Temple*), combined to give him a dubious reputation at this time. Finally at the election of 1837 he was returned to Parliament as Conservative member for Maidstone in Kent. His maiden speech in the House of Commons (December 7, 1837) was a failure. A combination of elaborate metaphors, affected mannerisms, and foppish dress resulted in his being howled down. But he was not silenced. The last words of his speech, shouted high above the clamour, were both defiant and prophetic: "I will sit down now, but the time will come when you will hear me."

Parliamentary debut

Disraeli was never a man to ignore the lessons of experience, and before long he became a speaker who commanded attention. His social position was consolidated by his marriage in 1839 to Mrs. Wyndham Lewis (*née* Mary Anne Evans). She was 12 years older than Disraeli and highly eccentric, but she had a life interest in a London house and in a fortune that produced some £4,000 a year and was deeply devoted to her second husband. It would be absurd to deny that Disraeli's motives in proposing were principally material, but there is no reason to doubt the truth of her own half-jesting remark: "Dizzy married me for my money but if he had the chance again he would marry me for love."

Breach with Peel. The Conservative leader, Sir Robert Peel, anxious to cultivate talent in his party, seems to have encouraged Disraeli, but when in 1841 the Conservatives won the election Disraeli was not given office. Much mortified at the rebuff, he began to assume an ever more critical attitude toward Peel and the type of Conservatism that he represented.

Disraeli's pen had not been idle during the past ten years. He had written a number of novels, of which *Venetia*, a fictional account of the lives of Lord Byron and Percy Bysshe Shelley, remains very readable and *Henrietta Temple* is one of the most attractive of all his romances. During these years, too, he wrote a series of political works, including *Vindication of the English Constitution* and *The Letters of Runnymede*, a series of pseudonymous letters, satirizing the characters of the principal Whig (the future Liberal) ministers. In the 1840s Disraeli's Toryism took on a new colour with the emergence of a group of youthful Tories (of which George Smythe, later Lord Strangford, and Lord John Manners, later duke of Rutland, were the most important members) who were nicknamed Young England and who looked to Disraeli for their inspiration. In 1844 he wrote *Coningsby; or, The New Generation*, the novel that immortalizes their attitudes and aspirations, and the hero is a portrait of Smythe. Young England was hostile to the cool pragmatic humdrum middle class conservatism of Peel. It was romantic, aristocratic, nostalgic, and escapist, looking to a nonexistent golden past in which the people and the nobility were united in a generous alliance supporting an enlightened throne and a conscientious church. A year later (1845) Disraeli published *Sybil; or, The Two Nations*, in which the causes and nature of Chartism, the working class political movement, were vividly analyzed and depicted.

The Young England group

In the autumn of 1845 the combination of the Irish famine and the arguments of Richard Cobden decided Peel to repeal the protective duties on foreign imported grain known as the Corn Laws. Here was an issue on which Disraeli, now in more or less open revolt against Peel, could rally not merely the romantic scions of noble houses but the great mass of country squires who formed the backbone of the Conservative Party. Acting ostensibly as lieutenant to Lord George Bentinck, who nominally led the rebels, Disraeli in a series of brilliant speeches consolidated the opposition to Peel. But his invective greatly embittered politics and created lasting resentment among the supporters of his former chief. Disraeli and the protectionists could not stop the repeal of the Corn Laws because the bill was backed by the Whigs, but they were able to put Peel in a minority on another issue, forcing him to resign (June 1846).

Conservative leader in the House of Commons. The Corn Laws crisis was a turning point in Disraeli's career. The loyalty of nearly all the Conservative former ministers to Peel and the death of Bentinck in 1848 gave Disraeli an indisputable claim to the leadership of the opposition in the House of Commons—a claim that even Lord Stanley, who led the Conservative Party as a whole, could hardly contest. Disraeli devoted the next few years to endeavouring to extricate his party from what he by then recognized to be "the hopeless cause of protection." The policy was sensible enough, but it inspired mistrust among his followers, as did Disraeli's pride in and insistence upon his Jewish ancestry. The party could not, however, dispense with his talents, whatever it thought of his character. Besides, with his election to Parliament as member for the county of Buckinghamshire in 1847 and his purchase of Hughenden Manor, near High Wycombe, in 1848, his social and political position was fortified. His finances, however, despite his father's death (1848) and his wife's income, remained shaky in the extreme.

The Whig government fell in 1852 and the Earl of Derby (as Lord Stanley had become since his father's death in 1851) formed a short-lived minority government. Disraeli was chancellor of the exchequer despite his protest that he knew nothing of finance. His budget in fact brought the government down (December 1852), but Disraeli can hardly be blamed for this. The free-trade majority in the house was determined to defeat measures that relieved agriculture, even though the method chosen did not involve protection; and yet Disraeli could not escape bringing forward some such proposals if he was to placate his own followers. For the next six years the Tories were in opposition. In 1858, however, Derby once again formed a minority government with Disraeli as chancellor of the exchequer. Disraeli with his notions of popular Toryism had for some time felt that there was no reason to leave parliamentary reform as a Whig monopoly. Accordingly he introduced a moderate reform bill in 1859. But its provisions seemed too obviously designed to help his own party, and the majority of the house combined to defeat them and turn out the government. Again the Tories were out of office for six years.

Chancellor of the exchequer

In 1865 the Whig-Liberal leader Lord Russell brought forward a measure of parliamentary reform. Moderate though this was, it aroused the strong opposition of the Tories and, more seriously, a revolt within his own party. Russell was defeated, and Derby formed his third minority government with Disraeli again chancellor of the exchequer. Disraeli's opposition to the Liberal bill had been based on opportunism rather than principle. Although the initiative for a new Conservative reform bill came from Queen Victoria and Lord Derby, Disraeli plunged into the battle with an enthusiasm combined with a mastery of parliamentary tactics unsurpassed by any other statesman of the day. He believed strongly that the measure should be a sweeping one, subject to certain safeguards, and he was determined that the bill should be carried by a Conservative government. But with the Liberals in a majority in the House of Commons, Disraeli was obliged to accept their amendments, and in the end nearly all the safeguards intended to protect Conservative interests vanished. The resulting measure doubled the existing electorate and was undoubtedly far more democratic than most Conservatives had envisaged. It was indeed "a leap in the dark," as Derby described it; but Disraeli could fairly claim that the measure had gone far toward "realizing the dream of my life and re-establishing Toryism on a national foundation."

The "top of the greasy pole." In February 1868 Derby retired from politics and Disraeli became prime minister. "Yes," he said in reply to a friend's congratulations, "I have climbed to the top of the greasy pole." It was in a sense a "caretaker" premiership, for everyone awaited the general election that was due to be held as soon as the new electoral register was completed, and the autumn election, as expected, was won by the Liberals. Disraeli set a precedent by resigning without waiting for Parliament to meet.

"Caretaker" premiership

The whole face of politics had changed in the three years since the death of Lord Palmerston, the great Whig prime minister, who had dominated political life in his last years. From a chaotic collection of ill-defined shifting groups influenced as much by personalities as politics, there emerged two great parties with intelligible coherent policies. Their respective leaders, Disraeli and W.E. Gladstone, seemed singularly well chosen to polarize the confused forces that had hitherto animated British politics. For 12 years they fought a deadly duel. Each detested the other, and history records scarcely a flicker, on either side, of those generous emotions toward the political enemy that have so often softened the seeming acerbity of parliamentary life.

For the moment Disraeli deemed it wise to play a comparatively peaceful role, gradually creating for the Conservative party a new image that would in the course of time, he hoped, persuade the new electorate. His seeming apathy, however, disturbed his supporters; and misgivings were increased by the publication of *Lothair*, described by G.E. Buckle, Disraeli's biographer, as "a gaudy romance of the peerage, so written as to make it almost impossible to say how much was ironical or satirical, and how much soberly intended." To many grave persons it must have seemed an undignified production for a former prime minister.

From 1872 onward, however, Disraeli took charge of his party with a firm hand. On three broad issues he sharply differentiated Conservative from Liberal policy. In a speech in April 1872 he defended the monarchy, the House of Lords, and the church against the radical threat that he claimed to discern in Gladstone's policy. Three months later he asserted for the first time the Conservative belief in consolidating the empire, with special emphasis on India. And in the same speech he dwelled on the importance of social reform: his old "condition of the people" question, which he had earlier analyzed so vividly in *Sybil*. During these years of opposition, yet a fourth distinctive Conservative policy began to emerge: the belief in a strong foreign policy, especially toward Russia, and in the greatness of Britain.

Bereavement

In December 1872 Disraeli suffered a personal bereavement none the less poignant for having been long expected. His wife died of cancer after many months of illness. Her oddities never disturbed Disraeli, while her affection, her management of domestic affairs, and her deep devotion to him had rendered his life happy as it had never been before. There were more material losses too. He no longer had her house in London, and her fortune, which was entailed, passed to her cousins. Disraeli was now 68 and his health was not good. It is a testimony to his indomitable will that he determined to continue the political battle despite the personal distress and confusion into which he was thrown. The society of women remained essential for him. From this period dates his romantic friendship with the sisters, Lady Bradford and Lady Chesterfield, to whom he poured out the secrets of politics and the affections of his heart in a profuse and fascinating series of letters that only ceased with his death.

At this juncture public affairs at last began to move in his favour. Gladstone's ministry was defeated in the House of Commons in 1873. Gladstone promptly resigned, but Disraeli shrewdly refused to take office on the ground that he could not dissolve Parliament for some months owing to the amount of uncompleted business and that a minority government could only damage his party's prospects. Gladstone reluctantly returned to office, but less than a year later himself decided on a sudden dissolution. Disraeli had not neglected the vital problems of party organization and machinery so necessary in the new era of mass democracy. The result was a triumphant conservative electoral victory.

Second administration, 1874–80. It was Disraeli's tragedy that power came to him too late. He had never possessed the vitality and the iron constitution of statesmen such as Gladstone, and he aged rapidly during his second premiership. Certainly he achieved much, but he could have done far more, even a few years earlier. He formed a strong cabinet and was further fortified by his friendship with the Queen, who was politically conservative and had already acquired a considerable dislike for Gladstone. Disraeli treated her as a human being, whereas Gladstone treated her as a political institution. In doing so he indulged in a degree of flattery to which the word Oriental has been legitimately applied.

Friendship with Queen Victoria

Embarking on a program of social reform that touched the real interests of the masses more effectively than most of Gladstone's measures, Disraeli was at last able to show that Tory democracy was not a mere phase. The Artizans' and Labourers' Dwellings Improvement Act was the first measure to make effective slum clearance possible. The Public Health Act of 1875 codified the complicated law on that subject and translated into reality Disraeli's own half-joking witticism in a speech of 1872: "Sanitas sanitatum, omnia sanitas" (an alliterative misquotation of *vanitas vanitatum, omnia vanitas*, "vanity of vanities, all is vanity," taken from the Book of Ecclesiastes). Equally important were an enlightened series of factory acts preventing the exploitation of labour and two trades union acts that clarified the doubtful legal position of those bodies.

Important though Disraeli's domestic program was, it took second place in the public eye to his imperial and foreign policy. His first great success was the affair of the Suez Canal shares. The extravagant and spendthrift ruler of Egypt, the khedive Ismāʿīl Pasha, owned slightly less than half the Suez Canal Company's shares and was anxious to sell them. This fact was communicated by a patriotic journalist to the Foreign Office, which regarded a purchase as absurd. Disraeli, however, at the end of November 1875 boldly bought them for the government, the money being put up by his friends the Rothschilds until Parliament could confirm the bargain. Financially it was an excellent bargain, and in terms of imperial prestige it seemed a notable triumph.

Disraeli's next effort, early in 1876, was to bring in a bill conferring on Queen Victoria the title empress of India. The measure passed, but there was a good deal of opposition; Disraeli himself would gladly have postponed it, but the Queen insisted. For some time his health had been poor, and it became clear to him in the summer of 1876 that the strain of leading the House of Commons was too much. Accordingly in August he accepted a peerage, taking the title earl of Beaconsfield.

From 1876 to 1878 his life was dominated by a major issue of foreign policy. The conflict between Russia and Turkey that had lain dormant since the end of the Crimean War in 1856 was abruptly reopened by the revolt of the Christian subjects of the Ottoman Empire against intolerable misrule. Russia declared war on Turkey in April 1877, and its troops reached the gates of Constantinople early the next year. Their victories renewed British fear for the safety of the route to India. Disraeli, knowing that the Russian forces were exhausted, rightly assumed that the threat of British intervention would be enough. As a result of his actions, Russia was obliged to submit the highly pan-Slavist Treaty of San Stefano, which it had forced upon Turkey, to a European congress. Beaconsfield attended the Congress of Berlin in the summer of 1878 and made a great impression, obtaining nearly all the concessions he wanted. He was able to return in triumph to London, declaring that he had brought back "peace with honour."

"Peace with honour"

This was the climax of his career. The Queen offered him a dukedom, which he refused, and the Order of the Garter, which he accepted. His prestige was at its height. Thereafter his fortunes waned. The aggressive policy of Lord Lytton, the viceroy of India, brought disaster in Afghanistan, and in South Africa an unwary British force was slaughtered by the Zulus. Agricultural distress combined with an industrial slump to bring the government into disfavour. At the general election of 1880 the Conservatives sustained a heavy defeat. Beaconsfield courageously agreed to retain the party leadership and despite this burden found time to finish *Endymion*, a mellow political novel in which he surveyed his own early career through the rose-coloured spectacles of

romantic nostalgia. His health was, however, failing rapidly, and he died at his London home on April 19, 1881. A few days after his burial in the family vault at Hughenden, Queen Victoria came in person to lay a wreath upon the coffin of her favourite minister.

MAJOR WORKS

NOVELS (MAINLY POLITICAL): *Vivian Grey*, 5 vol. (1826–27); *The Young Duke*, 3 vol. (1831); *Contarini Fleming*, 4 vol. (1832); *Henrietta Temple*, 3 vol. (1837); *Venetia*, 3 vol. (1837); *Coningsby; or, The New Generation*, 3 vol. (1844); *Sybil; or, The Two Nations*, 3 vol. (1845); *Tancred; or, The New Crusade*, 3 vol. (1847); *Lothair*, 3 vol. (1870); *Endymion*, 3 vol. (1880).

POLITICAL WRITINGS AND SATIRE: *Vindication of the English Constitution* (1835); *The Letters of Runnymede* (1836); *The Spirit of Whiggism* (1836); *Lord George Bentinck: A Political Biography* (1852).

MISCELLANEOUS POEMS: *The Revolutionary Epick* (1834); *The Tragedy of Count Alarcos* (1839).

BIBLIOGRAPHY. *Novels and Tales by the Earl of Beaconsfield*, Hughenden edition, 11 vol. (1881); *Lord Beaconsfield's Letters, 1830–1852*, ed. by R. DISRAELI (1887, reprinted 1928); *Letters from Benjamin Disraeli to Frances Anne, Marchioness of Londonderry, 1837–1861*, ed. by the MARCHIONESS OF LONDONDERRY (1938); *The Letters of Disraeli to Lady Bradford and Lady Chesterfield*, ed. by the MARQUIS OF ZETLAND, 2 vol. (1929); *Selected Speeches of . . . the Earl of Beaconsfield*, ed. by T.E. KEBBEL, 2 vol. (1882). The official biography is W.F. MONYPENNY and G.E. BUCKLE, *The Life of Benjamin Disraeli, Earl of Beaconsfield*, 6 vol. (1910–20; rev. ed., 2 vol., 1929, reprinted 1968). See also ANDRE MAUROIS, *La Vie de Disraeli* (1927; Eng. trans., rev. ed. 1947); ROBERT BLAKE, *Disraeli* (1966) and *Disraeli and Gladstone* (1970); and RICHARD A. LEVINE, *Benjamin Disraeli* (1968), a study of the novels, particularly *Coningsby*, *Sybil*, and *Tancred*. A popular biography is HESKETH PEARSON, *Dizzy* (1951). For his early life, see B.R. JERMAN, *The Young Disraeli* (1960). The best short life is HAROLD BEELEY, *Disraeli* (1936).

(B.)

Distilled Liquor

Distilled liquors are all alcoholic beverages in which the concentration of ethyl alcohol (the intoxicating agent) has been increased above that of the original fermented mixture. The production of distilled liquor is based upon fermentation, the natural process of decomposition of excess vegetative materials. Fermentation occurs in nature whenever the two necessary ingredients, carbohydrate-bearing vegetation and yeast, are available. Yeast, a vegetative micro-organism living and multiplying in media containing carbohydrates and particularly in simple sugars, has been found throughout the world, including frozen areas and deserts.

The principle of alcoholic distillation is based upon the different boiling points of alcohol (78.5° C, or 173.3° F) and water (100° C, or 212° F). If an alcohol-containing liquid is heated to a temperature above 78.5° C but below 100° C, the alcohol will vaporize and separate from the original liquid and can be gathered and recondensed into a liquid of much greater alcoholic strength.

History. Because the two ingredients necessary to alcoholic fermentation are widely spread and always appear together, civilizations in almost every part of the world early developed some form of alcoholic beverage. By 800 BC the Chinese were distilling a beverage from rice beer, and in the East Indies arrack was distilled from sugarcane and rice. The Arabs developed a distillation method which they used to produce a distilled beverage from wine. Greek philosophers reported a crude distillation method, and the Romans apparently produced distilled beverages, but no reference to these has been found in their writings before AD 1000. Production of distilled liquors was reported in Britain before the Roman conquest, but it is likely that the metal containers and metal-working techniques brought by the Romans greatly accelerated the development of distillation operations. Spain, France, and the rest of western Europe may have produced distilled liquors at an earlier date, but production apparently gained impetus in the 8th century, after contact with the Arabs.

The first distilled liquors were made from natural sugar-based materials, the two most important of which were grapes—producing brandy—and honey—the raw material of distilled mead. The earliest use of starchy grains in distilled-liquor production is not known, but their use was flourishing by the late Middle Ages and, by the middle of the 17th century, was of sufficient importance to attract widespread government control. The idea of rectification, or distilling in such a way as to divide the fermented mixture into various liquid fractions, first appeared in France at the beginning of the 16th century (this important process is described in detail below). By the early 19th century large-scale continuous stills, basically identical with those used in industry today, were operating in France and England. As production methods improved, the distilled-liquor industry became a source of government revenue, often with rigid controls on both production and sale.

Table 1 shows the orgin of some distilled liquors.

Table 1: Origins of Various Distilled Liquors

	date	raw material	fermented liquor	distillate
China	—	rice and millet	tchoo	sautchoo
Ceylon and India	800 BC	rice and molasses or palm sap	toddy	arrack
Asiatic Tatary	—	cow's milk	kefir	arika
Caucasia	—	mare's milk	koumiss	skhou
Japan	—	rice	sake	sochu
Britain	AD 500	honey	mead	mead distilled
Italy	1000	grapes	wine	brandy
Ireland	1100	oats and barley malt	beer	usquebaugh
Spain	1200	grapes	wine	*aqua vini*
France	1300	grapes	wine	cognac
Scotland	1500	malted barley	beer	aqua vitae, whiskey

DISTILLED LIQUOR PRODUCTION

Raw materials. In general, the raw material for a distilled liquor is a natural sugar—such as those found in honey, ripe fruit, sugarcane, beetroot, and milk—or a substance of an amylaceous (starchy) nature that may be easily converted into a sugar. This change is simply effected, for some of the active agent is already present in cereals containing the necessary starch or is easily developed in them; when it is absent, as in potato starch, a suitable malted cereal can be added.

Enzyme activity

The active agents that break down starches into sugars and sugars into alcohol are known as enzymes; these active proteins are catalysts, and even small amounts will cause a fundamental change in large quantities of the material being processed. Enzymes are easily poisoned by certain substances, such as hydrocyanic acid, mercury salts, and Formalin; are sensitive to temperature variations (being rendered inactive in certain conditions); and are specific in action in that each is effective for certain classes of substance only. They convert complex insoluble material into simpler assimilable substances: amylase, for instance, will reduce starch to dextrin and then to the sugar maltose; maltase will reduce maltose to the simpler sugar dextrose; and zymase will ferment the dextrose to alcohol and carbon dioxide. All these changes take place in aqueous solution.

Sugary materials. Grapes, cultivated in most of the subtropic and warm temperate zones of the world, are the major fruit employed as the raw material of distilled liquor, and the final product of their fermentation is brandy. Other natural fruits, such as apples and peaches, are used to a lesser extent, and many fruits are limited to local importance. Grapes made into wine for use in brandy production require only simple crushing and pressing. Apples and citrus fruits are put through crushers, and the fermentable juices are either pressed out for fermentation, as in apples, or the entire mass is fermented.

Sugary vegetables include sugarcane, sugar beets, and *Agave tequilana* (a type of cactus). Sugarcane and its products, including cane juices, molasses, and sugar, are the most important of the vegetable group. Grown

throughout the tropics and semitropics, sugarcane is used in making rum and an alcohol derived from rum. Sugarcane juice can be pressed from the cane for use as the base raw material for fermentation, or the juice may be concentrated for sugar production, with the molasses residue from the sugar crystallization used as a base for fermentation. This process is also applied to sugar beets.

Starchy materials. For many centuries, it was only feasible to employ local grain crops for liquor production, and, in this way, the basic characteristics of the local distilled beverage was established. Improved transportation has removed this restriction, and today economic considerations frequently determine grain selection, with the principal grain used being the one available at the lowest price per unit of fermentable materials.

Cereal grains used for liquor production

Corn (maize) is the most important cereal grain employed; principal producing countries include the United States, Argentina, Brazil, Mexico, South Africa, Thailand, and France. Rye grain, though less efficient in fermentation than corn, is used extensively in whiskey production, primarily for the flavour characteristics it imparts to the final product. It is particularly employed in Canada and the United States, accounting for as much as 25 percent of total grains used. Rice, a widely grown cereal, has limited use in distilled liquor production outside of eastern and southeastern Asia from India to Japan. Barley grain, probably the first cereal employed for distillation in large quantities, was formerly a major crop throughout Britain, Scotland, Ireland, and western Europe. Wheat, because of its high cost, is used only where corn is in short supply and is then limited to production of grain alcohol for blending or in production of liqueurs. Potatoes have been used in distilled-liquor production primarily in central Europe; in the tropics, certain starchy roots are also employed.

Preparation for fermentation, similar for all grains, involves grinding and mashing. In grinding, the grains are reduced to a coarse meal in various types of mills to allow thorough wetting of their starch cells. In mashing, the starch cells of the grain, enclosed in their own protective coatings, are broken to allow wetting and liquefaction of the entire starch mass. Mashing usually begins with the grain most difficult to treat. When corn is used, the ground meal is wetted at temperatures approximately 150° F (66° C), and the temperature is then raised to boiling or sometimes higher while under pressure. Temperature is reduced when the starch cells are broken. The grain ranking second in cell resistance (usually rye) is added next. Temperature is again reduced before ground malt meal, either in dry form or as a water slurry (insoluble mixture), is added. The time and temperature for each operation varies according to the grain used and the final distillates desired. The completed mashing process produces a mixture in which the starches have been converted to fermentable sugars, suitable for utilization by the yeast (see below). The sugars, principally dextrose and maltose, vary in concentration among producers but, generally, are sufficiently concentrated to make a final product ranging from 7 to 9 percent alcohol. Other starchy substances, such as potatoes, are usually crushed and heated, exploding the starch cells; the liquefied starch obtained is converted to fermentable sugar by the action of amylase enzymes.

Use of malt

Malt. Malt, a major raw material in distilled-liquor production, consists of grain, softened by steeping in water and allowed to germinate; it can be produced from any cereal grain. As the grain germinates for the production of a new plant, changes take place within the kernel. Embryo development begins, and enzymes within the kernel initiate the process of making carbohydrate, protein, and other constituents into a soluble substance available to the embryo. Although many enzymes affect this process, the most important for distilled-liquor production are the amylases, enzymes converting the high-molecular-weight carbohydrates into simple soluble form for utilization by the embryo. This process also allows the embryo to develop a root structure capable of taking needed nutrients from the soil. The fermenting industry employs the same process, under controlled conditions, to produce malt. In malting, grain is thoroughly soaked in water, then allowed to germinate until the embryo has partially developed. At this stage the enzymes, particularly amylases, have increased beyond the amounts necessary for conversion of the starch within the individual grain. After partial germination, the grain is dried at low temperatures. The resulting malt contains sufficient enzymes to convert approximately ten times its weight in other unmalted grains (see also BREWING).

Barley is the grain most commonly used to produce malt for the distilled-liquor industry. Of the two amylases—alpha and beta—developed during the malting process, the alpha is the most important for conversion of other grains. In addition to converting starches from other carbohydrates to sugars, barley malt contains soluble proteins (amino acids), contributing flavour to the distillate secured from fermentation and distillation of grain–malt mixtures.

Fungal amylase, less important to distilled-liquor production, is derived from mold growth. Although producing high concentrations of amylases, this type of enzyme conversion is used primarily in alcohol production and not for the distilled liquors, which derive their flavour from grain mixtures.

Yeast and yeast culture. As mentioned above, yeasts are found throughout the world; more than 8,000 strains of this vegetative micro-organism have been classified. Approximately nine or ten pure strains, with their subclassifications, are used for fermentation of grain mashes; these all belong to the type *Saccharomyces cerevisiae*. Each strain has its own characteristics, imparting its special properties to the distillate derived from its fermentation. A limited number of yeasts, all in the classification *Saccharomyces ellipsoideus*, are used in the fermentation of wines, from which brandy is distilled. Strains used in the fermentation of grain mashes are also used in fermentation for rum and beer production.

In grain-based products, yeast cells are grown in grain mixtures. The preparation of a cooked mash of rye and barley malt is most common. The mash is sterilized, then inoculated with lactic acid bacteria to increase acidity. When the desired acidity is reached, the mixture is again sterilized and a pure yeast culture is added. The yeast has been grown under controlled conditions until it reached the optimum point for mixing with the grain mash. In liquid fermentation, as from fruits and sugarcane, the yeast is generally grown in a mixture similar to the one it will be used to ferment; for example, a yeast culture to be used for molasses fermentation is usually grown in molasses.

Fermentation. In the fermentation process, simple sugars, including dextrose and maltose, are converted to ethyl alcohol by the action of yeast enzymes. Several intermediate compounds are formed during this complex chemical process before the final ethyl alcohol is obtained (see ALCOHOLS, PHENOLS, AND ETHERS: *Alcohols*).

Yeast functions best in a slightly acid medium, and the prepared grain mash, fruit juice, molasses, or other mixture must be checked for adequate acidity (pH value). If acidity is insufficient, acid or acid-bearing material is added to achieve the necessary adjustment. The previously prepared yeast is then added, and final dilution of the mixture is made. The final concentration of sugars is adjusted so that the yeast fermentation will produce a finished fermented mixture containing between 7 and 9 percent alcohol.

Equipment for fermentation

Commercial fermentation is carried on in large vats. Although open vats may be used, closed vats are preferred for sanitary purposes and are easier to equip for temperature control. The time required for completion of fermentation is mainly dependent upon the temperature of the fermenting mash. Normal yeast is most effective in breaking down all of the fermentable sugars at temperatures ranging from 75° to 85° F (24° to 29° C), and, in this range, completion of fermentation requires from 48 to 96 hours. Fermentation at lower temperatures requires longer periods. The mash is ready for distillation upon completion of fermentation and, if allowed to continue past this period, will be adversely affected by

bacterial action beginning conversion of the ethyl alcohol to acids, reducing the alcoholic content and influencing the flavour and aroma of the finished product.

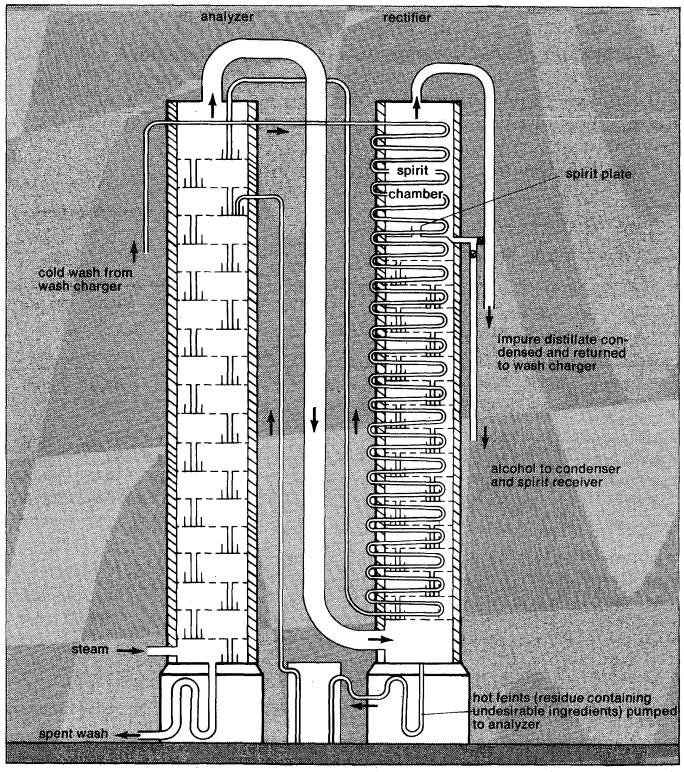
Distillation. As mentioned above, the difference in the boiling points of alcohol and water is utilized in distillation to separate these liquids from each other. Basic distillation apparatus consists of three parts: the still or retort, for heating the liquid; the condenser, for cooling the vapours; and the receiver, for collecting the distillate.

Equipment. Early distillation methods were crude, and satisfactory equipment for large-scale commercial use was not developed until after the year 1000. The first crude pot still was based upon primitive methods. The simple pot still is a large enclosed vessel, heated either by direct firing on the bottom or by steam coils within the vessel, with a cylindrical bulb at its top leading to a partially cooled vapour line. The bulb and vapour line separate entrained liquid particles from the vapour on its way to the final condenser. The usual pot-still operation involves a series of two or three pot stills. The vapour from the first is condensed; any vapour falling below a predetermined alcoholic content is fed into a second still; and condensed vapour from the second still falling below the required alcoholic content is fed to the third. The condensed vapours of the desired alcoholic content from all three stills are comingled in a single receiving container.

The pot still

The pot still, used primarily in Scotland and Ireland for whiskey production and in France for brandies, has had only brief use in distilled-liquor production elsewhere and is gradually becoming obsolete. Even in countries in which the pot still has long been used, it has been replaced by continuous distillation for the major portion of alcoholic-liquor production, and its current use is limited to production of flavouring whiskies and other flavouring ingredients.

The continuous still, coming into use in the early 19th century, consists of a tall cylindrical column filled with perforated plates (see illustration). These plates serve as a series of small pot stills, one on top of the other. Live steam, used as the heat source, is fed to the bottom of the still, and the liquid to be distilled is fed near the top. Steam pressure holds the liquid on the plates, and, with any overflow caught by the plate below, the liquid level on each plate is maintained. Use of a sufficient number of plates assures that the concentration of alcohol in the vapour leaving the top of the still will be appropriate for the desired product and that the liquid leaving the bottom of the still has been stripped of any alcohol.



Continuous still for preparing distilled liquors.

Many distillation operations combine column and pot stills. The condensed distillate from the column still is fed to the doubler, a kind of pot still heated by closed steam coils, and redistilled.

The rectification process. Rectification is the process of purifying alcohol by repeatedly or fractionally distilling it. It was mentioned above that a fermentation mixture primarily contains water and ethyl alcohol and that distillation involves increasing the percentage of ethyl alcohol in the mixture. Water vaporizes very easily, however, and, unless care is taken, the distillate of a fermentation mixture will contain unacceptably large quantities of water. Hence, the column through which the vaporized fermentation mixture passes is provided with baffles onto which the water vapour condenses, while the alcohol vapour passes through. In this way, the distillate obtained will have a high alcohol content. The fermentation mixture furthermore contains small quantities of complex constituents known as congeners or congenerics that give each distilled liquor its individual character. These congeners, described below in greater detail, have lower boiling points than water and thus also pass through the distillation column into the distillate. Discovery of this process of rectification permitted production of distilled liquors as they are known today; it is used widely in other industries.

Maturation and storage. The products obtained directly from the distillation of fermented materials are usually undesirable for human consumption and require further processing or maturation. The maturation process involves a series of chemical reactions producing changes in the original distillate that improve the aroma and taste of the final product. Some changes result from reactions of compounds after long exposure to each other within the distillate itself; others result from oxidation of some of the compounds because of the presence of small quantities of oxygen taken from the air; still other changes result from the presence of extracts from the wood of the barrel, either as direct flavouring ingredients or produced by chemical reaction between the barrel extracts and some of the ingredients in the distillate. Generally, a longer maturation period is required when the concentration of heavy compounds is high, and distillates produced at very high proof (alcohol content; see below) require a much shorter maturation period than those produced at lower proofs. Maturation systems vary according to the type of alcoholic liquor.

The maturation of whiskies falls into two categories, according to the type of container employed. Sometimes in Canada, and by law in the United States, charred new white-oak containers are used for maturation of products to be called straight bourbon or straight rye whiskey. These containers, holding approximately 50 gallons or about 200 litres, are usually stored in dry warehouses, often having controlled temperature and humidity. During maturation, wood extracts, caramelized wood, and other ingredients are extracted from the container; maturation also results from contact of oxygen from the outside air with the ingredients in the alcoholic mixture. White oak is one of the few woods holding liquids while allowing the process of breathing through the pores of the wood. This breathing process is caused by temperature changes of the liquid in the barrel. Temperature increases cause pressure to develop within the barrel, forcing some of the liquid into the barrel staves; temperature reduction decreases internal pressure, bringing part of this liquid back into the main body of liquid along with a small amount of air. The more frequently this breathing process occurs, the more rapidly maturing progresses. Modern warehouses are often mechanically equipped to allow control of the heating and cooling cycle.

The breathing process

Outside the United States, reused cooperage is common. Since used containers have already yielded their initial

oak extracts, the resulting product is low in extracted flavouring ingredients, which is desirable in some beverages. This maturation method, typified by Scotch and Irish whiskies, can be carried on in casks holding up to 132 gallons or about 500 litres. These casks have usually had previous use for storage or maturation of other whiskies or wines and may be reused for many maturation cycles. Maturation in dry warehousing increases alcoholic content of the liquid in the container, but the more common practice for Scotch and Irish whiskies of maturation in high humidity warehouses reduces alcoholic concentration.

The maturation procedure for brandies is similar to that of some whiskies, but the brandies are usually matured in fairly large casks or oak containers. Most brandies are matured for three to five years, but some remain for as long as 20 to 40 years or even longer.

Rum is usually matured in reused oak containers; high concentrations of oak extracts are not considered desirable. Normal maturation time is two to three years, but rum, generally a blended product, may contain a percentage of older rums.

In distilled liquors, the term age refers to the actual duration of storage, while maturity expresses the degree to which the chemical changes described above have taken place. Most governments specify storage time for various products. The United States requires a two-year storage period for most whiskies but has no requirement for any pure alcohol or neutral spirits (close to 100 percent alcohol) added to such whiskies in the production of blended whiskey. Canada requires storage of two years for all distilled liquors; Scotland and England require a three-year storage; while Ireland requires five years for all products classified as whiskey but has no requirements for vodka and gin.

Packaging. Distilled liquors react upon exposure to most substances, extracting materials from the container which tend to destroy the liquor aroma and flavour. For this reason, glass, being nonreactive, has been the universal container for packaging alcoholic liquors. Packaging economics require containers that are standardized in size and shape and that lend themselves to automatic processes. In the early 1970s suitable plastics became available and were approved by food and health departments in many countries. Containers made from such plastics, however, were not economical for use.

Mechanized packaging

Early hand methods of filling, labelling, corking, and other operations have been replaced by highly mechanized bottling lines, with bottles cleaned, filled, capped, sealed, labelled, and placed in a shipping container at a rate as high as 240 bottles per minute. This progress became possible with the development of high-strength glass, plastic caps with inert liners, and high-speed machines. Even specialized packaging, long a hand operation, has been replaced by standardization of containers, allowing production on automatic lines.

COMPOSITION OF DISTILLED LIQUOR

Though the major components of distilled liquors are ethyl alcohol (C_2H_5OH) and water, a simple combination of these two substances does not produce an attractive beverage. The minor constituents or congeners mentioned above, including higher alcohols, aldehydes, ethers, esters, volatile acids, furfural (a liquid aldehyde of penetrating odour, usually made from plant materials), and other organic compounds influence the character, aroma, and flavour of the beverage out of all proportion to their quantities. As they are too great or too small, they invest the liquor with fuller or feebler flavour than may be desired.

Proof systems. Mixtures of alcohol and water (such as are found in distilled liquor) have values intermediate between those of the pure components of some of their physical characteristics—such as viscosity, heat of vaporization, specific heat, thermal expansion, vapour pressure, electrical conductivity, boiling point, refractive index, and specific gravity. Although the variations may not be strictly proportional to the amount of alcohol present, they are nevertheless generally progressive for certain ranges; this fact has provided some convenient physical means of determining spirit strength, and it is common practice to derive such information from specific gravity (the ratio of the density of one substance to the density of another substance when both are weighed in air). Provided that the liquid is free of or has been freed by distillation from saccharine and other matter that would interfere, the specific gravity is determined by immersing in it a hydrometer, which may be graduated (1) in terms of specific gravity; (2) with arbitrary scales (United Kingdom, Soviet Union); (3) in percentages by volume of alcohol (Belgium, France, Italy, Norway, Sweden); (4) in percentages by weight of alcohol (Germany); (5) in percentages by volume of proof spirit (United States), as described below; and (6) with a scale based upon fractions of the volume of the instrument (The Netherlands). In all cases, these instruments are calibrated at standard temperatures; when the liquid to be tested is cooler or warmer (as it usually is), reference is made to tables to ascertain the true value for the spirit at the standard temperature.

Determining specific gravity

Spirit strength may be designated in several ways—percentage by weight, weight per gallon, gallons per hundredweight, or percentage by volume, all these having reference to absolute (*i.e.*, pure) alcohol and water. But there are other standards in common use; *e.g.*, United States proof spirit, which is 50 per cent by volume of alcohol and water, and British proof spirit, which is of such a strength that at 51° F (11° C) its weight is $12/13$ of an equal volume of water. Thus in the United States, each degree of proof represents 0.5 percent of alcohol, so that a liquor having 50 percent alcohol is termed 100 proof. British proof is based on a specific concentration of alcohol, a 50 percent alcoholic content being equivalent to 114.12 United States proof. In the United States proof system, distilled liquors are expressed in direct proof figures; in the British system, distilled liquors are expressed as being over or under 114.12 proof. The metric Gay-Lussac system simply states the percentage by volume of alcohol in a distilled liquor.

Classification of distilled liquors. Chemical characteristics of distilled liquors vary widely, and a detailed chemical analysis cannot be spelled out for any one type or group. Most distilled liquors, especially those produced from natural fermentations, contain a large number of chemical compounds. Table 2 lists the most widely distributed distilled liquors, with the chemical analysis of each expressed in industry terms, including alcohol, acids, esters, fusel oils, and solids.

Table 2: Distilled Liquor Analyses
(grams per 100 litres)

	bourbon		Canadian	Scotch	Irish	blend	brandy
	bottled in bond	straight					
Percentage of alcohol by volume	50.0	43.0	40.0	43.0	43.0	43.0	40.0
Total acids	79.2	67.2	26.4	12.0	16.8	67.2	33.6
Esters	60.0	51.2	14.1	17.6	12.3	22.9	30.0
Total fusel oil	165.0	145.0	34.9	76.1	113.0	78.9	150.0
Solids	189.6	167.2	205.2	58.0	99.2	180.0	550.0

BIBLIOGRAPHY. HAROLD J. GROSSMAN, *Grossman's Guide to Wines, Spirits, and Beers*, 4th rev. ed. (1964), a popular work prepared for both the trade and the consumer that includes definitions pertaining to all alcoholic beverages, an excellent section on both distilled liquor in general and specific types, and chapters on mixed drinks, beverage service, and merchandising; KARL M. HERSTEIN and MORRIS B. JACOBS, *Chemistry and Technology of Wines and Liquors*, 2nd ed. (1948), includes a treatment of the technology of distilled liquor production, giving extensive information on the science, art, and history of the distillation process; ALEXIS LICHINE *et al.*, *Alexis Lichine's Encyclopedia of Wines and Spirits* (1967), a comprehensive encyclopaedia of alcoholic-beverage terminology, with introductory chapters treating distilled liquor history, development, and production.

(F.M.S.)

Distribution, Theory of

The division of the social product

The theory of distribution deals with the way in which a society's product is distributed among the members of that society. It involves three distinguishable sets of questions. First, how is the national income distributed among persons? How many persons earn less than $2,000, how many between $2,000 and $4,000, how many between $4,000 and $6,000, and so on? Are there regularities in these statistics? Is it possible to generalize about them? This is the problem of personal distribution. Second, what determines the prices of the factors of production? What are the influences governing the wage rate for a specific kind of labour? Why is the general wage level of a country not lower or higher than it is? What determines the rate of interest? What determines profits and rents? These questions have to do with functional distribution. Third, how is the national income distributed proportionally among the factors of production? What determines the share of labour in the national income, the share of capital, the share of land? This is the problem of distributive shares. Although the three sets of problems are obviously interrelated, they should not be confused with one another. The theoretical approaches to each of them involve quite different considerations.

ASPECTS OF DISTRIBUTION

Personal distribution. Personal distribution is primarily a matter of statistics and the conclusions that can be drawn from them. When incomes are charted according to the number of people in each size category, the resulting frequency distribution is rather startling. Generally the top 10 percent of income receivers get between 25 and 35 percent of the national income, while the lowest 20 percent of the income receivers get about 5 percent of the national income. The inequality seems to be greatest in poor countries and diminishes somewhat in the course of economic development.

There are various explanations of the inequality. Some authorities point to the natural inequality of human beings (differences in intelligence and ability), others to the effects of social institutions (including education); some emphasize economic factors such as scarcity; others invoke political concepts such as power, exploitation, or the structure of society.

Distributive shares. For a long time economists were pessimistic as to the possibilities of any substantial improvement in the lot of those at the bottom of the income distribution. They generally held that the scarcity of productive land and the tendency of population to increase faster than the means of subsistence imposed limits on distributive justice. David Ricardo, in *On the Principles of Political Economy and Taxation* (1817), held that the landlords would receive an increasing part of the national income while capitalists would get less and less and that this shift in distribution would lead to economic stagnation. Karl Marx prophesied that the workers would be increasingly exploited and made miserable and these conditions would lead to the downfall of capitalism. Neither prediction materialized. Thus in the Western world the share of rents has dwindled to a few percent of the national income, while the share of labour has gradually increased. For some time, economists believed that the share of labour was more or less constant, but investigations show that economic development is accompanied by an increasing share of labour. Though the statistics are complicated by technical problems, it is safe to say that in the U.S., the share of wages rose from more than half the national income at the beginning of the century to more than 70 percent in the 1970s.

Contemporary approaches to this aspect of income distribution vary. Some are highly abstract and are closely related to the study of the whole, the modern macroeconomics of saving and investment. These will not be dealt with here. A simple common-sense approach employs an equation that starts by writing labour's share as the quotient of the total wage bill and the national income, labour's share $= \frac{\text{total wage bill}}{\text{national income}}$, and then writes the wage bill as the product of the wage level and the amount of labour (wage bill = wage level × amount of labour); next the national income is written as the product of the national output and the price level (national income = national output × price level); the result is that the share of labour equals the quotient of the average real wage rate and labour productivity, share of labour $= \frac{\text{average real wage rate}}{\text{labour productivity}}$, the latter being the quotient of the national output and the amount of labour: labour productivity $= \frac{\text{national output}}{\text{amount of labour}}$. If these two variables move parallel, the share of labour is constant. If the real wage rate increases faster than labour productivity, the share of labour goes up. Similar reasoning applies to the shares of capital and land. This simple arithmetic is useful for an understanding of what happens in the real world, but for a profounder analysis one must turn to the theory of functional distribution.

Functional distribution. The theory of functional distribution, which attempts to explain the prices of land, labour, and capital, is a standard subject in economics. It sees the demand for land, labour, and capital as derived demand, stemming from the demand for final goods. Behind this lies the idea that a businessman demands inputs of land, labour, and capital because he needs them in the production of goods that he sells. The theory of distribution is thus related to the theory of production (*q.v.*), one of the well-developed subjects of economics. The reasoning that synthesizes production and distribution theory is called neoclassical theory, and it will be the main subject of this article.

THE NEOCLASSICAL THEORY OF DISTRIBUTION

The theory of marginal productivity

The basic idea in neoclassical distribution theory is that incomes are earned in the production of goods and services and that the value of the productive factor reflects its contribution to the total product. Though this fundamental truth was already recognized at the beginning of the 19th century (by the French economist J.B. Say, for instance), its development was impeded by the difficulty of separating the contributions of the various inputs. To a degree they are all necessary for the final result: without labour there will be no product at all, and without capital total output will be minimal. This difficulty was solved by J.B. Clark (*c.* 1900) with his theory of marginal products. The marginal product of an input, say labour, is defined as the extra output that results from adding one unit of the input to the existing combination of productive factors. Clark pointed out that in an optimum situation the wage rate would equal the marginal product of labour, while the rate of interest would equal the marginal product of capital. The mechanism tending to produce this optimum begins with the profit-maximizing businessman, who will hire more labour when the wage rate is less than the marginal product of additional workers and who will employ more capital when the rate of interest is lower than the marginal product of capital. In this view, the value of the final output is separated (imputed) by the marginal products, which can also be interpreted as the productive contributions of the various inputs. The prices of the factors of production are determined by supply and demand, while the demand for a factor is derived from the demand of the final good it helps to produce. The word derived has a special significance since in mathematics the term refers to the curvature of a function, and indeed the marginal product is the (partial) derivative of the production function (see below).

One of the great advantages of the neoclassical, or marginalist, theory of distribution is that it treats wages, interest, and land rents in the same way, unlike the older theories that gave diverging explanations. (Profits, however, do not fit so smoothly into the neoclassical system.) A second advantage of the neoclassical theory is its integration with the theory of production. A third advantage lies in its elegance: the neoclassical theory of distributive shares lends itself to a relatively simple mathematical statement.

An illustration of the mathematics is as follows. Suppose that the production function (the relation between all hypothetical combinations of land, labour, and capital on the one hand and total output on the other) is given as $Q = f(L,K)$ in which Q stands for total output, L for the amount of labour employed, and K for the stock of capital goods. Land is subsumed under capital, to keep things as simple as possible. According to the marginal productivity theory, the wage rate is equal to the partial derivative of the production function, or $\partial Q/\partial L$. The total wage bill is $(\partial Q/\partial L) \cdot L$. The distributive share of wages equals $(L/Q) \cdot (\partial Q/\partial L)$. In the same way the share of capital equals $(K/Q) \cdot (\partial Q/\partial K)$. Thus the distribution of the national income among labour and capital is fully determined by three sets of data: the amount of capital, the amount of labour, and the production function. On closer inspection the magnitude $(L/Q) \cdot (\partial Q/\partial L)$, which can also be written $(\partial Q/Q)/(\partial L/L)$, reflects the percentage increase in production resulting from the addition of 1 percent to the amount of labour employed. This magnitude is called the elasticity of production with respect to labour. In the same way the share of capital equals the elasticity of production with respect to capital. Distributive shares are, in this view, uniquely determined by technical data. If an additional 1 percent of labour adds 0.75 percent to total output, labour's share will be 75 percent of the national income. This proposition is very challenging, if only because it looks upon income distribution as independent of trade union action, labour legislation, collective bargaining, and the social system in general. Obviously such a theory cannot explain all of the real economic world. Yet its logical structure is admirable. What remains to be seen is the degree to which it can be used as an instrument for understanding the real economic world.

SPECIAL PROBLEMS

Application of the theory to reality

Returns to scale. Neoclassical theory assumes that the total product Q is exactly exhausted when the factors of production have received their marginal products; this is written symbolically as $Q = (\partial Q/\partial L) \cdot L + (\partial Q/\partial K) \cdot K$. This relationship is only true if the production function satisfies the condition that when L and K are multiplied by a given constant then Q will increase correspondingly. In economics this is known as constant returns to scale. If an increase in the scale of production were to increase overall productivity, there would be too little product to remunerate all factors according to their marginal productivities; likewise, under diminishing returns to scale, the product would be more than enough to remunerate all factors according to their marginal productivities.

Research has indicated that for countries as a whole the assumption of constant returns to scale is not unrealistic. For particular industries, however, it does not hold; in some cases increasing returns can be expected, and in others decreasing returns. This situation means that the neoclassical theory furnishes at best only a rough explanation of reality.

One difficulty in assessing the realism of the neoclassical theory lies in the definition and measurement of labour, capital, and land, more specifically in the problem of assessing differences in quality. In macroeconomic reasoning one usually deals with the labour force as a whole, irrespective of the skills of the workers, and to do so leaves enormous statistical discrepancies. The ideal solution is to take every kind and quality of labour as a separate productive factor, and likewise with capital. When the historical development of production is analyzed it must be concluded that by far the greater part of the growth in output is attributable not to the growth of labour and capital as such but to improvements in their quality. The stock of capital goods is now often seen as consisting, like wine, of vintages, each with its own productivity. The fact that a good deal of production growth stems from improvements in the quality of the productive inputs leads to considerable flexibility in the distribution of the national income. It also helps to explain the existence of profits.

Substitution problems. Another difficulty arises from the fact that marginal productivity assumes that the factors of production can be added to each other in small quantities. If one must choose between adding one big machine or none at all to production, the concept of the marginal product becomes unworkable. This "lumpiness" creates an indeterminacy in the distribution of income. From the viewpoint of the individual firm, this objection to neoclassical theory is more serious than from the macroeconomic viewpoint since in terms of the national economy almost all additions to labour and capital are very small. A related problem is that of substitution among factors. The production function implies that land, labour, and capital can be combined in varying proportions, that every conceivable input mix is possible. But in some cases the input mix is fixed (*e.g.*, one operator at one machine), and in that situation the neoclassical theory breaks down completely because the marginal product for every factor is zero. These cases of fixed proportions are scarce, however, and from a macroeconomic viewpoint it is safe to say that a flexible input mix is the rule.

This is not to say that substitution between labour and capital is so flexible in the national economy that it can be assumed that a 1 percent increase in the wage rate will reduce employment by a corresponding 1 percent. That would follow from the neoclassical theory described above. It is not impossible, but it requires a very special form of the production function known as the Cobb–Douglas function. The pioneering research of Paul H. Douglas and Charles W. Cobb in the 1930s seemed to confirm the rough equality between production elasticities and distributive shares, but that conclusion was later questioned; in particular the assumption of easy substitution of labour and capital seems unrealistic in the light of research by Robert M. Solow and others. These investigators employ a production function in which labour and capital can replace each other but not as readily as in the Cobb–Douglas function, a change that has two very important consequences. First, the effect of a wage increase on the share of labour is not completely offset by changes in the input mix, so that an increase in wage rates does not lead to a proportionate reduction in total employment; and second, the factor of production that grows fastest will see its share in the national income diminished. The latter discovery, made by J.R. Hicks (1932), is extremely significant. It explains why the remuneration of capital (interest, not profits) has shrunk from 20 percent or more a century ago to less than 10 percent of the national income in modern times. In a society where more and more capital is employed in production, a continually smaller proportion of the income goes to the owners of capital. The share of labour has gone up; the share of land has gone down dramatically; the share of capital has gradually declined; and the share of profits has remained about the same. This picture of the historical development of income distribution fits roughly into the frame of neoclassical theory, although one must also make allowance for the short-run effects of inflation and the long-run effects of technological progress.

RETURNS TO THE FACTORS OF PRODUCTION

The analysis of earnings

The demand side of the markets for productive factors is explained in large degree by the theory of marginal productivity, but the supply side requires a separate explanation, which differs for land, labour, and capital.

Rent. The supply of land is unique in being rather inelastic; that is, an increase in rent does not necessarily increase the amount of available land. Landowners as a group receive what is left over after the other factors of production are paid. In this sense, rent is a residual, and a good deal of the history of the theory of distribution is concerned with the issue whether rent should be regarded as part of the cost of production or not (as in Ricardo's famous dictum that the price of corn is not high because of the rent of land but that land has a rent because the price of corn is high). But inelasticity of supply is not characteristic only of land; special kinds of labour and the size of the total labour force also tend to be unresponsive to

variations in wages. The Ricardian issue, moreover, was important in the context of an agrarian society; it lacks significance now, when land has so many different uses.

Wages. In analyzing the earnings of labour, it is necessary to take account of the imperfections of the labour market and the actions of trade unions. Imperfections in the market make for a certain amount of indeterminacy in which considerations of fairness, equity, and tradition play a part. These affect the structure of wages—*i.e.*, the relationships between wages for various kinds of labour and various skills. Therefore one cannot say that the income difference between a carpenter and a physician, or between a bank clerk and a truck driver, is completely determined by marginal productivity, although it is true that in the long run the wage structure is influenced by supply and demand.

The role of the trade unions has been a subject of much debate. The naive view that unions can raise wages by their efforts irrespective of market forces is, of course, incorrect. In any particular industry, exaggerated wage claims may lead to a loss of employment; this is generally recognized by union leaders. The opposite view, that trade unions cannot influence wages at all (unless they alter the basic relationship between supply and demand for labour), is held by a number of economists with respect to the real wage level of the economy as a whole. They agree that unions may push up the money wage level, especially in a tight labour market, but argue that this will lead to higher prices and so the real wage rate for the economy as a whole will not be increased accordingly. These economists also point out that high wages tend to encourage substitution of capital for labour (the cornerstone of neoclassical theory). These factors do indeed operate to check the power of trade unions, although the extreme position that the unions have no power at all against the iron laws of the market system is untenable. It is safe to say that basic economic forces do far more to determine labour's share than do the policies of the unions. The main function of the unions lies rather in modifying the wage structure; they are able to raise the bargaining power of weak groups of workers and prevent them from lagging behind the others.

Interest and profit. The earnings of capital are determined by various factors. Capital stems from two sources: from saving (by households, financial institutions, and businesses) and from the creation of money by the banks. The creation of money depresses the rate of interest below what may be called its natural rate. At this lower rate, businessmen will invest more, the capital stock will increase, and the marginal productivity of capital will decline. Although this chain of reactions has drawn the attention of monetary theorists, its impact on income distribution is probably not very important, at least not in the long run. There are also other factors, such as government borrowing, that may affect the distribution of income; it is difficult to say in what direction. The basic and predominant determinant is marginal productivity: the continuous accumulation of capital depresses the rate of interest.

One type of earning that is not explained by the neoclassical theory of distribution is profit, a circumstance that is especially awkward because profits form a substantial part of national income (20–25 percent); they are an important incentive to production and risk taking as well as being an important source of funds for investment. The reason for the failure to explain profit lies in the essentially static character of the neoclassical theory and in its preoccupation with perfect competition. Under such assumptions, profit tends to disappear. In the real world, which is not static and where competition does not conform to the theoretical assumptions, profit may be explained by five causes. One is uncertainty. An essential characteristic of business enterprise is that not all future developments can be foreseen or insured against. Frank H. Knight (1921) introduced the distinction between risk, which can be insured for and thus treated as a regular cost of production, and uncertainty, which cannot. In a free enterprise economy, the willingness to cope with the uninsurable has to be remunerated, and thus it is a factor of production. A second way of accounting for profits is to explain them as a premium for introducing new technology or for producing more efficiently than one's competitors. This dynamic element in profits was stressed by Joseph Schumpeter (1911). In this view, prices are determined by the level of costs in the least progressive firms; the firm that introduces a new product or a new method will benefit from lower costs than its competitors. A third source of profits is monopoly and related forms of market power, whether deliberate as with cartels and other restrictive practices or arising from the industrial structure itself. Some economists have developed theories in which the main influence determining distributive shares is the relative "degree of monopoly" exerted by various factors of production, but this seems a bit one-sided. A fourth source of profits is sudden shifts in demand for a given product—so-called windfall profits, which may be accompanied by losses elsewhere. Finally, there are profits arising from general increases in total demand caused by a certain kind of inflationary process when costs, especially wages, lag behind rising prices. Such is not always the case in modern inflations.

DYNAMIC INFLUENCES ON DISTRIBUTION

How incomes are affected by changing conditions

Prices. Neoclassical theory throws light upon the long-run changes in distribution of income. It fails to take account of the short-run impact of business fluctuations, of inflation and deflation, of rapidly rising prices. This failure is an omission, though it is true that distributive shares do not fluctuate as much as employment, prices, and the state of business generally. This lagging in the behaviour of shares can be understood by remembering that they are determined by the quotient of the real remuneration of the factor and its productivity; both variables move, according to marginal productivity theory, in the same direction. Yet inflation and deflation do have a certain impact upon distribution: if purchasing power shrinks, profits are the first income category to suffer; next come wages, particularly through the effects of unemployment. In a depression, the recipients of fixed-money incomes (such as interest and pensions) gain from lower prices. In an inflation the opposite happens.

The traditional inflationary sequence was that as prices rose, profits would increase, with wages lagging behind; this would tend to diminish the share of labour in the national income. Experience since World War II, however, has been different; in many countries wage levels tended to run ahead in the inflationary spiral and profits lagged behind, although most entrepreneurs eventually succeeded in shifting the burden of wage inflation onto the consumers. The result of the postwar inflation was a slight acceleration of the increase in the share of labour, while the shares of capital and land decreased faster than they would have in the absence of inflation. Profits as a whole held their own. The struggle among the various participants in the economic process no doubt added fuel to the inflationary fires.

Technology. Another dynamic influence is technological progress. The concept of the production function assumes a constant technology. But in reality the growth of production is much less the consequence of increased quantities of labour and capital than of improvements in their quality. This element in increased production is distributed in a way not fully explained by neoclassical theory. Part of the change in distribution that is caused by technological progress can be analyzed as resulting from changes in the elasticities of production. If $\frac{K}{Q} \cdot \frac{\partial Q}{\partial K}$ goes up, technological change is said to be "capital-using," and the share of capital will increase. This is what, in fact, may have happened; the change in technology has offset, though it has not neutralized, the decline in the share of capital caused by the employment of a higher amount of capital per worker. But another part of the fruits of technological progress is garnered by profit receivers, probably quite a substantial part. Businessmen who are quick innovators make high profits; in a rapidly changing society, profits tend to be high, a circumstance that is fortunate because profits are the mainspring of economic

change. The high rate of growth experienced by the post-World War I Western world stemmed from this profit-innovation-profit nexus.

PERSONAL INCOME AND NEOCLASSICAL THEORY

The neoclassical theory outlined above endeavours to explain the prices of productive factors and the distributive shares received by them. It does not come to grips with a third category of distribution, that of personal income, which is much more affected by institutional arrangements and by characteristics of the social structure. Profits in particular may be shared in various ways: they may accrue to stockholders, to workers, to management, or to the government or they may be retained in the corporation. What happens depends on dividend policy, tax policy, and the existence of profit-sharing arrangements with workers. Neoclassical theory has little to say on these matters or on the fact that in present-day capitalist society the managers of big business are virtually in a position to fix their own personal incomes. Managers have so much power vis-à-vis the stockholders and their total share of profits is so relatively little that their ability to pay themselves high salaries is limited only by the conventions of the business world. These high incomes cannot be explained by the categories of the neoclassical theory, and they do not constitute an argument against the theory. They may well argue for changes in society's institutions, but that is a matter on which the neoclassical theory of distribution does not pontificate. A great deal of change could occur in the legal and social order without any disturbance to the theory.

BIBLIOGRAPHY. JOHN BATES CLARK, *Distribution of Wealth* (1899), the classic work on marginal productivity theory whereby distribution is viewed as a harmonious process in which the factors of production receive as income what they contribute to the product; PAUL H. DOUGLAS, *The Theory of Wages* (1934), marginalist theory based on statistical research that sets forth the famous Cobb-Douglas function; JOHN R. HICKS, *The Theory of Wages*, 2nd ed. (1964), a sophisticated treatment of marginal productivity theory; NICHOLAS KALDOR, "Alternative Theories of Distribution," in *Essays on Value and Distribution* (1960), a discussion of various theories from Ricardo to Keynes; FRANK H. KNIGHT, *Risk, Uncertainty and Profit* (1921, reprinted 1971), an analysis of profits viewed as a result of imperfect foresight and as a remuneration for risk-bearing; KARL MARX, *Capital: A Critique of Political Economy*, vol. 1 (1867, many reprints), the process of distribution seen as pure conflict; DAVID RICARDO, *Principles of Political Economy and Taxation* (1817, reprinted 1933), the classical subsistence theory of wages; JOSEPH SCHUMPETER, *Theory of Economic Development* (1912), an analysis of economic development as a result of the innovations of entrepreneurs motivated by profit; K.J. ARROW *et al.*, "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics*, 43:225–250 (1961), an econometric study explaining the falling share of capital in the national income by the elasticity of substitution.

(J.P.)

Distribution of Organisms

The way in which animals and plants are distributed has long held the attention, and exercised the imagination, of naturalists. Quite apart from the intrinsic scholarly interest in the distribution of organisms, however, are aspects of direct benefit to man. The fact that certain organisms are found within definite limits of environmental conditions can be of considerable value to man. A farmer, for example, who wants to know whether an area is suitable for growing a certain crop, can be aided by knowledge of the naturally occurring plants of that area, as indicators of soil and moisture conditions. (Ironically, it is often the weedy plants, usually considered the enemy of the farmer, that give him precisely the information he wishes to know.)

Significance of distribution

The subject of distribution embraces the individual organism as a unit of study as well as the highest taxonomic unit. As far as scale is concerned, the range is equally great, from a drop of water to the largest ocean. Not surprisingly, the study of distribution encompasses several overlapping disciplines. Ecology is concerned with the numbers and distribution of organisms on a local scale, and biogeography is concerned with distribution on a regional or even world-wide scale. Other disciplines contribute much to the study of distribution and are in turn helped by it. The two most prominent of these are systematics and the study of evolution. Systematics, the study of the relationships between organisms, in the widest sense embraces classifying and naming organisms, information essential to biology in general and to the study of distribution in particular. Systematists, in turn, use distribution information in reaching decisions on the relationships between organisms. The study of evolution has benefitted considerably from distribution information, as the most well-known pioneers in the field, Charles Darwin and Alfred Russel Wallace, freely acknowledged. In return, a knowledge of the processes of natural selection has made large contributions to the understanding of the reasons for changes in distribution with time.

PATTERNS AND PROCESSES OF DISTRIBUTION

It is convenient to distinguish two closely related modes of distribution that are, in fact, inseparable in nature: (1) the local arrangement—on a small scale—of members of a population to ensure their survival by maximizing the use of the environment, often in accordance with the time of day, week, month, or year; (2) the large-scale geographical spread—on a regional, or even global scale—of a species over a long span of time. Obviously, the distribution of organisms on a local scale eventually affects the geographical arrangement of the species.

Local distribution. Just as the body of an organism has a structure, so does a population of organisms. The structure that results from the distributions of organisms is termed a pattern. Patterns may be characterized either by the forces that produced them (physical, social, or other) or, as in this article, by the ability of the organisms themselves to adjust their distribution.

Spatial and temporal arrangements. In a homogeneous environment—one that varies little from place to place—all parts of the environment may be occupied or visited by organisms. The activities of protozoans may be distributed throughout a small pool of water, and all available space on a rock in the intertidal zone of a shore may be covered by mussels or barnacles. In a heterogeneous environment, the total volume occupied or visited by a population may contain parts that are not exploited. In the pelagic, or open water, realm of an ocean, minute plantlike organisms may occur patchily, and the animals that feed upon them may be distributed in the same manner. These extremes of distribution are difficult to characterize because of the problems of measurement and representation in three dimensions. If the third dimension is not occupied, or unimportant enough to ignore, then distribution in a single plane, such as a flat piece of ground, is easy to characterize.

The three patterns of dispersion

There are three patterns of internal distribution (or dispersion) of a population in a single plane (Figure 1). Organisms are distributed at random when individuals are free to choose a place to settle. But individuals often influence each other. When they attract each other, an aggregation, or clump, tends to form; when they repel each other, an overdispersion, or uniform spacing of the individuals, results. In the soil or in the leaf litter of a forest floor, almost all arthropod species are distributed in a clumped or aggregated fashion, with only a few species randomly distributed. Territorial animals, such as birds, and some plants, such as certain cacti, repel each other at a certain distance and thereby assume a uniform pattern of distribution.

The importance of scale is illustrated in Figure 1. It shows that on a small scale, members of a population are uniformly distributed, but on a large scale they are aggregated in clumps. This happens, for example, when many flocks of birds occur in a large area but retain their discreteness. A certain minimum distance is maintained between individuals in each flock by mutual avoidance, yielding uniform spacing within the flock. But the flocks themselves may be randomly distributed or

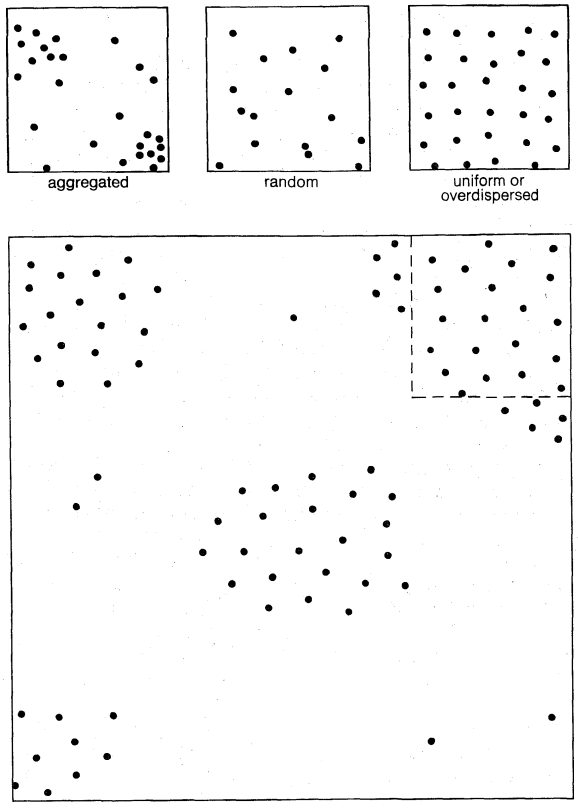


Figure 1: *Local distribution of organisms.* (Top) The three basic dispersion patterns. (Bottom) The relativity of dispersion to scale. On a local scale (inset) the individuals are uniformly distributed (overdispersed), but, on a larger scale, they are aggregated.

clumped (for further information on population dispersion see POPULATION, BIOLOGICAL).

For many species, distribution is a function of time. The members of a population make regular movements, so that their distribution is different from one hour to the next; one day, week, or month to the next. Short-term shifts in distribution are best exemplified by plankton, small, often microscopic, organisms that float in the surface waters of lakes and seas. They move vertically in response to daily changes in light intensity, so that they are at greatest depth at midday and are at the surface or close to it, at dawn and dusk.

More striking, perhaps, are the regular annual movements of some animals, which bring them to one region at the time of breeding and to another region in the nonreproductive season. Such movements are known as migrations (see MIGRATION, ANIMAL).

Environmental influences. Organisms cannot exist outside a certain range of environmental conditions. There is an upper temperature limit to the activities of protoplasm, beyond which proteins are denatured, and a lower limit, which for many organisms corresponds with the temperature at which water freezes. Many organisms, however, are capable of existing, at least for a short time, under inclement conditions. They can survive long enough to move out of the area (by migrating) or to become physiologically inactive and, thus, protected against adverse conditions (by hibernating, estivating, or encysting). Some "cold-blooded" animals have the ability to adjust or acclimate physiologically by altering their metabolic activities in such a way as to operate best under the prevailing temperatures. Physiological adjustments of this sort are possible only when the environmental conditions change relatively slowly and by a relatively small amount.

Within this framework, many processes operate in the establishment of a pattern of distribution of a species on a local scale. Perhaps the easiest to envisage is a distribution determined largely by the characteristics of reproduction. In many plants, and some lower animals, the reproductive products (seeds, spores, etc.) remain close to the parents because their dispersal capabilities are poorly developed, thus creating an aggregated distribution. In contrast are the intertidal organisms, many of which, although themselves fixed firmly to the substrate, shed their reproductive products into the water at high tide. The larvae develop in the pelagic realm of the sea and are carried to shore at the time they are metamorphosing to the adult stage.

The distribution pattern of adults that results from this combination of dispersal and settlement is determined by two kinds of processes, which are termed vectorial and stochastic. The vectorial process refers to directional dispersal caused by environmental motion, as in wind and water currents. The stochastic process refers to essentially random forces whose operation does not allow a prediction to be made as to where organisms will be carried and will settle.

Directional versus random forces

As was mentioned earlier, the pattern of distribution may be determined also by the influence of one organism upon another. The social process that gives rise to the pattern is one of signalling. The messages are either "come here," leading to aggregation, or "go away," leading to uniform spacing. Alternatively, these two patterns, as well as a random one, may be produced by each organism responding individually to environmental features such as nutrients, which are themselves distributed in random, clumped, or uniform fashion.

Interactions with other organisms. Finally, the distribution of organisms of a species may be determined by interactive processes with other species. Heavy grazing by animals in a certain area often leads to a sparseness of plant species in the grazed area. Similarly, competitor species may influence each other, the one causing the other to be either rare or absent in one part of a local area. Throughout Europe, North Africa, and some of the Canary Islands, for example, the chaffinch (*Fringilla coelebs*) occurs in both deciduous and coniferous forests, but on two of the Canary Islands, Tenerife and Gran Canaria, it occurs only in deciduous forests, a closely related species, the blue chaffinch (*Fringilla teydea*), occupying the coniferous areas. The distribution of the chaffinch on Tenerife and Gran Canaria appears to be influenced, in a restrictive way, by the blue chaffinch, with which it could compete for space and food. This type of influence, while not easy to recognize, is probably exercised commonly between species that live in the same habitat as well.

Differences in the way in which the three basic patterns of distribution—random, clumped, or uniformly spaced—are set up have been emphasized, but a single mechanism could be responsible for any one. Consider plants, for example. If the spatial pattern of germinating seeds is random, if they thrive only in space not occupied by other root systems, and if they grow to a size that depends upon the space available, then the resulting distribution will depend solely upon the density, or numbers of seeds. Going from low to medium to high density, the distribution changes from clumped to random to uniformly spaced. Thus, the factors that govern the abundance of a species within a local area indirectly determine the distribution of the species in that area. It is possible that the factors that govern the abundance of a species also limit the distribution of that species. The processes leading to the establishment of a distribution range (area or volume), as opposed to the pattern of distribution within that range, are discussed in the next section.

Large-scale distribution. More is known about distribution on large scale than on a local scale. The reasons are partly practical and partly historical; it is easier to collect a few isolated specimens of plants or animals over a wide area than to study in detail their distribution in a clearly circumscribed area such as a field or a grove. In the past taxonomists and students of evolution have spent much energy trying to determine the extent of geographical variation in morphological characters of organisms, for which purpose collected specimens are sufficient.

The shape and extent of spread. Most species occupy areas of irregular shape. The irregularities are often determined by obvious environmental features. The irregular pattern of distribution in North America of the pika (*Ochotona princeps*), a small rabbitlike animal, for example, corresponds with the irregular pattern of distribution of its habitat—rock slides and talus slopes at high altitude in the Rocky Mountains. When the distribution of certain habitat features is essentially linear, so too is the distribution of the species; the emergent vegetation at the edge of a lake and the intertidal mollusks in a narrow band along a rocky seacoast are cases in point.

Continuous versus discontinuous distribution

Viewed on a local scale, the distribution of organisms may be continuous, or without interruption, but when viewed on an intermediate regional scale, the distribution may be discontinuous, or spotty, because of unsuitable habitat intervening between areas of suitable habitat. Viewed on a large scale, a continental or worldwide scale, the distribution may appear continuous once again, as the narrow bands of unsuitable habitat that produce the discontinuities slip out of focus. But discontinuities on a worldwide scale are evident for some species. King crabs (family Limulidae) are found on the Atlantic coast of North America from the state of Maine to Mexico.

These crabs are not found on the west coast of that continent, however, but on Asian coasts, from India to Japan, thousands of miles away. Another example is provided by beetles of the genus *Catops*, which are distributed around the North Temperate Zone in four widely separated areas (Figure 2).

Adapted from Jeannel, *La Genese des faunes terrestres* (1942); Presses Universitaires de France, Paris



Figure 2: Discontinuous distribution of three beetles of the genus *Catops* in the North Temperate Zone.

Few species—of which man is one—have a worldwide distribution. Of those that do, several are associated closely with man and have achieved their present distribution as a result of this relationship. Nevertheless, there are cosmopolitan species that have achieved worldwide distribution by their own efforts. Such are the short-eared owl (*Asio flammeus*) and the osprey (*Pandion haliaetus*), two birds, both predators, conspicuous on all continents excluding Antarctica.

It is much more common to find species with restricted distributions. Aside from the numerous species confined to single, small islands, are continental species found in extremely small areas on large landmasses. The dusky seaside sparrow (*Ammospiza nigrescens*), for example, occurs only in the salt marshes around Merritt Island, Florida, and on the adjacent peninsula.

Another form of restriction is indicated by the distribution of *Catops* beetles (Figure 2): restriction to a climatic zone, in this case the North Temperate Zone. The booby (*Sula leucogaster*), a seabird, has a pantropical distribution, breeding only—but extensively—within the tropical zones.

Two closely related species of the marine wormlike *Priapulus* are restricted to the polar regions, one in the Arctic, the other in the Antarctic. In all these instances, and in variations upon these patterns, the restriction is the result of the inability of the organisms to exist beyond a range of environmental conditions. This is contrasted with restriction caused by the inability of organisms to travel over barriers, which has led to regional differentiation of faunas and floras. (This phenomenon is considered in more detail in the article BIOGEOGRAPHIC REGIONS.)

The factors limiting spread. Although more is known about the gross distribution of species than about their distribution within local areas, little has been established with certainty concerning the determination of the limits of an area occupied. In some cases the reasons for limitation of geographic range are obvious: the distribution of terrestrial organisms stops at the edge of sea, lake, or river; freshwater organisms occupy the length of a river, but extend no further than its estuary. But within the major environments the reasons for the observed distributions are not always so easy to discern.

At the outer edge of a distribution range the continued existence of organisms depends upon either reproduction or immigration from more central regions. The requirements of the individual for existence and reproduction must be met. One obvious way in which the distribution is limited is by insufficiency of the supply, or rate of supply, of these requirements. Any number of environmental factors, acting singly or in combination, can act in this limiting way.

Limiting factors

The northernmost limit of distribution of the frog *Rana sylvatica*, in North America, is almost certainly set by a combination of temperature and time. Complete development of the tadpole stage in one summer is the rule. At their northern limit, these frogs have just enough time in the short Arctic summer, at the prevailing low temperatures, to complete development. Temperature has also been implicated as a factor in limiting the southern distribution of introduced pheasants (*Phasianus* species) in North America. If the environmental temperature is high before incubation of the eggs, the eggs will not hatch. The southern limit to the distribution of the pheasant in North America corresponds approximately with those spring temperatures that are critically high.

In Scandinavia, the distribution of a carabid beetle, *Bembidion aeneum*, corresponds well with the distribution of soil that was lightly salted when the land was covered by the sea about 9,000 years ago. The beetle appears to depend upon the salt, either directly or, perhaps more likely, indirectly, through plants or other animals that require the salt and that themselves serve as food for the beetle.

In general, gross climatic and soil conditions determine the distribution of plants, singly and as assemblages; animals, being dependent upon plants for energy, follow the plant distribution. There are some highly specialized associations between plants and animals, as between some pollinators and the plants they pollinate; the edges of the distributions of the two interdependent species coincide.

Many animal species, however, are not so much limited in distribution by a single plant species as by an assemblage of plant species, such as those constituting a savannah or a tidal marsh, for example.

Factors limiting the distribution of a species are more easy to recognize when they are discrete rather than graded, as is temperature. The lizard *Uta stansburiana*, for example, which occurs in the deserts of southwestern North America, breeds only in the abandoned nests

of the pack rat (*Neotoma micropus*); the number of such nests thus sets an upper limit to the number and the distribution of breeding pairs of lizards.

The Australian ecologists H.G. Andrewartha and L.C. Birch have suggested that the factor or factors that limit the abundance of an animal species within its distribution range also limit the range itself. Exceptions to this are easy to find, particularly where the limit of distribution is set, at least in part, by a physical barrier, such as a mountain range. Nevertheless, the statement is likely to be widely true, especially for arthropods, and it is easy to appreciate when the species fluctuates in numbers markedly from one year to the next. A fine example is provided by the cutworm moth (*Porosagrotis orthogonia*), a pest species in North America, whose larvae feed on the roots of wheat. In Montana, for instance, in years of great abundance the species occurs throughout most of the state, but in years of scarcity it is either rare or absent from some parts.

Not only are distributional limits at one point set by an interaction of factors, but different boundaries are set by different factors. The saguaro cactus (*Cereus giganteus*) occurs in the Sonoran Desert of the United States and Mexico, but is not found either at the highest or lowest altitude within its area of distribution. The upper limit appears to be set by low winter temperatures, and the lower limit is set by a variety of factors, most important among which seems to be the fineness of the soil; this in turn probably affects the plants in a complex way through its properties of stability and of retention of nutrients and water.

One species can prevent the spread of another by virtue of competitive superiority in the range that it occupies, thereby setting the limit to the distribution of the other species. The distribution of two species of pocket gophers in western North America is representative. *Thomomys talpoides* occurs in northern states and *Thomomys bottae* in southern ones. An outline of their distributions shows that there is little overlap and that their boundaries are often complementary (Figure 3). No conspicuous environmental feature coincides with the highly distinctive boundaries, so it appears that the species have set them themselves by interacting.

Adapted from W.H. Burt and R.P. Grossenheider, *A Field Guide to the Mammals* (1964); Houghton Mifflin Company, Boston

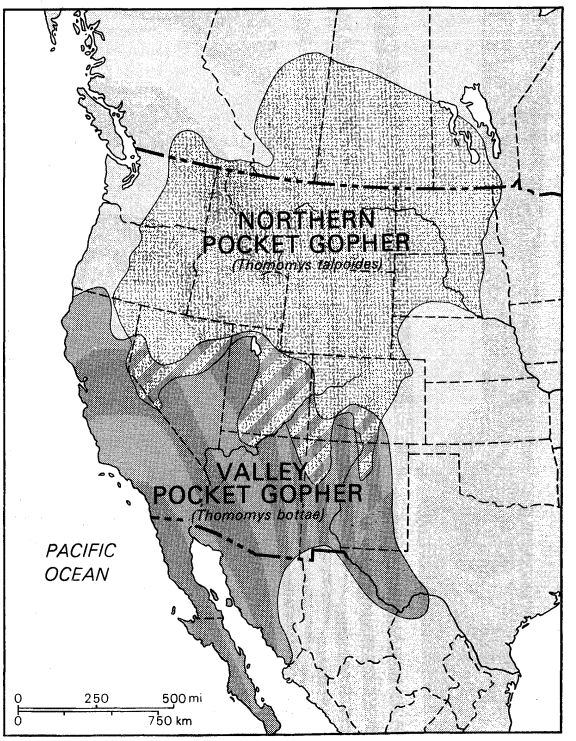


Figure 3: Complementary distribution and overlap of two species of pocket gopher in North America.

THE NATURE OF DISPERSAL

The requirements of an organism are a place to live, food, and, for sexually-reproducing organisms, a mate. The requirement of space is not easily met in a densely populated area. Animals, of course, possess the ability to seek out unoccupied space, but for plants there is a strong element of chance in the finding of suitable space. In both, however, the dispersal characteristics of the organisms are of fundamental significance and, as such, are strongly under the influence of natural selection. These characteristics partly determine whether a species is widely or narrowly distributed, and whether it can establish itself in isolated areas. Dispersal characteristics also influence the evolutionary process. Since the evolution of two populations from a single ancestral population usually requires spatial isolation, the dispersal frequency of members of the populations determines whether the degree of isolation is sufficient for evolution to take place.

The role of natural selection

Mechanisms of dispersal can be considered under two headings, active and passive. The seeds of some plants are actively dispersed by mechanical means, as when a seed pod dries out, splits open, and sends a shower of seeds to the ground in the vicinity of the parent plant. Generally, however, plants rely on passive means of dispersing, whereas both passive and active means are used widely by animals.

Active dispersal. The conditions under which active dispersal occurs are not uniform throughout the life of organisms. Dispersal usually occurs at the time of reproduction, when the individual seeks a mate or, in the case of many vertebrates, seeks a place in which it can attract a potential mate. The first patch of suitable terrain or water is chosen by the dispersing individual, and this may be close to the place where it was born. If all the space in the vicinity of home is already occupied by other individuals, however, then that individual will be forced to disperse more widely, to increase its chances of securing a mate and food. It is these far-dispersing individuals that are likely to significantly alter the distribution range of the species. The rule is almost certainly that most of these individuals perish as they encounter unsuitable conditions beyond the limit of distribution of the species. Still, some reach suitable pockets of habitat beyond and somewhat isolated from the original distribution range, and occasionally these give rise to new populations.

The most spectacular of dispersal movements, however, are the so-called irruptions of birds. These movements are not the result of a search for a place to breed, but are induced by a combination of poor food supply and possibly other adverse environmental conditions coupled with unusually high density at the end of the breeding season. An example of such an irruption is observed in Britain, which is beyond the breeding range of Pallas's sandgrouse (*Syrrhaptes paradoxus*), but which has been visited irregularly by sizable numbers of this species. As far as vertebrates are concerned, it seems that all individuals in a population have an approximately equal ability to disperse. Whether they do disperse, and how far, is dependent upon density in relation to the required resources of food and water, and the availability of mates.

Somewhat distinctive dispersal abilities are exhibited by members of many insect populations. Perhaps the best known examples are found among locusts, ants, and termites. In central and east Africa the locust *Schistocerca gregaria* occasionally swarms and consumes all green plant material in its path. The species actually occurs in two forms of quite different behaviour. The bright green form is solitary and sluggish, while the darker form, the one observed in the plagues, is highly mobile and gregarious. If the young produced by the green form are raised at high density with frequent physical contact between individuals, they develop into adults of the dark form and not the green form. This phenomenon, known as phase polymorphism, is admirably suited to the species. As the population increases and depletes its food supply, a developmental and behavioral mechanism comes into play, resulting in the production of individuals that are predisposed to dispersing.

Polymorphism: forms to fit the need for dispersal

The caste polymorphism of ants and termites serves a

similar function. In these highly social insects, in which different members of the society are structurally and behaviourally specialized to perform different tasks, dispersal to new areas and the founding of new colonies is the task of the winged forms.

A third form of polymorphism exhibited by insects is life-history polymorphism as exhibited among butterflies, dragonflies, and wasps. In all of these insects the wingless larvae are the feeding phase and disperse little, whereas the adults are the dispersal phase and feed little. In fact, the adults of some insects, such as mayflies, do not feed at all, but disperse, mate, lay eggs, and die.

There are exceptions to the statement that larval insects disperse little, which introduces a fourth kind of polymorphism, termed behavioural polymorphism. In a population of the larch budmoth (*Zeiraphera griseana*) in Switzerland, there are at least two types of larva that differ in structure and behaviour. One is intolerant to crowding; it disperses readily and is usually the first to appear on the principal food plant, larch trees. The other is tolerant to crowding and disperses little. The proportion of these two types in the population varies from time to time in accordance with the overall density. A balance of advantages and disadvantages allows the two forms to persist together without one replacing the other.

As indicated by the above example, natural selection does not always favour the individuals that disperse the most. If the chances of dispersing and finding a new, suitable, and unexploited environment are very low, selection will act against the strong dispersers. Such seems to be the case in highly restricted and well-isolated environments such as oceanic islands and mountain tops. In these environments the number of flightless birds and insects is striking, and is much greater than in continental regions. Charles Darwin pointed out that the possession of large wings and well-developed powers of flight could be a disadvantage to animals on oceanic islands, particularly small ones, because an occasional strong wind might carry an already airborne animal away from the island and out to sea. The zoogeographer P.J. Darlington has emphasized that wings are also used in flight from predators, but predators are scarce or absent on these islands, so the need for dispersal powers within the island environment is not as great as within continental environments.

Passive dispersal. Organisms are dispersed passively by three principal agents: wind, water, and other organisms.

Dispersal by wind and by water. The capability of wind to transport organisms is well known, but the extent and the scale of such transportation is only now becoming clear. Microscopic organisms, such as viruses and protozoa, are readily carried by the wind, but even some of the smaller vertebrate animals such as frogs can be carried along on strong winds. Spiders, mites, and insects have been collected on airplane flights across the Pacific Ocean at distances up to 3,000 kilometres (1,900 miles) from land. Because of the method of capture, specimens are dead, so there is no means of distinguishing whether they were or were not living when they were collected. Nevertheless, in view of these remarkable results, it is no longer surprising that even the most isolated islands have an extensive insect fauna. There are, for example, close to 4,000 insect species native to the geographically rather remote Hawaiian Islands. It has been suggested that these species evolved from 250 ancestral types that could have been windborne from continents.

In a similar way, winds carry insects from lowland regions to mountain tops, where they may be deposited in the snow. Those incapable of existing in this climatically rigorous environment (probably almost all) are eaten by the native arthropod fauna. In fact, the native arthropods may be dependent upon this wind-drifted food, just as some stream-dwelling insects are dependent for food upon the organic matter drifting down from upper levels of the stream.

A study of the spotted alfalfa aphid (*Therioaphis maculata*) in California indicates that an amazingly large number of organisms are passively transported by wind. Millions of these winged aphids alone float passively with the current in the spring months.

Adaptations for dispersal

Structural adaptations facilitating dispersal by wind are strongly developed in some groups of spiders. Gossamer, the silk secreted by the spider, is blown by the wind, and the spider, attached to it, is carried along, assisted sometimes by rising air currents.

Water currents carry organisms in the same manner as wind. In the aquatic environment, however, the terrestrial organisms must stay afloat, avoid becoming saturated with water, and maintain their internal salt balance. Some outstanding examples of long distance passive dispersal by ocean currents are known. Marine invertebrates are carried from West Africa to South and Central America on the main equatorial current of the Atlantic. Some water in this current comes from the Niger and Congo rivers, so it is possible for estuarine and freshwater animals to be transported to another continent by this means as well.

The distribution of several species of terrestrial animals is evidence that long distance dispersal has been achieved by major ocean currents. One dramatic example is the distribution of a flightless staphylinid beetle, *Micralymma marinum*, believed to have been carried passively from the Western hemisphere to the Eastern by means of the Gulf Stream.

The odds are considerable against terrestrial organisms surviving the ordeal of a prolonged journey in the surface waters of the sea. Some vertebrates may actively participate in the travel by swimming, thereby increasing their chances of survival. Others may cling to floating debris or natural rafts of land or ice. Pieces of river banks or fringing swamp vegetation and soil break off and are carried out to sea, particularly when flood conditions prevail, and with the rafts go the animals that were stranded on them. Pack ice, breaking loose, can similarly function as a raft for polar animals, such as polar bears, foxes, etc. A remarkable floating island was observed in the Atlantic off the coast of North America in 1892. It was estimated to be 100 feet square, with trees 30 feet high, and to have drifted at least 1,000 miles. Whether it reached land and its plant and animal passengers were able to colonize the new land is not known, but this example at least demonstrates the feasibility of distribution of organisms on rafts.

The distinction between airborne and waterborne means of dispersal is easy to make, but the two are not alternatives, because many organisms that start a journey on air complete it by water. Insects cross the Baltic Sea from Estonia to Finland, originally participating in mass flights, but reaching the other side in the water, many of them alive.

Dispersal by other organisms. The third means of dispersal—by other organisms—may be regular, in the case of a host transporting a parasite, or irregular, as in the chance attachment of one organism to another. For parasites, finding a host is a major problem, and an efficient means of dispersal from one host to another or to the environment in which another can be found, is essential. Consequently, parasites have developed remarkable specializations to the life history characteristics of the host that enable the parasite to disperse at the most favourable time. Dispersal is often assisted by the use of an additional species as a transporting agent, or carrier, from one host to another. A case in point is the myxoma virus, which parasitizes rabbits and uses mosquitoes as carriers. When aided by wind, an individual mosquito can carry the virus as far as 40 miles before infecting a rabbit; by this process of transporting and infecting, the mosquito population can disperse the virus rapidly over a large area.

Many organisms are distributed by attachment to mobile organisms. Small invertebrates such as snails and flatworms, for example, are accidentally transported on the legs of migrating birds. Undigested food items, such as the seeds of some plants, can be widely dispersed by their consumers in excrement deposited far from the source. There are numerous cases of mutualistic associations between animals and plants in which the plants are

dispersed. One example is provided by flightless weevils of the genus *Gymnopholus,* which carry fungi, lichens, algae, or liverworts in depressions on their back. From this association the plants gain a place to grow and a vehicle of dispersal. Plant dispersal is also carried out by mites that live for part of the time in the plant association growing in the weevil's back. The advantage gained by the weevils is that of camouflage; to a predator they appear to be part of the vegetation-covered tree bark on which they occur (see also BIOTIC INTERACTIONS).

Introductions by man

Probably the most effective agent of dispersal is man, by virtue of his numbers, widespread distribution, long-distance travel and shipping, and the objects he uses. It is not surprising to find, for example, that ports of entry are the centres of distribution of some arthropod groups newly introduced to a continent. The arthropods, other small animals, and plant seeds are transported in boats with cargo. It used to be the custom for vessels to take on soil as ballast in England and unload it in Newfoundland, which resulted in the introduction to Newfoundland of a number of soil organisms not previously found there. These examples might be termed adventitious transport through the agency of man. By contrast, there are also numerous instances of deliberate introduction of "exotic" species of plants and animals to new areas. In many cases, the newly introduced species perishes or simply does not spread from the point of introduction, as has happened with the skylark (*Alauda arvensis*), introduced from Britain to southern Vancouver Island, Canada. In other instances, the introduced species has exploited the new environment and spread rapidly. It took only 50 years for the rabbit *Oryctolagus cuniculus,* exported from Britain, to spread the length and breadth of Australia from a single farm, even in the face of efforts to prevent the spread.

The dispersal of plants depends largely on their exploitation of animals as carriers. Their structural specializations to being carried and dispersed exhibit great variety. These include a variety of hooks, barbs, bristles, and sticky secretions to aid attachment to animals, fleshy palatable fruit for consumption by animals (but enzyme-resistant seed walls), very small seeds light in weight but large in area to facilitate wind dispersal, and resistance to saltwater to facilitate dispersal by ocean currents.

The means of transport determine the type of environment to which the plant part (seed, fruit, spore, etc.) will be carried. In a study of flowering plants on various islands and island groups, it has been found that different types of island situations receive dispersing plant species by different means. Atolls, for example, receive most of their plant colonists by oceanic drift. Plants have colonized high islands most frequently by bird transport. Adherence of bristly seeds to the feathers of migratory birds is responsible for the extensive plant populations of dry volcanic islands. Air transport, whether aided by birds or not, is most effective for carrying plants to islands near the mainland.

Reduction in dispersal powers

Both plants and animals exhibit a loss or reduction of dispersal powers on islands. Examples include an increase in fruit size without concomitant increase in appendages such as hooks that serve in dissemination, a diminution or actual loss of those appendages, and an alteration of the mechanism of the release of fruits. These evolutionary shifts prevent the loss of reproductive products to unsuitable neighbouring environments in which they may not survive. There also tends to be a corresponding reduction in the number of seeds or fruits produced. And, in the case of increases in fruit size, there is the additional advantage of a greater volume of stored food, of particular help to seedlings growing under shady forest conditions. There are probably other less conspicuous changes in the dispersal abilities of plants, affecting such attributes as resistance to seawater, resistance to the digestive enzymes of animals, and the length of viability of seeds.

COLONIZATION OF NEW AREAS

From what has been said about the dispersal powers of organisms, it is a wonder that all taxonomic groups are not found everywhere. They are not, for two principal reasons. Different groups have different dispersal abilities, and environments have a limited capacity for supporting organisms, in terms of both number of individuals and number of species.

Effects of dispersal abilities. The probability that an organism will colonize successfully a new area is lessened by the distance it must travel and increased by the suitability of the new area. The most rapidly successful colonizing species are those with well-developed dispersal powers and a wide tolerance for environmental conditions. Plants referred to as weeds have these properties, and in addition they have the ability to exploit recently disturbed terrain. As a community of plants and animals undergoes a natural successional change toward greater complexity and stability, these pioneer organisms are replaced by organisms inferior in dispersal abilities but superior in competitive abilities. Such pioneer species, both plants and animals, tend to be found in ecologically marginal and geographically peripheral regions, from which position they are well suited to disperse to new areas. Many of them have other advanced biological properties that relate to their role of ecological opportunism, including the ability to reproduce without fertilization.

The future of initial colonists is partly determined by the ability of other competing species to reach the new area. The barrier to be crossed in reaching the new area —be it mountain, sea, river, or climatic zone—is not to be considered an insuperable one; it simply presents an obstacle that reduces the probability of transgression by some organism. There may be what has been termed a filter route across the barrier; that is, a route suitable for only some members of a fauna or flora, those with the most well-developed dispersal powers (Figure 4). An island twenty miles from the mainland can be colonized by birds flying across the water barrier, a route suitable only for the more dispersive members of the mainland bird life. Alternatively, the barrier may be almost impenetrable. For some organisms, the Sahara must be such a formidable barrier. The route across it is a "sweepstakes route"—to use the term coined by the American paleontologist George Gaylord Simpson—for the probability of any organism crossing the barrier is very low indeed, and yet some do cross it. Chance plays the major role in determining which ones get across, and when.

From *Life: An Introduction to Biology* by George Gaylord Simpson and William S. Beck, © 1957, 1965 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

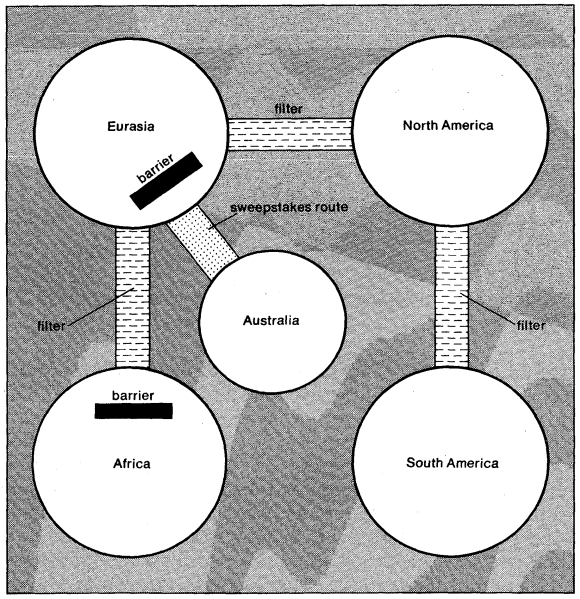


Figure 4: Barriers and possible filters for animal dispersal.

Chance in dispersal

The most strongly isolated environments may be so difficult to reach that the initial colonists have the opportunity to undergo evolutionary changes, unaffected by immigration of members of the same species or members

of different species. In both plants and animals, these changes involve the relinquishing of their opportunistic characteristics.

Effects of the new environment. On the other hand, an organism arriving in a new area may have little chance of establishing itself, either because the habitat is unsuitable or because numerous other species already occupy the area. There appears to be a maximum number of species that can be supported in a given area (or volume) of environment. Size of the environment itself is one determinant of this maximum; the larger the area, the larger the number of species (and individuals) present. Climatic and related physico-chemical factors are other determinants; these are responsible for the gradual change in species from the equator to the poles. This variation in species number with latitude is well illustrated for the mammals. It is equally well illustrated by mollusks, snakes, birds, ants, or corals. In contrast, variation in species number with longitude is less pronounced and less predictable. The number of species at any point along the latitudinal gradient is more or less the maximum possible, given present-day physico-chemical conditions and the total number of species in the world today. In other words, most areas are saturated with species; if a new species arrives in an area, it can establish itself usually only by displacing a resident species.

The process by which an area becomes saturated with species can be expressed in a simple model (Figure 5).

Adapted from R.H. MacArthur and E.O. Wilson, *The Theory of Island Biogeography* "Monographs in Population Biology," vol. 1 (1967); Princeton University Press

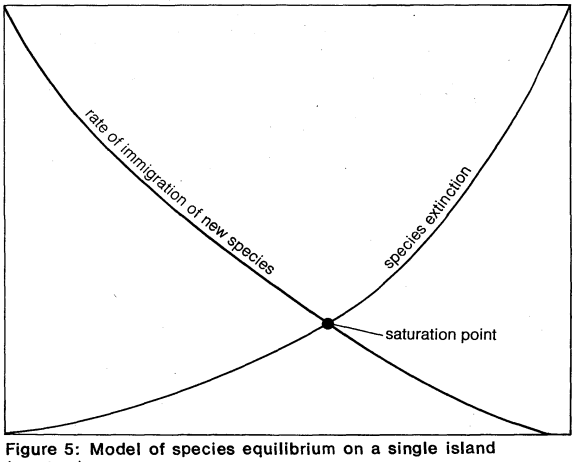


Figure 5: Model of species equilibrium on a single island (see text).

Immigration and successful establishment of new species depends on the number of species already present. The rate of establishment of new species declines and the rate of extinction of older species rises until a point is reached when these two processes are equal, and no further change in number of species takes place; the area is then saturated. There are interesting variations of this model that allow assessment of the importance of size of area and degree of isolation upon the two processes and upon the equilibrium number of species. Large size and weak isolation, for instance, are conditions conducive to a large equilibrium number (many species), and small size and large isolation yield a small equilibrium number (fewer species).

If the environment changes, however, the equilibrium number is subject to change. The shifting of average temperature over a long period of time in water, for example, may enable more species to exist at a given position in that body of water. In the terrestrial environment, the geographical distribution of climate has changed frequently in the past, which has caused a change in the distributions of organisms.

CHANGES IN DISTRIBUTION WITH TIME

Interpretation of past biological events will always remain a challenge because the necessary information will never be complete. The determination of past distributions of organisms, and changes in them, is biological detective work with generally few clues at hand. It is first of all dependent upon fossils that can be identified and located as to place and time of occurrence. Attention is focussed less on individual species than on large groups (such as a family or order) or an ecological assemblage (such as a community or population). It is, secondly, based upon the principle that biological and ecological processes operating today are the same as those that operated in the past.

Centres of evolutionary origin. A survey of many groups of organisms shows a common pattern. It is characterized by evolution and proliferation (adaptive radiation) into a diversity of types adapted to performing a variety of ecological roles. The adaptive radiation of marsupial mammals in Australia from a primitive insectivore is a good example. The evolutionary processes are accompanied by expansion of distribution ranges and a geographical radiation. There follows recession or contraction of distribution ranges, of the group as a whole or of individual species, as new groups arise and partly supplant their predecessors. Superimposed upon this pattern are large-scale shifts in distribution, without necessarily any expansion or contraction, as environmental conditions change; for instance, there was a steady southward shift of northern vegetation zones during much of the Tertiary Period (7,000,000 to 65,000,000 years ago), evidently caused by a cooling of northern climates. The distribution of organisms in the past can be viewed as an ever-changing kaleidoscope, the result of movements, countermovements, spreadings, competitions, extinctions, and replacements of a diversity of plants and animals over the whole of the diverse surface of the world.

An examination of the distribution of fossil material and the distribution of present-day organisms suggests that there were regular aspects to the kaleidoscope. It has been recognized, for example, that there were major centres of origin and evolution of terrestrial vertebrates, although the identification of these centres has not met with unanimous approval because the fossil evidence by itself is not convincing. Some biologists have suggested that the North Temperate Zone was the centre of origin because its variable climate continually presented challenges to vertebrates, to which the vertebrates responded evolutionarily by diversifying. Others have argued instead for the tropics as a centre of origin and evolution, in view of the greater number and variety of species in the tropics than in either temperate zone today. Furthermore, North Temperate forms are not usually peculiar to the north, but are representatives of groups also present, often more abundantly, in the tropics; among birds, the hawks, owls, and finches may be cited as among these forms. Finally, many groups present in the tropics are absent in the north.

The tropics as the evolutionary centre

There are other reasons for considering the tropics to be an evolutionary centre, and here historical clues play a large part. It seems reasonable to expect that (1) the region in which the largest number of species and genera existed in the past is where the majority of them originated, (2) the region of greatest degree of differentiation might be the place of origin of a group, and, finally, (3) the region of largest area would have been occupied longest. The fossil evidence interpreted in the light of these three statements points strongly to the Old World tropics as the main origin of terrestrial vertebrates.

The subsequent spread of organisms. *Environmental influences.* Major shifts in the distribution of whole faunal and floral assemblages can usually be correlated with inferred environmental change. At different times in the history of the earth there have been different patterns of climate and, more particularly, the average temperature has been higher or lower than at present. Climatic changes have had profound effects upon animal and plant distribution, both directly and indirectly. In colder periods, for example, the earth was covered with more ice and less water, and the sea level was sufficiently depressed that landmasses, now separated by a water barrier, were connected.

The best known of these climatic changes is the most recent Ice Age, comprising several glacial (cold) and interglacial (warm) periods. The sea level was as much as 120 metres (390 feet) lower than at present during glacial periods, and as much as 50 metres (165 feet) higher than at present during interglacial periods. These fluctuations affected organisms most profoundly in polar and temperate regions, least in the tropics.

At the start of the Tertiary (Paleocene Epoch), there were great differences between the faunas and floras of North America and Eurasia. There then occurred an interchange of faunas, and the geological and fossil evidence suggests that the interchange occurred across a land bridge in what is now the Bering Sea. Initially, the interchange involved large parts of the continental faunas and many major groups, most noticeably among the mammals. Subsequently, exchanges were increasingly selective and tended to involve smaller fractions of faunas, progressively less distinctive types of animals, lower taxonomic categories only, and ecological types less novel in the invaded regions. This has been interpreted to mean that the land bridge was increasingly difficult to cross, perhaps the result of an increasingly effective climatic barrier; additionally, each continent may have become increasingly saturated with the major types of animals, making it progressively more difficult for a potential colonist to establish itself. This whole process, extending down through the Tertiary Period and up to the recent Ice Age (Pleistocene Epoch), was evidently interrupted three or more times by submergence and subsequent elevation of the land bridge.

The dispersal of man

One of the species of mammals taking advantage of the land bridge across the Bering Sea, it is thought, was man himself. By late Pleistocene times man had become a dominant species throughout the Old World through the evolution and development of ecological versatility. In several waves of dispersal, man entered North America, probably at the last continental glaciation. Dispersal and changing (mainly expanding) distribution have continued to be characteristic of the species.

In contrast to the events in the far north, North and South America were separated during most of the Tertiary Period. They were connected by land possibly at the beginning, and then again late in the Pliocene Epoch. The effect of this Pliocene connection upon the distribution of organisms can be illustrated with mammals, of which many fossils are available. Before the connection was made, North America had about 27 families of terrestrial mammals, and South America had 29 families. Only one or two families were common to both continents. After the connection, the number of families in common rose to a maximum of 22. Incursions, withdrawals, and extinctions led to North America possessing 23 families and South America, as before, 29. North America benefitted by the arrival of opossums, three families of ground sloth, two stocks of armadillos, and many others. South America benefitted by the arrival of several species or genera of cats (including saber-toothed tigers), horses, camels, elephants, tapirs, and rabbits. Most have survived to the present day, although tapirs, camels, elephants, horses, and sabre-toothed tigers have since disappeared from North America, and South America has lost a portion of its rich fauna also.

Contemporary patterns of distribution

The recurring expansion and contraction of distribution ranges have produced some patterns exhibited by contemporary organisms that at first sight seem inexplicable. Continuous distributions are not difficult to understand; but what past events led to the highly discontinuous or disrupted distributions evident today, such as the distribution of king crabs in North America and southeast Asia? Migration from one area to another is possible but unlikely; it is more likely that the distribution was once more widespread and either continuous or moderately discontinuous, and that a shrinking of distribution range has occurred. The distribution of fossils indicates that king crabs were widespread and abundant in the Paleozoic Era, and that a restriction of range occurred in the Mesozoic, possibly as a result of competition with newly evolved marine invertebrates.

Other species were profoundly affected by climatic disturbances, such as the recent Ice Age, surviving partly in relatively small areas of suitable habitat, or refuges, and partly in larger areas farther away, to which they dispersed when the climatic change rendered the previously occupied area unsuitable. The evidence for this consists of numerous instances of much differentiated forms isolated from the main range of the group. These are referred to as relicts. Examples are Arctic plants found isolated on mountain tops in north and middle Europe.

Even before Darwin's time, their occurrence was explained as due to survival in refugia, areas untouched by glaciers but once surrounded by them.

Certain groups of coastal plants still may be confined to their glacial refugia because they are not evolutionarily versatile enough to break out from them; their variability became restricted during the past glaciation when new variants would have been at a disadvantage during the harsh conditions. In contrast, there are some species, termed disharmonious relicts, which occur in unusual and isolated habitats where they are thought to have survived because of broad ecological tolerance in the past.

Relict patterns of distribution are found on islands as well. These are in irregular and unpredictable patterns interpreted as having been formed by partial extinction of an old fauna or flora formerly common to all the islands. They are most pronounced on oceanic islands. An example is the distribution of frogs on islands in the Indian Ocean. The pattern is replete with discontinuities and mixed relationships, which indicates that the fauna is partly a relict one, and that existing geographical relationships are partly the result of extinction and survival rather than of direct dispersal. The relict pattern of distribution contrasts with an immigrant pattern, characteristic of continental islands, in which the number of species in a taxonomic group decreases in more or less orderly fashion with distance from the continent, although modified by island area.

Geological influences. Much has been said about the distribution of terrestrial organisms, as if the organisms have moved and the land has stood still. But, in fact, there is a growing body of evidence to indicate that the land also has moved, and that the continents have not always been in their present positions. If this is true, the distributions of organisms, in relation to the Earth's axis, have undergone even more change than has been referred to so far.

The theory of continental drift (*q.v.*) proposed by the geologist A.L. Wegener holds that the continents were once connected, but that they split apart before or during Early Mesozoic times (150,000,000 to 225,000,000 years ago) and gradually drifted apart. Often maligned, the theory is currently in favour as a result of new evidence, and mechanisms of continental drift have been proposed in detail. Some of the geological evidence is based on the shapes of continents. The outline of West Africa, for example, is complementary to that of eastern South America. If the two continents are fitted together like a jigsaw puzzle, the Paleozoic and Early Mesozoic rocks of the two continents in the vicinity of their (present) coasts show remarkable correspondence. The paleontological evidence is that the flora and fauna of the two continents showed a great degree of resemblance in those early times, and this resemblance may have diminished progressively thereafter. The fossil evidence is equivocal, however, and the similarity of geological strata on the separate continents has been attributed alternatively to similarity of climate (on separate continents) in times past, so the theory is far from being universally accepted. If continental drift was a reality, however, it must be viewed as a major contributor to changes in the distribution of organisms in time.

Man's impact on distribution

Human influence. It is appropriate to conclude with a few remarks about man. The agricultural, industrial, and commercial activities of man have had an enormous impact upon the distribution of organisms. In the last 10,000 years man has replaced the climate as the most

important agent of change. Even in the Pleistocene Epoch man may have been an important agent of change, for it has been suggested that large-scale extinctions of mammals at that time were not the result of climatic change but of "overkill" at the hands of man.

Some organisms are more widely distributed now as a result of man's activities, many more are more restricted, and a sizable number have disappeared altogether. Since 1600, almost 100 species of birds and 40 species of mammals are known to have become extinct, and it is probable that man exercised an important, if not decisive, influence in these extinctions. Moreover, many species are now on the verge of extinction, again largely because of man's activities.

Man's influence upon the distribution of organisms has been, and continues to be, needlessly destructive. Admittedly, some destruction is probably inevitable, since species like man, with novel ecological features, spread and cause a reduction in numbers of other species. But this process leads to a state of dynamic equilibrium, to something approaching stability. Unfortunately, modern man shows no sign of approaching an equilibrial relationship with his environment in the immediate future.

BIBLIOGRAPHY. Local distribution is considered in detail in H.G. ANDREWARTHA and L.C. BIRCH, *The Distribution and Abundance of Animals* (1954); and V.C. WYNNE-EDWARDS, *Animal Dispersion in Relation to Social Behaviour* (1962). Large-scale distribution is dealt with in S.P. EKMAN, *Tiergeographie des meeres* (1953; Eng. trans., *Zoogeography of the Sea*, 2nd ed., 1967); P.J. DARLINGTON, JR., *Zoogeography: The Geographical Distribution of Animals* (1957); P. DANSEREAU, *Biogeography: An Ecological Perspective* (1957); R.H. MACARTHUR, *Geographical Ecology: Patterns in the Distribution of Species* (1972); L.F. DE BEAUFORT, *Zoogeography of the Land and Inland Waters* (1951); L.C.M. CROIZAT, *Panbiogeography*, 3 vol. (1958); W.B. GEORGE, *Animal Geography* (1962); R. GOOD, *The Geography of Flowering Plants*, 3rd ed. (1964); M.J. DUNBAR (ed.), *Marine Distributions* (1963); N.V. POLUNIN, *Introduction to Plant Geography* (1960); D.J. DE LAUBENFELS, *Geography of Plants and Animals* (1970). The factors behind dispersal are examined in M.D.F. UDVARDY, *Dynamic Zoogeography* (1969); S.J. CARLQUIST, *Island Life* (1965); C.S. ELTON, *The Ecology of Invasions by Animals and Plants* (1958); C.H. LINDROTH, *The Faunal Connections Between Europe and America* (1957); L. VANDERPIJL, *Principles of Dispersal in Higher Plants* (1970); R.H. MACARTHUR and E.O. WILSON, *The Theory of Island Biogeography* (1967).

(P.R.G.)

Divination

Divination, the alleged art or science of foretelling the future by various natural, psychological, and other techniques, is a phenomenon found in all civilizations in all times and areas. It is known especially in the modern Western world in the form of horoscopic astrology.

NATURE AND SIGNIFICANCE

In the context of ancient Latin language and belief, divination was concerned with discovering the will of the gods. Today, however, scholars no longer restrict the word to that earlier root meaning. Divinatory practices and the beliefs undergirding them are greater in scope than the ancient methods of discerning the will of the gods and the fatalistic view of the human condition that inspired so much of early Mediterranean religious thought. In some societies, in fact, divination is a practice to which many persons frequently resort, but never in terms of discovering the will of the gods. The idea of a godly providence controlling human affairs, in such societies, is unusual, although humbler spirits are often thought to intervene in troublesome ways.

Concerns of divination

Divination is universally concerned with practical problems, private or public, and seeks information upon which practical decisions can be made; but the source of such information is not conceived as mundane, and the technique of getting it is necessarily fanciful. The mantic (divinatory) arts are many, and a broad understanding can only emerge from a survey of actual practices in various cultural settings. A short definition, however, may be offered as a preliminary guide: divination is the effort to gain information of a mundane sort by means conceived to transcend the mundane.

Though the act of divination is attended by respect and the attitude of the participants in the divinatory act may be religious, the subject matter of divination (like that of magic) is ephemeral—*e.g.*, a trouble case, an illness, a worrisome portent, an object lost. Divination is a consultative institution, and the matter posed to a diviner may range from a query about a few lost coins to high questions of state. The casual or solemn nature of the matter is normally matched by that of the diviner in terms of attitude, technique, and style. Where the diviner is a private practitioner, the quality of the procedure may be linked to the fee. This can become a matter for bargaining or implicit contention, so that worldly motives are patent. In contrast to the pecuniary motive of some diviners, the ancient Etruscans in Italy and the Maya in Mexico viewed the calling of diviner-priest as sacred, and his concern was for the very destiny of his people. Divination has many rationales, and it is difficult to describe the diviner as a distinctive social type. He may be a shaman (private curer employing psychic techniques), a priest, a peddler of sorcery medicines, or a holy man who speaks almost with the voice of prophecy. To appreciate the significance of the diviner's art in any culture or era, one must know the culture or era's fundamental beliefs about man and the world. In Christian times Europe has moved from a horror of necromancy (conceived not as consultation with a ghost but as a literal "raising of the dead") to an amused tolerance (among the educated) of spiritualism as a sort of parlour game. To assert that European beliefs about God and man have remained the same throughout the Christian Era would be to ignore the broad impact of natural science and secularization in the modern era. On the other hand, to suppose that divination has been doomed by science and secularism would be to ignore the abiding popularity of horoscopic astrology (divination by means of relating the movement of stars, planets, the Sun, and the Moon to a person's birth date) and recurrent fashions for other mantic disciplines —and perhaps to misjudge the security of "modern" beliefs.

THE STRUCTURE OF DIVINATION

Importance of world views and philosophical undergirdings for divination

The extent to which a practice such as divination should be called a corollary of the beliefs entailed and the extent to which the opposite might be true (*i.e.*, the beliefs deriving from the practice as an after-the-fact explanation) is difficult to ascertain. Among the great cultures, the Chinese tradition has given the broadest scope to divination; yet there is no single Sinic religious cosmology, or theory on the ordering of the world, comparable to those of the Maya, Sanskritic (Hindu), or Judeo-Christian traditions, from which the variety of popular practice can be seen to derive. Sometimes, as with the flourishing business of astrology in Christian countries since the Renaissance, the metaphysical (transcendent) presuppositions of mantic practice may have been de-emphasized in order to minimize conflict with official religious and scientific doctrines. Generally, however, it may be said that the philosophical underpinnings of divination need not be deep or well worked out, but, where they are, they will afford clues to fundamental beliefs about man and about visible or invisible nature. Some traditions of divination—such as astrology, geomancy (divination by means of figures or lines), or the Chinese divinatory disciplines—are so old and firmly established that it is virtually impossible to discover their original contexts. Over the centuries such practices have survived many changes and they have become—as they have passed over numerous local and cultural boundaries—timeless answers to the constant questions of the human condition.

Established long ago in the hieratic (priestly) discipline of primitive theocracies, such a tradition still bears the marks of the specialists who worked out its systematic techniques and its cogent symbolic structures. Since the practice is now observed only as a folk or plebeian tradition, however, it would be rash to suppose that the true philosophical tradition undergirding divination survives.

Only in the case of the *I Ching*, the Chinese *Book of Changes*, have scholarly commentaries accumulated over the millennia in a context of considerable intellectual continuity. Systematic studies of geomancy are recent, and the literature of astrology is as perishable as it is massive. Babylonian astrology, from which later forms are derived, arose in an agrarian Mesopotamian civilization concerned with the vicissitudes of nature and the affairs of state. The mercantile, seafaring, and individualistic Greeks absorbed the mantic system of the collectivistic, floodplain civilization of Mesopotamia, elaborated on it by adding the horoscopic discipline, and transmitted it, through Hellenistic (3rd century BC–3rd century AD), Egyptian, and Islāmic science, to Europe. In the course of this transformation, the dual influence of world view upon the structure of divination may be seen: a most varied band of priests, scholars, and innovators have made their contributions to the system; yet there also is a clear correspondence between the general character of a culture and the uses it finds for divination. That is, the world view or views implicit in the divination system itself may reflect the historical rather than the current context of use. It requires only practical understanding to consult a ouija board or use a forked stick to decide where to drill for water. Hence, men of very different beliefs may adopt the same practices, and a full correspondence between practice and belief can only be expected where both have developed in the same cultural context. Where much of the popularity of the mantic art derives from its "exotic" flavour, its symbolism may be little understood. By its very nature, however, divination tends to develop as a discipline, becoming the tradition of an organized body of specialists. This is because the means to which the diviner must resort are conceived as setting him apart. That is the case even among such peoples as the Azande of the Nile-Congo Divide in Africa, where the resort to divination is frequent, and the most common techniques utilized are recognized to be within the competence of ordinary men. There, on a sensitive or contentious issue, an extraordinary credibility is desired, and the ultimate reliability of an oracle reflects the political standing of its owner—the king's oracle, for example, is viewed as the final authority, and the royal court is scrupulously organized to guard this vessel of power (divinatory and other) from contamination. Few societies are as enthusiastically given to divination as the Azande, who routinely employ it to explore their thoughts and who will not consider any important undertakings without oracular confirmation in advance. Among the Azande, the ordinary man could be considered a divinatory specialist. Elsewhere, men are content to reserve divination for special crises, and consultation must be with a recognized expert in order to distinguish an authentic answer from a spurious one. The diviner builds up his credibility through a dramatization of his role, the use of impressive paraphernalia, striking behaviour, or a combination of special accoutrements used for effect and an unusual rhetoric or rapport.

Types of divination. As schools of dramatic art range from those relying on explicit technique to those teaching intuitive identification with a role, mantic skills range from the mechanical to the inspirational but most often combine both skills in a unique, dramatically coherent format. The comparative study of divinatory practices is at least as old as the 1st-century-BC Roman orator and politician Cicero's treatise *De divinatione* (*Concerning Divination*), and the convenient distinction there drawn between inductive and intuitive forms designates the range. An intermediate class, interpretive divination, allows a less rigid classification, since many divinatory disciplines do not rely strongly either upon inductive rigour or upon trance and possession.

Inductive, intuitive, and interpretive forms of divination

Inductive divination presupposes a determinative procedure, apparently free from mundane control, yielding unambiguous decisions or predictions. The reading of the "eight characters" of a Chinese boy and girl before proceeding to arrange a marriage—the year, month, day, and hour of birth of the two persons to be betrothed—illustrates this class of procedures. The "characters" are all predetermined by the accidents of birth date and hour; and it is supposed that all proper diviners would come to the same conclusions about them.

Interpretive divination requires the combination of correct procedure with the special gift of insight that sets a diviner apart from his fellows. The contemporary Maya diviner of Guatemala, seeking to diagnose an illness, will carefully pass a number of eggs over the patient's body in order to draw into them an essence of the affliction. The intact contents are then collected in water, and the diviner withdraws into a darkened corner to bend over the receptacle and read the signs of the eggs. His recitation then explains the origin and nature of the disease, according to his reading.

Intuitive divination presupposes extraordinary gifts of insight or ability to communicate with beings in an extramundane sphere. The "Shaking Tent" rite of the Algonkian Indians of Canada illustrates the use of uncanny phenomena to lend credence to a mediumistic performance. The diviner, bound and cloaked, is no sooner placed in his barrel-like tent than the tent begins to shake with astonishing vigour and to fill the air with monstrous noises; and this continues with great effect until, all of a sudden, the communicating spirit makes his presence known from within the tent and undertakes to answer questions. The virtuoso performance illustrates the difficulty of explaining the phenomena of spirit possession as products of deliberate instruction. Evidently, such skills must be actively sought by the neophyte, and they answer to deep tendencies in his personality.

Cosmological and psychological conditioning

The cosmological and psychological conditioning that affects divinatory practices within a cultural tradition will influence in a similar fashion all of its religious practices. The Greeks tended to the intuitive or "oracular" style, and the Etruscans, in contrast, elaborated upon the more systematic but less versatile inductive practice of Mesopotamia—developing an authoritative state religion in which the positions were monopolized by the ruling class. Greek divination was eccentric in that sanctuaries were located apart from the centres of political power; the Etruscan system, on the other hand, was concentric, focussed at the summit itself. Rome eclectically incorporated both Greek and Etruscan elements, the ecstatic cult and the expert "reading" of livers—*i.e.*, haruspimancy. Rome, however, never allowed divination to become the central preoccupation of society as it had been for Etruria, nor did it become an autonomous force in society as it had been for the Greeks. In this, Rome represented a balance that is more congenial to modern Western thought. Throughout the ancient Mediterranean world, the notable exception being Egypt, divination was tied to expiation and sacrifice: fate was perceived as dire but not quite implacable, and the function of divination was to foresee calamity in order to forestall it. In trans-Saharan Africa, religion centres about expiation and sacrifice, and divination is a pivotal institution, but the Mediterranean notion of fate is not developed. Instead, the trouble of a person is attributed to witchcraft, sorcery, or ancestral vexation—all of which are believed to be arbitrary and morally undeserved. Divination is employed to discover the source of trouble in order to remove it, whether by sacrifice, countersorcery, or accusation and ordeal. The mind is turned to past events or hidden motives of the present time, however, and not to the future—that would be to borrow trouble.

The function of divination. The function of divination needs to be understood in its motivational context. It is not enough to say that information won from the diviner serves to allay uncertainty, locate blame, or overcome man's impotence before misfortune. Divination is motivated to the extent that information, whether spurious or true, will please a client. Unless one assumes that the information is usually not misleading or palpably false, one should expect that clients would be displeased and skeptical. A careful assessment of the kinds of information that divinatory systems are required to yield is thus in order. The two main kinds are general information about the future and specific information about the past as it bears upon the future.

The first kind of information is yielded by horoscopic divination. It is usually so general that it cannot be properly tested, for it bears the stamp of destiny. If such information were specific enough, the prediction could interfere with its own fulfillment, acting as a warning or breeding overconfidence. The other kind of information demanded from diviners is specific enough to be tested and often is; but testing a particular diviner's competence is seldom conceived to put the institution to test. It is common in trans-Saharan societies for a troubled client to consult a series of diviners until he finds one who is convincing. Again, many systems of divination have a double check built into them: the question is posed first in the positive and then in the negative, and the oracle must (obviously without manipulation) answer consistently. The chances are actually even that any oracle will fail to do so, yet the credibility of such oracles seems not to be lost. Technically, this means that untenable or invalid information can be obtained by a client without disconfirming his belief in the source. Early students of divinatory practice concluded that clients must be gullible, superstitious, illogical, or even "prelogical"; *i.e.*, culturally immature. Ethnographic studies do not confirm this, suggesting rather that what a client seeks from the diviner is information upon which he can confidently act. He is seeking, in so doing, public credibility for his own course of action. Consistent with this motive, he should set aside any finding that he thinks would lead him into doubtful action and continue his consultations until they suggest a course that he can take with confidence. The diviner's findings are judged pragmatically.

Occasions of divinatory consultation

Men seek out a diviner when they are unsure how to behave—when there is illness, drought, or death and the fear of death; when there is suspicion of malevolence, secret thieving, or breach of faith; when dreams or other symptoms are disturbing or the signs of the time seem bad. Divination universally serves the purpose of circumscription, of marking out and delimiting the area of concern: the nature of the crisis is defined, the source of anxiety is named. Concern becomes allegation, bafflement decision. The diviner may function as a stage manager, speeding up the action, rejecting false moves in advance, or indicating the secret fear or the hidden motive. Finally, when the client is content to act upon the diviner's information—such as accusing a thief or sanctioning his wife for adultery—his course of action is self-explanatory. Where divinatory practice is a recognized resource, a man who ignores it is considered arbitrary, and one who heeds it needs no further justification. In this sense, the ultimate function of divination is the legitimation of problematic decisions.

VARIETIES OF DIVINATION

Because dramatic effect is important, divination takes many forms and employs a wealth of devices. In a general way, it may be said that inductive divination employs nonhuman phenomena, either artificial or natural, as signs that can be unambiguously read. The prime condition is that the signs appear to be genuine, not manipulated. Interpretive divination commonly combines the use of nonhuman phenomena with human action, employing devices so complex, subtle, or fluid that the special gifts of the diviner seem required if the meaning is to be known. It is here that divination takes its most characteristically dramatic forms. Intuitive divination usually places little reliance upon artificial trappings, except for dramatic effect. The performance exhibits and may crucially test deep human qualities: the impressive performer may want gifts like those that in a different context would have made him an effective actor, writer, or political leader. Where the diviner can produce voices other than his own, the impression is that the gods or spirits are speaking.

Inductive divination. To speculate that inductive divination from natural phenomena must be very old—*i.e.*, that it arose from early man's intimate acquaintance with nature—is tempting. In fact, however, evidence of an awareness of nature as a system among preliterate peoples is spotty, and this is particularly true in respect to astral observation. Divination from the skies—involving such phenomena as configurations of the stars, planets, the Sun, and the Moon—is concerned pre-eminently with the future but presupposes a concern with cycles of time and history. Quite distinctive attitudes were taken toward the celestial clock by the ancient Maya astronomers and those of Mesopotamia; and distinct but related forms of astrology were developed in the Western, Indian, and Chinese civilizations. But the foundation of astrology in scientific astronomy is quite apparent, and the two "sciences" were inseparable in the West until early modern times.

Forms of inductive divination

Associated with the observation of the heavens is the reading of signs in the weather and the movement of birds. The interpretation of lightning as a decipherable message from the gods—not simply as an outburst of divine anger—was brought to the level of a pseudoscience by the Etruscans. Winds and clouds, being suited to less exact observation, invited interpretive rather than inductive divination. Weather phenomena were also conceived to be in a special status relative to man, in that rain, drought, and natural disasters are forces that man seeks not simply to read but to control. Nonetheless, Hindu Scripture discusses the art of interpreting "castles in the air"—celestial cities seen in towering clouds.

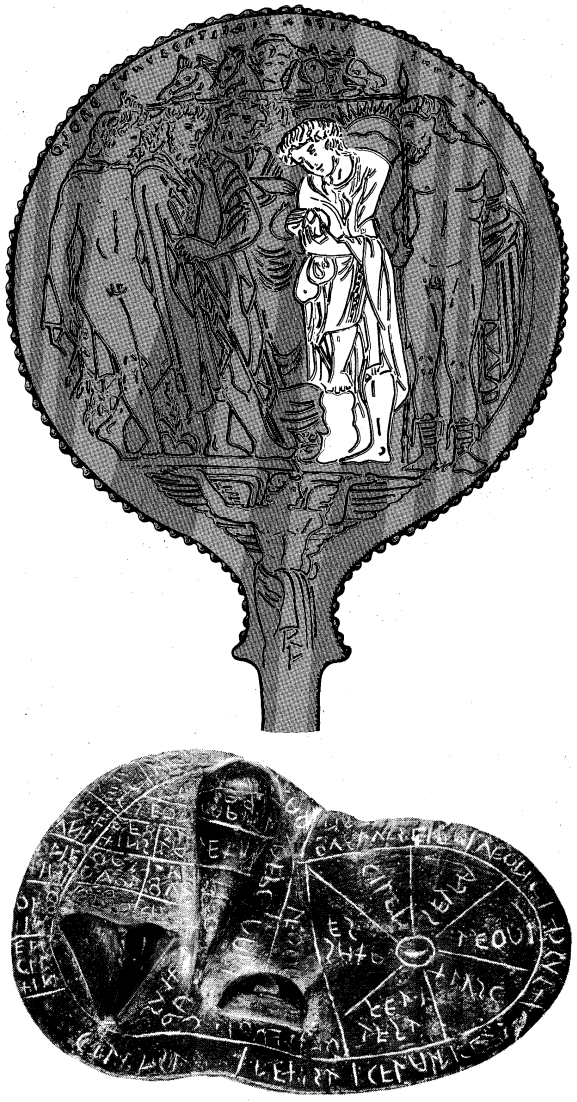
Augury, the art of interpreting omens, is the attempt to discover divine will in phenomena of animate nature. In Mesopotamia, augury was associated with sacrifice and perhaps developed from it. As the priests watched the rising smoke to divine the answer to a ritual query, they observed the movement of birds as auspicious or inauspicious. As a further augury the viscera of the sacrificial victim were examined, and particularly the liver, which (rather than the heart) was conceived as the vital centre. The discipline of augury diagrammed cosmic space with the sacrificial altar at the centre, and each sector was assigned a definite meaning. Every event in the heavens could thus be charted and pondered. In parallel fashion, haruspicy, the meticulous examination of the liver, was developed by mapping that organ in the manner of a microcosm and reading it as one may read the palm.

Inductive divination from nature is associated with the reading of artificially contrived events, such as the movement of sacrificial smoke, the fall of an arrow shot upward, or the cast of dice or lots. A much-used natural–artificial technique consists in the braising of bone or shell to produce a system of signs. Scapulimancy—divination from a fire-cracked shoulder blade—was widespread in North America and Eurasia. The related but more elaborate Chinese technique of tortoiseshell divination was inspired by the idea of equating the carapace (back) and ventral (lower) shell with their view of a rounded sky over flat earth. Only the "earth" was inscribed and heated to produce signs. In general, however, artificial systems of signs are likely to be manipulatory, as they will be used in an artful way by the professional diviner—and in such cases interpretive techniques have to be taken into account.

Interpretive divination. Interpretive divination involves, in the main, the reading of portents, omens, or prodigies. To the scientifically minded person who views the world as subject to an ordering of matter and events, accidents are meaningless. Yet, the accidental does occur within an ordered world and thus is subject to various kinds of interpretation. Manipulated accident is the essential dramatic element of interpretive divination, but the less active forms depend upon projection, introjection, and free association—thus being associated, to some degree, with intuitive techniques.

Forms of interpretive divination

Pyromancy, divination by fire, may be highly dramatic in a society dependent on fire for light and safety at night. In some trans-Saharan societies the diviner may test an accusation at a seance around the fire, which will suddenly explode upon the "guilty" one. Elsewhere, objects may be overtly cast into the fire and signs read in the reaction. Hydromancy, divination by water, is usually less dramatic, ranging from the reading of reflections in a shallow surface, like the crystal gazer, to construing the movements of floating objects, as in the reading of tea leaves.



(Top) Etruscan mirror showing a ritual scrutiny of a sacrificed liver. The various sections of the liver were believed to be a reflection of the divisions of heaven, each of which had a favourable or unfavourable meaning. The man (second from right) holds the liver so that the lobes can be examined and the details for foretelling the future are distinct. From Tuscania, Italy, 3rd century BC. In the Museo Archeologico, Florence. (Bottom) Bronze model of a sheep's liver used for divining. It is divided on its upper surface into about 40 sections, each with the name of an Etruscan divinity inscribed on it. From Piacenza, Italy, 4th–3rd century BC. In the Museo Civico, Piacenza, Italy.

By courtesy of (bottom) the Museo Civico, Piacenza, Italy; from (top) O.W. von Vacano, *The Etruscans in the Ancient World*

A range of related mantic practices may be grouped under the terms cleromancy, divination by lots, and geomancy, which may involve the casting of objects upon a map or a figure drawn on the ground. Cleromantic practices in trans-Saharan Africa may rely on the supposed magical—or indeed horrifying—qualities of objects in the diviner's bag or basket. When they are thrown, the proximity of one piece to another—for example, a dried bit of intestine from a murdered child and a man-eating animal's tooth—may be regarded as having meaning; or the position of a particular piece at the centre or apart from the others may be picked out. Often, the diviner must first prove his ability by discovering the client's problem, through a line of patter accompanying the throws—suggesting this, questioning that, leaping from one matter to another until the reactions of the client betray his interest. Here the diviner may be said to introject ideas and attitudes, while the lots act for the diviner and client alike as a projective device, the meaning of which is only half-formed in the objective pattern cast. A far more elaborate practice is the geomancy of West Africa, in which elegant equipment is combined with impressive erudition to produce a seance in which lots are used to select verses, wherein the client is expected to find his answers. The nature of the lots employed, the number lore on which the selection of verses is based, and the verses themselves are entirely distinct from their counterparts in the Chinese yarrow (an herb with finely dissected leaves) tradition embodied in the *I Ching*, but the general equivalence of the two elaborations is noteworthy. The parallel has perhaps been obscured by the use of "geomancy" in China to signify only a specialized art by which propitious locations are selected.

Sometimes a diviner can be said to interpret signs so characteristic of his client that the practice falls between interpretive and intuitive arts. Somatomancy, body divination, is clearly interpretive in most forms, whether in China or the West, though the system of signs employed comprises private attributes of the client's physique. Examples are phrenology, employing features of the head that are normally unnoticed; and the reading of moles, where the body is treated as a microcosm bearing astrological signs. But oneiromancy, dream interpretation, employs explicitly psychic phenomena; and here the diviner may be said to assist the intuition of meaning by his client as often as he can be said to introject. The Ojibwa and Bella Coola Indians of North America were characteristically preoccupied with the meanings of their dreams.

Intuitive divination. The prototype of the intuitive diviner is perhaps the occasional shaman or curer who uses trance states—which are achieved idiopathically (*i.e.*, arising from the self spontaneously) or induced by drugs or by autokinetic (self-energized) techniques, such as hand trembling among the Navajo, a large North American Indian tribe. As a mantic art, trance is associated with oracular utterance and spirit possession. An impressive performance will be taken to represent the actual voice of a god or spirit, addressing the client directly; and divination in this mode is known from diverse religious traditions, including Christianity. The idea that the gods may be importuned to speak to a matter of temporal human concern is, most likely, very ancient. In early Egypt incubation was practiced—*i.e.*, sleeping in the temple in the hope of being inspired by the resident god. The idea behind Maya maiden sacrifice was the same: a number of maidens were cast into the sacred cenote or deep well, and those who survived after some hours were brought back to recite the messages received during their ordeals —a virtual enactment of the journey into the underworld. As oracular utterance comes to be regularly desired, special techniques or contraptions are developed for making the god's image show assent or denial or for amplifying the sound of an unseen priest's voice. In nomadic societies today, however, the diviner still may achieve his authority by passing into trance before his fellows, trembling and speaking "as if possessed"—that is, as if his own spirit had ceased to inhabit his body and had been replaced by another.

Forms of intuitive divination

Related to possession is the conviction that malevolent persons are essentially unlike innocent folk, though not in outward appearance. When a test is devised for discovering malevolence, commonly conceived as witchcraft or as a nonhuman force dissembling itself in human form, the test takes the form of an ordeal. This may be a demonstration of immunity from harm—since the presence of blessed qualities is viewed as inconsistent with malevolence; among the many types of ordeals are walking on coals and retrieving an object from boiling liquid. The ordeal may also be mortal: in the ordeal by water, a witch was expected to float and so not be spared burning, but an innocent person would be accepted by the water and drown. In trans-Saharan poison ordeals the innocent person is expected to survive, but the malevolent one, it is believed, will be struck dead by the poison.

Intuitive divination may also be a wholly private affair. A Roman who heard a warning from the gods in a bit of innocent conversation was projecting his fears, as was the Aztec dismayed by an animal's howl. The North American Indian who sought a private vision through isolation,

self-mutilation, and fasting would preserve the memory of that vision through youth and later life, turning to it as his unique guardian spirit.

Popularity of horoscopic astrology

Divination today. The immense popularity of horoscopic astrology in the urban West during the latter half of the 20th century illustrates the almost exclusive concern with individual fortune-telling that characterizes divination in a mobile and competitive mass society. Chiromancy, Tarot (fortune-telling) cards, and crystal gazing represent respectively body divination, sortilege (divination by lots), and trancelike performance in styles suitable for what might be called the casual encounter with the idea of fate. Necromancy, in the modern spiritualist form, is suitable for a more serious and sustained effort to establish contact with extramundane beings. But astrology, of the various popular forms, is best suited to mass distribution as printed matter, since it is based on a well-articulated body of lore, touches matters of high destiny as well as individual fortune, and "personalizes" its introjective advice without the necessity of interviewing the client. On the other hand, the more esoteric mantic arts have the appeal of discipline—an individual may enter into the lore deeply until it becomes a part of his own world view and affects his attitude toward life and himself. Such involvement is demonstrated on the part of those individuals who study the *I Ching* for divinatory purposes.

BIBLIOGRAPHY. The scope of divination, historically and ethnographically, is well represented in a recent work by specialists in many fields, *La Divination*, 2 vol., ed. by ANDRE CAQUOT and MARCEL LEIBOVICI (1968). For a briefer introduction the reader may wish to consult the pertinent section in W.A. LESSA and EVON Z. VOGT (eds.), *Reader in Comparative Religion*, 2nd ed. (1965). The classic treatment of modern times is A. BOUCHE-LECLERCQ, *Histoire de la divination dans l'antiquité*, 4 vol. (1879–82). Recent studies of Mesopotamian, classical Mediterranean, and Chinese systems of divination are: J. NOUGAYROL (ed.), *La Divination en Mésopotamie ancienne et dans les régions voisines* (1966); R. FLACELIERE, *Devins et oracles grecs* (1961; Eng. trans., *Greek Oracles*, 1965); G. DUMEZIL, *La Religion romaine archaïque* (1966; Eng. trans., *Archaic Roman Religion*, 2 vol., 1970); WILLIAM A. LESSA, *Chinese Body Divination* (1968); C.K. YANG, *Religion in Chinese Society* (1961); and HELMUT WILHELM, *Die Wandlung: Acht Essays zum I-Ging* (1958; Eng. trans., *Change: Eight Lectures on the I Ching*, 1960).

Three ethnographic studies of African divination are significant: HENRY CALLAWAY, *The Religious System of the Amazulu* (1870); E.E. EVANS-PRITCHARD, *Witchcraft, Oracles, and Magic Among the Azande* (1937); VICTOR W. TURNER, *Ndembu Divination* (1961). Detailed studies of West African geomancy are WILLIAM BASCOM, *Ifa Divination* (1969), and BERNARD MAUPOIL, *La Géomancie à l'ancienne côte des esclaves* (1943). Another ethnographic study is EVON Z. VOGT and RAY HYMAN, *Water Witching, U.S.A.* (1959).

(G.K.P.)

Djakarta

Djakarta (Jakarta, formerly Batavia) is the capital of the Republic of Indonesia and the largest and most consistently growing city in that country. Coextensive with the metropolitan district of Djakarta Raya, it has an area of 223 square miles (578 square kilometres) and lies at the mouth of the Tji (river) Liwung (Chiliwong) on the northwestern coast of Java. Its population in 1971 was 4,576,009.

In 1966 the city was declared to be a special capital region (Daerah Chusus Ibukota, or DCI), thus gaining a status approximately equivalent to that of a state or province. The city has long been a major trade and financial centre; it has also become an important industrial city and an important centre for education.

History. The first settlements were established at the mouth of the Tji Liwung, perhaps as early as the 5th century. Its official history, however, starts in 1527, when the Sultan of Bantam defeated the Portuguese there and called the place Djajakerta, meaning Glorious Fortress.

The Dutch, under the leadership of Jan Pieterszoon Coen, captured and razed the city in 1619, after which he established the capital of the Dutch East Indies—a walled township named Batavia—on the site.

The colonial history of the city can be divided into three major periods. First was that of the Dutch East India Company, when most of the activities of the city centred around the fortress and the company warehouses. At that time the city somewhat resembled a typical Dutch town, complete with canals. The second period began in the early 1800s, when the city was extended to include higher and more healthful areas to the south, which would later become the seat of the new colonial government. A brief interval of British control during the Napoleonic Wars, ending in 1815, interrupted the second period. During the third period, which lasted from about the 1920s to 1941, the city became a truly modern tropical city.

The colonial era

The colonial era ended with the entry of Japan into World War II, when Indonesia was occupied by Japanese forces. After the war the city was briefly occupied by the Allies and then was returned to the Dutch. When Indonesia gained full independence on December 27, 1949, the city was renamed Djakarta and proclaimed the national capital.

The present city. Djakarta lies on a low, flat alluvial plain, with extensive swampy areas; it is easily flooded during the rainy seasons. The parts of the city further inland are slightly higher. As most of the soil is of old volcanic origin, the area is quite fertile.

Climate and environment. Djakarta is a tropical, humid city, with temperatures ranging between the extremes of 75° and 93° F (24° and 34° C) and a humidity of between 75 and 85 percent. The average mean temperatures are 79° F (26° C) in January and 82° F (28° C) in October. The annual rainfall is more than 67 inches. Temperatures are often modified by sea winds.

The draining of swamps for building purposes has increased the probability of floods. With an excess of water in the soil, Djakarta experiences a shortage of clean drinking water, for which there is a high demand. Air pollution has not yet become a problem.

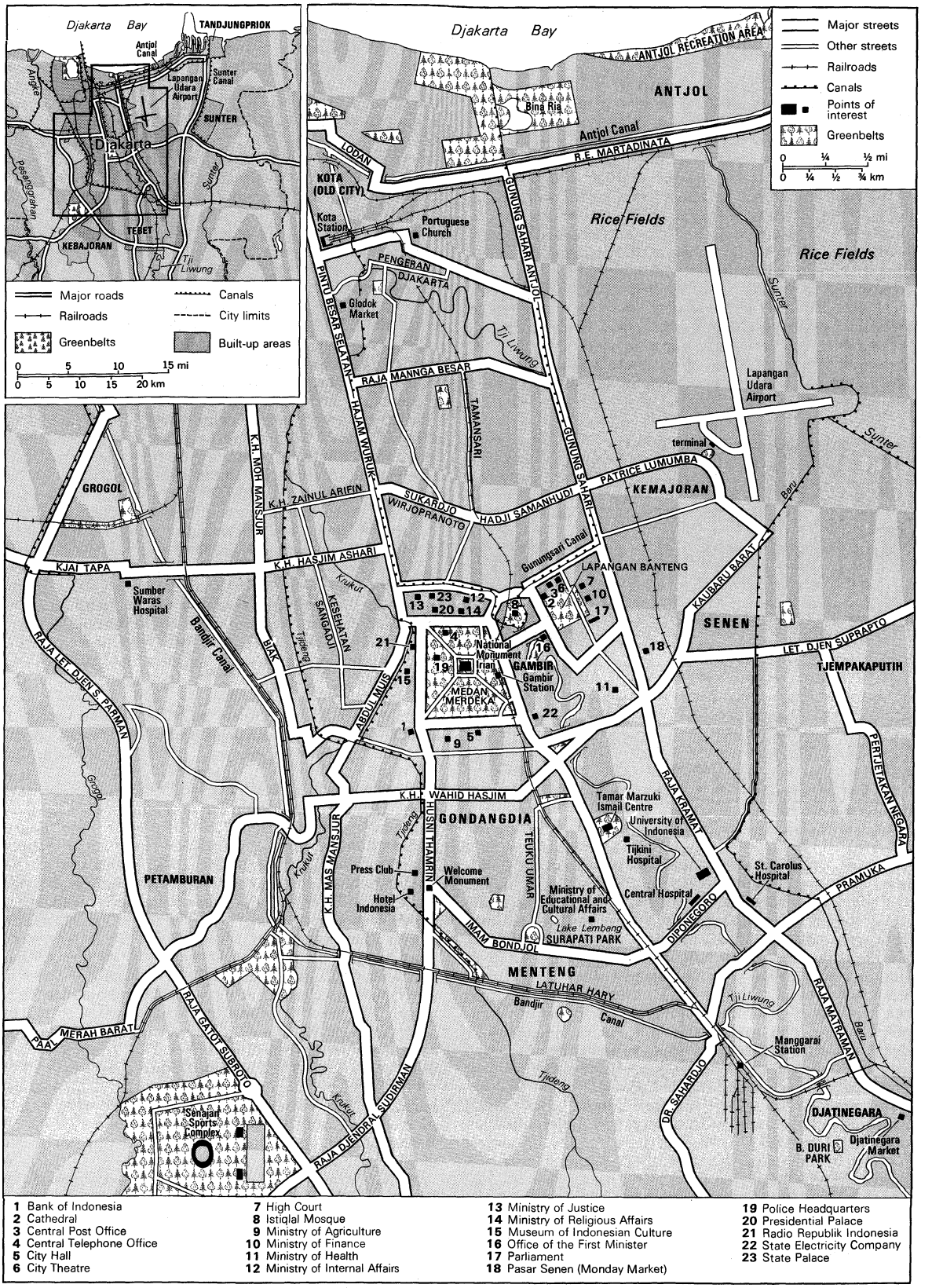
City plan and traditional neighbourhoods. Although the Dutch were the first to attempt to plan the city, the contemporary city layout is probably more British than Dutch in character, as can be seen from such large squares as the Medan Merdeka (Freedom Field) and Lapangan Banteng (Place of the Gaur [large wild ox]). The Oriental style, or "indische" style, as the Dutch call it, is, however, not only apparent in the city's way of life but also in the types of houses, the wide, tree-lined streets, and the spacious gardens and house lots. In Kebajoran, a satellite town built since World War II on the southwest side of the city, the houses and garden lots are much smaller than in the older colonial districts.

A city of new settlers

Djakarta has always been a city of new settlers who assimilated local ways and became Djakartans themselves. Some traditional neighbourhoods can, however, be identified. The Kota (Fort), or Old City, for example, sometimes also called the downtown section, is the central business district and also the financial capital of Indonesia. It houses a significant part of the Chinese population. The area of the Kemajoran (Progress) and Senen sections, originally on the eastern fringe of the city, is now almost central in its location and increasingly is becoming the major retail area of the city.

The Djatinegara (Real Country) section, originally a Sundanese settlement but later incorporated as a separate town, after which it became a Dutch army camp, has now merged with the rest of Djakarta and includes many new settlers. The Menteng and Gondangdia section is a fashionable residential section near the central part of the city. To the west, the Tanahabang Petamburan section is, like the Kemajoran and Djatinegara sections, densely developed. Tandjungpriok is the community attached to the harbour.

Demography. *Population and population growth.* The population of Djakarta at the 1971 census was 4,576,009. Pronounced growth had occurred only since 1940. According to the 1930 census, the city's population was 530,000; the estimate for 1940 was still only 540,000, but by 1950 it had risen to 1,400,000, and at the 1961 census it was 2,900,000; the 1965 estimate was 3,800,000. About half of this extraordinary rate of increase (about



Djakarta and (inset) its metropolitan area.

4 percent annually) is attributed to immigration. Although government regulations close the city to unemployed new settlers, better economic conditions inevitably attract new people. In addition, much of the population is young and fertile, resulting in a very high natural increase potential. In the near future, therefore, the same rate of increase is expected to continue. Meanwhile, most urban services are operating at capacity.

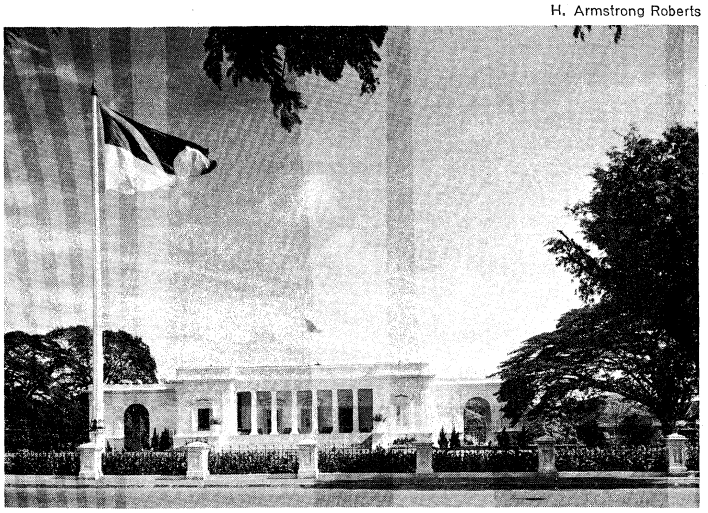
Composition and employment. Analysis of the immigrant stream shows that after the West Javanese, the largest groups are the Central and East Javanese; a sizable number also are from Sumatra. Other population groups—Arabs, Indians, Europeans, and Americans—are minor.

Among the working force the largest group is engaged in commerce, trade, banking, or related occupations;

the next group works in service industries (including government); and in the third largest group are industrial and transportation workers (including a large group of *betjak* drivers).

Housing. Because of the relatively equable climate, housing, as such, is not considered a primary problem. The most common type of house in the city is the *kampong* ("village") house; most are built of materials such as wood or bamboo mats, but this does not necessarily mean that they are substandard. Another common type of housing, often used to house government workers, is the colonial urban house, or *rumah gedongan;* these are mostly single-family detached or semidetached houses, each standing on a separate lot. Apartment buildings constitute a newer category; although they are more economical in the use of land than single-family types, their architectural and construction costs often make them fairly expensive. Housing is generally overcrowded. It is estimated that slightly more than 60 percent of all households have only one room; there are, on the average, about five persons per household and about three persons per room.

H. Armstrong Roberts



Presidential Palace, Djakarta.

Buildings from colonial times

Historic and architectural features. Some of the buildings, such as the old Portuguese Church (1695) in the Old City, are of architectural or historical interest. Some of the buildings around the city square in the Kota also date from colonial times, including the old city hall, restored as the municipal museum. The National Archives building was originally the palace (1708) of a Dutch governor general, van Riebeek. In the present city centre the Ministry of Finance building, facing Lapangan Banteng Square, and the Presidential Palace, facing Merdeka Square, both date from the same period. Merdeka Square is the site of the National Monument Irian (at 360 feet, the highest building in Djakarta); there is also a monument on Lapangan Banteng Square. The Istiqlal Mosque, near the northwest corner, still under construction in the early 1970s, was designed to be the largest mosque in Indonesia, if not in all of Southeast Asia.

After World War II Djakarta experienced a building boom. The Hotel Indonesia (the city's first high-rise building) and the Senajan Sports Complex were built for the Asian Games in 1962. Most high-rise buildings stand along Husni Thamrin Road.

Economic life. Economically, Djakarta can be identified first as the national capital and a central place of control for the national economy, then as an administrative centre in its own right, and, finally, as a significant industrial hub. In addition, its location as a port makes it an important centre for trade.

The cost of living in the city in the early 1970s tended to rise. Land was very expensive, and rents were high, so that industrial development and the construction of housing usually was undertaken on the outskirts, while commerce and banking remained concentrated in the city centre. The Indonesian Chamber of Commerce is active in promoting trade with other countries; the annual Djakarta Fair also serves to promote trade.

City-operated markets

To meet the needs of the local city population, the municipality operates several markets. The Central City Markets (Pasar Kota), like the markets of Pasar Senen to the east of the Central City and Pasar Glodok in the Kota area, are major retail centres. The Pasar Djatinegara is primarily a food supply centre. The district markets are fairly large, with each one catering to a whole section of the city. There are also small neighbourhood markets, each serving only a limited area. Special markets include one selling fish, one selling automobile parts, and the Pasar Rumput "fly market." Djakarta also has about 90 general markets.

Transportation. *Roads and railways.* Major road arteries lead west from the centre of the Old City and east and south from the centre in Gambir. To the east, a major railroad connects the city with all of the island of Java. There is also a highway, primarily a regional supply road, running between Djakarta and the productive areas of east and central Java. To the south, a major road and railroad connect Djakarta with Bogor, Sukabumi, and, ultimately, also Bandung. To the west, a railroad and road run to Banten and to the harbour in Merak, which is connected by ferry to Lampung in Sumatra.

Sea and air links. Djakarta's Tandjungpriok Harbour is the largest in Indonesia, handling exports from Djawa Barat (West Java) and a large proportion of Indonesia's import trade; many goods are later transshipped to other islands or harbours.

Djakarta is served by international airlines, Garuda Indonesian Airway (the national airline, with international and domestic service), and other domestic airlines. In the early 1970s the city's airports were at Kemajoran and Halim Perdanakusuma; a new jet airport was planned at Tjengkareng, about eight miles east of the city.

Terminals and urban traffic. The central bus terminal, located on Lapangan Banteng, serves all the city, intercity, and regional bus lines; there are also five suburban bus terminals—in Djatinegara (Tjililitan), Kebajoran, Grogol, Kota, and Tandjungpriok. The major railroad terminal is in the Old City. Traffic jams occur particularly during the morning and afternoon rush hours. Although the number of vehicles is increasing, the number of automobiles is relatively small. Public transportation, by either bus or *betjak* (a two-passenger tricycle taxi), is still reasonably convenient and inexpensive, which may be why the automobile-type taxis are still almost absent.

The *betjak*, or tricycle taxi

Government and services. The mayor of the city has the same status as the governor of a province. The city government is composed of the executive and the electorate. The executive consists of a governor assisted by four vice governors, an executive staff, and a regional secretary; there are also a number of city directorates, bureaus, and agencies. The electorate consists of 35 to 40 members, including representatives of eight political parties, four representatives of the armed forces, and nine representatives of the so-called functional groups. It is headed by a council of five members, one chairman, and five vice chairmen.

Public utilities. Public utilities are usually operated or owned by the Indonesian government. The State Electricity Company and the subsidiary State Gas Company both supply Djakarta. Postal and cable services and telephone services are supplied by state companies working under the aegis of the Ministry of Communications. Djakarta's electricity comes from several sources; these include the thermal plant in Antjol, close to Tandjungpriok Harbour, smaller diesel plants in various parts of the city, and the Djatiluhur hydroelectricity project located close to Purwakarta, about 70 miles east of Djakarta.

The city government is responsible for the water supply; as already mentioned, there is a shortage of drinking water. The city water is obtained in part from freshwater springs in the Bogor area, but most of the supply

comes from the Pedjompongan water treatment plant. Water is required mainly for domestic purposes but is also needed for industry and to supply shipping. The removal of garbage and the provision of other sanitation services are also the responsibility of the municipality.

Health and safety. There are three major hospitals—one operated directly by the Ministry of Health, one by the Roman Catholic Church, and one by a Protestant mission. Three municipal hospitals each serve a separate area of the city—the Sumber Waras Hospital in west Djakarta, the Fatmawati Hospital in Kebajoran (south Djakarta), and the Persahabatan (Friends) Hospital in east Djakarta. Altogether Djakarta has almost 40 general or special hospitals, with a capacity of about 2,000 beds in the early 1970s. In addition, there are more than 400 general clinics or polyclinics scattered throughout the city. A quarantine hospital is in operation in Tandjungpriok. The city also operates a hospital and rehabilitation centre for the mentally ill and socially deprived, and there are almost 50 family-planning clinics.

Hospitals and clinics

There are 18 fire-brigade posts, nine of which are in the central area. Djakarta has a low crime rate. The city police are more concerned with maintaining general order and regulating traffic than with combating organized crime.

Education. To meet the needs for primary education, many new elementary schools and secondary schools were built in the late 1960s; a number of old school buildings were also renovated. Altogether, there were about 270 kindergartens, 1,200 elementary schools, 230 *madarsah* ("religious schools"), 300 secondary schools, and 120 high schools. There were also about 550 vocational and special schools and more than 100 universities, academies, and institutes for higher learning. The largest and best known university is the Universitas Indonesia; its most important departments are the school of medicine, the school of law, and the school of economics.

Cultural life, recreation, and mass media. Among other cultural activities, the Tamar Marzuki Ismail Centre has facilities for traditional or classical art performances as well as theatres for presenting modern plays and concerts; the centre also has a planetarium. Traditional performances include *wayang* dance and drama, gamelan music, and *wayang* puppet shows. Traditional performances representing the culture of other parts of Indonesia are included in the programs presented at the annual Djakarta Fair.

Extensive public recreation areas include the Bina Ria seaside recreation area and the Ragunan zoo, close to Pasarminggu. Playgrounds include, among others, the Taman Ria complex at the Djakarta Fair. Public recreation facilities include a bowling centre, a greyhound racecourse, and a horse racecourse.

There are about 25 major newspapers in the city. Of these the *Berita Yudha, Kompas, Indonesia Raya, Merdeka, Pedoman,* and *Sinar Harapan* have the largest circulation. There are two English-language newspapers, the *Indonesian Observer* and the *Djakarta Times. Indonesia* is a newspaper printed in Chinese. There are three national news agencies; numerous foreign and international news agencies and reporters are also based in Djakarta. There are about five general weekly magazines.

In addition to the Radio Republik Indonesia, the national radio system, some private radio companies also broadcast programs from Djakarta. The national television (TVRI) operates from the city and has a network covering most of the island of Java. Since the opening of the earth-satellite station in Djatiluhur, the Indonesian television network can be connected with programs in European or other Asian countries.

BIBLIOGRAPHY. On the history of Djakarta, see F. DE HAAN, *Uit Oud-Batavia* (1898); and DJAKARTA, KOTAPRADJA, *Sedjarah pemerintahan kota Djakarta* (1958). A contemporary description is given in the guidebook, *Djakarta,* issued by the PETUNDJUK DCI (1969). Census and population information may be found in BIRO PUSAT STATISTIK, *Sensus penduduk 1961 D.C. I. Djakarta Raya* (1963); and in H.J. HEEREN (ed.), *The Urbanisation of Djakarta* (1955). See also *Djakarta: Its Rehabilitation and Development* (n.d.), issued by the Badan Perentjana Pembangunan (Development Planning Body) of the City Government; and PAULINE D. MILONE, *Urban Areas in Indonesia* (1966).

(W.J.W.)

Dnepr River

The Dnepr (conventional English spelling Dnieper, Ukrainian Dnipro, Belorussian Dnepro, the Borysthenes of ancient Greek authors) is the second river in length and basin area in the European part of the U.S.S.R. and the third longest in Europe, after the Volga and Danube. It is 1,400 miles (2,200 kilometres) long and drains an area of about 195,000 square miles (504,000 square kilometres).

The Dnepr rises at an altitude of 721 feet in a small peat bog on the southern slope of the Valdai Hills, 46 miles northwest of Vyazma, and flows into the Dnepr Estuary of the Black Sea. For the first 300 miles, it passes through the Smolensk (Smolenskaya) *oblast* of the Russian Soviet Federated Socialist Republic (R.S.F.S.R.), first to the south and then from Dorogobuzh to the west; near Orsha it turns south once more and for the next 370 miles flows through the Belorussian Soviet Socialist Republic. Next, it flows through Ukrainian territory: south to Kiev, southeast from Kiev to Dnepropetrovsk, and then south-southwest to the Black Sea.

The Dnepr watershed includes the Volyno-Podolsk Upland (Volino-Podolskaya Vozvyshennost), the Valdai Hills, the Belorussian Ridge (Belorusskaya Gryada), the Central Russian Upland, and the Priazovsk Upland (Priazovskaya Vozvyshennost). The centre of the basin consists of broad lowlands. Within the forest area, and to some extent within the forest–steppe area, the basin is covered with morainic and fluvioglacial deposits; on the steppe, it is covered with loess. In some places, where the basin borders upon the basins of the Western Bug and the Western Dvina, there is a flat swampy area. This facilitated the cutting of connecting water routes from the Dnepr to neighbouring rivers even in ancient times. At the end of the 18th century and the beginning of the 19th, the Dnepr was connected to the Baltic Sea by several canals; the Dnepr–Bug Canal (Dneprovsko-Bugsky Kanal) running by way of the Pripyat, the Western Bug, and the Vistula; the Oginsky Canal by way of the Pripyat and the Neman; and the Berezina water system by way of the Berezina and the Western Dvina. These canals later became obsolete.

The Dnepr watershed

The natural environment. The Dnepr is customarily divided into three parts: the Upper Dnepr as far as Kiev, the Middle Dnepr from Kiev to Zaporozhye, and the Lower Dnepr from Zaporozhye to the mouth. The basin of the Upper Dnepr is mainly within a forest area where peat–podzolic soils predominate (replaced in the southern portion of the upper course by podzolized, gray forest soils). The Upper Dnepr is characterized by excessive moisture and great swampiness. The river network is well developed here, for this is where 80 percent of the basin's annual runoff forms and the longest, most water bearing tributaries (the Berezina, Sozh, Pripyat, Teterev, and Desna) flow. The basin of the Middle Dnepr is in a forest–steppe area with black earth. There are forests in the watersheds and along the river valleys. The river network is thinner here, and the rivers contain comparatively little water. The principal tributaries of the Middle Dnepr are the Ross, Sula, Psyol, Vorskla, and Samara. The Lower Dnepr basin lies within the Black Sea Lowland, in the black-soil steppe area, which has now been completely plowed up. The grassy steppe vegetation has been preserved only in the national forests and in old ravines and gullies. Near the Black Sea there is wormwood–fescue vegetation of the semi-arid type in chestnut-brown soil mixed with saline solonetz and solonchak soils. The Lower Dnepr passes through a region of insufficient moisture, where irrigation and flooding are employed. The river network here consists for the most part of dry river beds (ravines) that fill up with water in the spring and after torrential rains. The largest tributary is the Ingulets.

The climate of the Dnepr Basin is, on the whole, moder-

ately warm, milder, and damper than that of more eastern regions of the R.S.F.S.R. located at the same latitude.

Climates of the river's course

The continental nature of the climate increases from northwest to southeast. The mean annual air temperature in the upper part of the basin is 41° F (5° C); in the middle (near Kiev), 45° F (7° C); and in the lower reaches of the Dnepr, 50° F (10° C). Winters in the northeast of the basin are long and of steady intensity, the mean temperature in January being 16° F (−9° C); whereas in the south they are shorter and milder with frequent thaws, the mean temperature in January being 27° F (−3° C). The amount of precipitation decreases from north to south. On the slopes of the Valdai Hills and the Minsk Heights (Minskaya Vozvyshennost), annual precipitation is about 30–32 inches, while in the Lower Dnepr region it is about 18 inches. The mean annual precipitation for the Upper Dnepr Basin (above Kiev) is about 28 inches, with about 27 inches for the entire basin, with 60–70 percent falling in the form of rain during the summer and fall.

From its source to Dorogobuzh, the Dnepr is a small river flowing past low wooded and, in some places, swampy banks. Downstream the banks rise, and the width of the depression before Orsha varies for the most part from two to six miles, narrowing to 0.3 mile in places. Its bed, from 130 to 400 feet wide, is sinuous but steady, with numerous sandbanks. Above Orsha the Dnepr crosses a layer of Devonian limestone, forming the so-called Kobelyaki Rapids (Kobelyakskiye Porogi), which hamper navigation into the lower waters. From Orsha to Shklov, the Dnepr flows between raised, sometimes steep banks overgrown with woods; then the left bank drops, while the right remains high as far as the Sozh Basin. The depression is wide, reaching six to nine miles in places. The river bed from Orsha to Mogilyov is comparatively steady; below Mogilyov, the Dnepr splits into branches, producing many islands and sandbanks. The width of the river from Orsha to the mouth of the Sozh is 260–1,300 feet, and from the mouth of the Sozh to the mouth of the Pripyat it is 1,600–2,000 feet. The tidal vegetation of the Upper Dnepr consists mainly of rather broad tidal meadows, thickets of willows and alders, and old lowland marshes.

Marked asymmetry of the depression is characteristic of the Middle Dnepr. The steep, high right bank (up to 130–260 feet) forms the escarpment of the Volyno-Podolsk Plateau, which stretches along the whole middle course of the river. The low and sloping left bank is formed by broad, ancient terraces. Single hills, reaching over 300 feet in height, sometimes rise on the low-lying left bank—for example, Pivikha Hill near Gradizhsk. On the southern portion of the Middle Dnepr, the river cuts through the Ukrainian crystalline masiff and flows for 56 miles in a narrow, almost unterraced valley bounded by high rocky banks. The famous Dnepr Rapids, which for centuries prevented continuous navigation, were once located here. The rapids were flooded by the backwaters of the Dnepr hydroelectric power station dam, which raised the level of the river by 121 feet.

The lower river

Below Zaporozhye, the Dnepr again passes into a very wide valley with a high right bank (130 feet near Nikopol, 260 feet near Kherson). The slopes of the river here are very slight. Before the development of the Kakhovka Reservoir (Kakhovskoye Vodokhranilishche), the waters of which overflowed a vast territory, the Dnepr split into a multitude of streams; flat swampy islands, overgrown with tidal vegetation and reeds, lay among the channels. Today, much of this is hidden under the waters of the Kakhovka "sea."

Below Kherson, the Dnepr forms a delta whose numerous streams flow into the Dnepr Estuary. Some have been deepened for navigational purposes.

The water conditions of the Dnepr have been rather thoroughly studied. The data on the river's annual runoff go back to 1818, while the maximum discharges, computed from the old high-water marks, go back more than 250 years. More than 300 hydrometric stations and posts operate in the Dnepr Basin. The Dnepr is among those rivers that have a spring high and a summer and winter low, with increased runoff in the fall. The principal source is the water entering the upper part of the basin as a result of the spring snowmelt. During the spring, from March to May, about 60 percent of the annual runoff passes through. The period of stable ice on open water in the Upper Dnepr sets in at the beginning of December, and in the Lower Dnepr at the end of December. Thaw starts at the beginning of April in the upper course, and in early March in the lower course. The mean annual runoff of the river is 13 cubic miles. In individual years, the variations in the runoffs can be quite considerable. The water of the Dnepr is low in minerals and soft. In a year the river carries an average of 8,600,000 tons of dissolved matter to the sea.

Flora and fauna

The Dnepr has rather diverse aquatic flora and fauna. In its upper course the plankton consists mainly of diatom and protococcal algae, rotifers, and *Bosmina*. Blue-green algae come from the mouth of the Pripyat. In its lower course, the amount of plankton decreases sharply under the influence of the reservoirs. Up to 66 species of fish live in the Dnepr. The following are of commercial importance: pike, roach, chub, ide, rudd, rapfen, tench, barbel, alburnum, golden shiner, goldfish, carp, catfish, burbot, pike perch, perch, and ruff. In the spring, the Lower Dnepr is enriched with migratory and semimigratory fish (sturgeon, herring, roach, and others). The reservoirs have been artificially stocked with fish of commercial importance, including whitefish, pike perch, golden shiner, and carp.

The human imprint. The Dnepr Basin has been populated since ancient times. It was of central importance in the history of the peoples of eastern Europe, particularly in the founding of the ancient Kievan state. Along this waterway a system of river routes developed in the 4th to 6th centuries, a "route from the Varangians to the Greeks," connecting the Black Sea with the Baltic and linking the Slavs with both the Mediterranean and the Baltic peoples. Most of the Dnepr (677 miles) passes through territory of the Ukrainian Soviet Socialist Republic, and the river is for the Ukrainians the same kind of national symbol that the Volga is for the Russians.

The first historical information about the Dnepr is recorded by the Greek historian Herodotus (5th century BC); the river is also mentioned later by the ancient writers Strabo and Pliny the Younger. It was first depicted on a map drawn by the Byzantine geographer Ptolemy in the 2nd century AD. Instrument surveys of the Dnepr were begun at the beginning of the 18th century.

River development

Under the Soviets, in line with the general plan for development of the water economy, a great deal of work has been done on the reconstruction of the Dnepr and the comprehensive use of its water resources. As early as 1932, in accordance with the country's electrification plan, as formulated by the State Commission for the Electrification of Russia (Goelro), the first hydroelectric power station was built at Zaporozhye in the region of the rapids. It was the largest power station in Europe, with a capacity of 560 megawatts, until the construction of the Volga power stations. Completely destroyed by the Germans during World War II, it was rebuilt in 1947 and its capacity was raised to 650 megawatts. By the early 1970s the following hydroelectric power stations and reservoirs had been built on the Dnepr: Kiev (1967), Kanev (construction being completed), Kremenchug (1961), Dneprodzerzhinsk (1964), Zaporozhye (1932, 1947), and Kakhovka (1955). The total capacity of the five stations in operation is more than 2,000,000 kilowatts, with an electrical power output of 8,000,000,000 kilowatt-hours per year. As a result of their construction, many problems have been solved: a continuous deep-water route from the mouth of the Pripyat to the Black Sea has been created; the water-supply problem of the Donbass and Krivoy Rog industrial regions has been solved; and irrigation of arid lands in the southern Ukraine and the Crimea has been made possible.

Regular navigation on the Dnepr extends as far as Orsha, and, when the water is high, to Dorogobuzh. On the Upper Dnepr the required depths are maintained by straightening and by dredging work. Below the Pripyat,

navigable locks make the passage of modern vessels possible. The principal cargoes are coal, ore, mineral building materials, lumber, and grain. The chief ports are Dorogobuzh, Smolensk, Orsha, Mogilyov, Rechitsa, Loyev, Kiev, Cherkassy, Kremenchug, Dnepropetrovsk, Zaporozhye, Nikopol, Kakhovka, and Kherson.

The Krivoy Rog region is supplied with water from the Kakhovka Reservoir by means of the Dnepr–Krivoy Rog Canal. The North Crimea Canal, which was completed in 1975, originates in the reservoir; the canal, 250 miles long, is designed for irrigation of the steppes of the Black Sea Lowland and the northern Crimea and for the creation of a water route from the Dnepr River to the Sea of Azov.

Plans for the future are comprehensive. The area of irrigated land will be increased and new navigable routes will be created. The construction of the Dnepr–Donbass Canal and other canals, the development of a modern water route from the Dnepr to the Baltic Sea through the Neman River, and the reconstruction of the river above the Pripyat are among the projects contemplated.

(A.P.D.)

Dnestr River

The Dnestr River (Greek Donaster, Romanian Nistrul, Turkish Turla, English conventional Dniester) is the second longest river in the Ukrainian Soviet Socialist Republic and the main water artery of the Moldavian Soviet Socialist Republic. It rises on the northern slopes of the Carpathian Mountains in the Lvov *oblast* (region) of the Ukrainian S.S.R. and runs 839 miles (1,350 kilometres) in a generally southeastern direction to the Black Sea near Odessa. For roughly half its length it flows through the territory of the Moldavian S.S.R.

The river basin

The Dnestr and its tributaries drain a long, narrow basin of about 28,000 square miles (72,000 square kilometres), bounded on the north by the Volyno-Podolsk Upland and on the south of the upper course by the Carpathian Mountains; farther to the south are hilly plains and the Moldavian plateau; and at the southernmost end of the basin is the Black Sea Lowland.

The Dnestr has many tributaries, only 15 of which are more than 60 miles long. In the first third of its course the tributaries include the Stry, Svicha, Lomnitsa, and Bystritsa, on the right, originating in the Carpathians. In the large middle part of its course the tributaries include the Zolotaya Lipa, Strypa, Seret, Smotrich, Ushitsa, and Murafa, on the left. In the lower reaches of the Dnestr, the tributaries are mainly on the right, including the Reut, the Byk, and the Botna.

The climate of the basin is humid, with warm summers. Annual precipitation varies within a range of from 40 to 50 inches in the Carpathians down to 20 inches near the Black Sea. A large proportion of the surface of the basin is under cultivation.

The Dnestr has three parts: the upper river as far as the village Nizhny, about 170 miles; the middle river from Nizhny to Dubossary, about 450 miles; and the lower river from Dubossary to the mouth, about 210 miles. For first 30 miles of the river's course, it is a rushing mountain stream, flowing through a deep gorge, but from the mouth of the Stry through the rest of its upper course the valley is several miles wide. In its middle course the river meanders a great deal, winding for 300 miles over the 145-mile distance between Galich and Mogilyov-Podolsky.

At Dubossary there is a hydroelectric station and reservoir. Below Dubossary the valley gradually widens to 12 miles with slopes several hundred feet high, bearing orchards, vineyards, and forests. After Bendery the floor of the valley becomes marshy and broken with small lakes and islands. The estuary of the Dnestr is formed by the incursion of the sea into the lower Dnestr Valley, forming a shallow basin (no more than ten feet deep) separated from the sea by a narrow strip of land. The length of the estuary is 27 miles, and its greatest width is seven miles.

The Dnestr frequently floods, causing extensive damage to settled areas. The water level in its middle course varies by 25–35 feet at different times of the year because of snow melting and rainfall in the upper part of its basin. The average discharge is about 10,000 cubic feet per second (300 cubic metres), but it has been known to reach 250,000 cubic feet per second or more in the middle course. Freezing usually occurs at the end of December or the beginning of January and lasts about two months, although in some years there is no ice-bound period.

Although the basin of the Dnestr is densely populated, there are no large towns on the river itself. Lvov, Ternopol, Stanislav, Kishinyov, and other urban centres lie above the main valley on tributaries.

The Dnestr is navigable for about 750 miles, from its mouth to Rozvaduv; regular passenger and freight lines run from Soroki to Dubossary and from Dubossary to the sea. Navigation is made difficult on the lower reaches by shallow water and sandbars. The river is used extensively for carrying logs, which are brought together at the mouths of the Carpathian tributaries and rafted downstream. Fishing is of little importance except near the coast. In the lower reaches and in the Dubossary Reservoir there are fish hatcheries for sturgeon, whitefish, pike perch, and carp.

Development of the Dnestr

The first description of the lower part of the Dnestr was published in 1711 by the Moldavian prince Dmitry Kantemir; the first hydrographic map of the Dnestr was made in 1787 under Catherine II. In the 18th century a plan was made to join the Dnestr with the Western Bug and the San (tributaries of the Vistula). Detailed studies and plans were made in the 19th century, having as their goal the protection of the upper part of the basin from flooding and the improvement of navigation conditions in the middle and lower parts.

The hydroelectric station at Dubossary, built in 1954, has a capacity of about 50,000 kilowatts. Several small water power installations exist in the upper reaches of the river. In the summer a large number of floating mills, operated by the swift current, are set up on the river. Plans in the early 1970s called for the construction of a series of five hydroelectric stations above Dubossary (at Kamenka, Yampol, Mogilyov-Podolsky, Zhvanchik, and Unizh) with a total capacity of 2,300,000 kilowatts, as well as the improvement of the irrigation system on both banks of the Dnestr.

(A.P.D.)

Dobzhansky, Theodosius

The research and publications of Theodosius Dobzhansky, prominent American biologist and geneticist, combined the study of heredity and evolution. His work has been a major influence on 20th-century thought and research on genetics, evolutionary theory, and human evolution. He also wrote extensively on the philosophical implications of biological evolution.

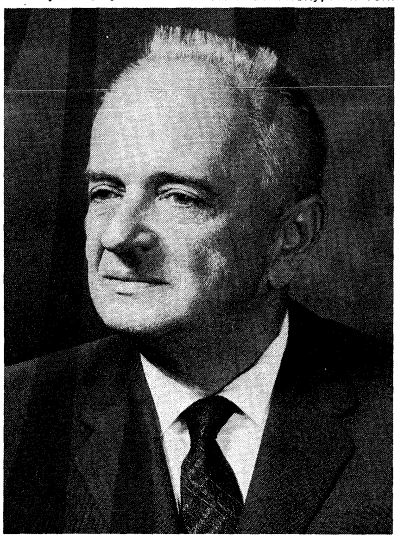
Dobzhansky was born on January 25, 1900, in Nemirov, Russia, where his father was a teacher of mathematics. In 1910 the family moved to Kiev, where he first went to school. In 1917 he entered the University of Kiev, where he studied during the difficult days of the Russian Revolution. After graduating in 1921, he remained to teach until 1924, when he moved to Leningrad. In 1924 he married Natalia Sivertzev, whom he met while teaching at Kiev.

In 1927 Dobzhansky went to Columbia University in New York City as a Rockefeller Fellow to work with the geneticist Thomas Hunt Morgan (*q.v.*). He accompanied Morgan to the California Institute of Technology in Pasadena and, on being offered a teaching position there, decided to remain in the United States, becoming a citizen in 1937. He returned to Columbia as a professor of zoology in 1940, remaining until 1962, then moved to Rockefeller Institute (later Rockefeller University). After official retirement, he moved in 1971 to the University of California at Davis.

Between 1920 and 1935, mathematicians and experimentalists began laying the groundwork for a theory combining Darwinian evolution and Mendelian genetics. Starting his career about this time, Dobzhansky was in-

volved in the project almost from its inception. His book *Genetics and the Origin of Species* (1937) was the first substantial synthesis of the subjects and established evolutionary genetics as an independent discipline.

By courtesy of The Rockefeller University, New York



Dobzhansky.

Until the 1930s, the commonly held view was that most species in the wild state carried only one, either the dominant or the recessive, gene (*i.e.*, the normal type in wild populations) for each characteristic. This wild-type gene, inherited from both parents, was present in double dose in most individuals (homozygous), and variant genes usually represented a chance mutation. Most mutations would be harmful when homozygous, but rarely a mutation would appear that was advantageous in the homozygous condition. The carriers of such a gene would leave more descendants; as a result, the gene would increase in frequency in a population from generation to generation, until it finally would become, in its turn, the wild-type gene. Natural selection thus produced, as it were, something close to the best of all possible worlds, and changes would be rare and slow and not apparent over one life-span, in agreement with the observed constancy of species over historical time.

Drosophila studies

Dobzhansky's most important contribution was to change this view. In observing wild (Californian) populations of the vinegar fly *Drosophila pseudoobscura*, Dobzhansky found extensive genetic variability. Furthermore, around 1940 evidence accumulated that in a given local population some genes would regularly change in frequency with the seasons of the year. For example, a certain gene might appear in 40 percent of all individuals in the population in the spring, increase to 60 percent by late summer at the expense of other genes at the same locus, and return to 40 percent in overwintering flies. Compared to a generation time of about one month, these changes were rapid and effected very large differences in reproductive fitness of the various types under different climatic conditions. (Fitness of a type in this sense is proportional to the average number of its offspring that survive to reproduce.) Furthermore, the course of many of the changes could be observed in laboratory populations, which could be maintained for years in small cages with a continuing supply of new food. Other experiments showed that, in fact, flies of mixed genetic makeup (heterozygotes) were superior in survival and fertility to either pure type—that is, homozygous dominant and homozygous recessive.

It was already known that these superiorities of such heterozygotes would ensure the preservation of both sets of genes in the population. Dobzhansky pointed out that newly arisen genes are rare at first and that an individual is exceedingly unlikely to receive such a gene from both parents. Hence, in the beginning, the only genes that can "get ahead" and become more widespread in the population are those that are "good mixers"—that is, those that produce superior genotypes when combined with a random gene from the population, and not necessarily when homozygous.

The picture thus developed by Dobzhansky and some of his students showed that most gene loci are represented by more than one gene in a given population, no two individuals having exactly the same genotype. Hence the so-called wild-type genotype was a myth. A genetic system of the kind proposed by Dobzhansky can change rapidly, in response to natural selection, if environmental conditions should change. Among the myriad genotypes appearing in each generation would be many that were adapted to the changed conditions and that would leave more descendants; thus, these genes would be more common in the next generation. In contrast, under the older idea of a fairly uniform population in which most gene variants occurred rarely, much more time would be needed before variants adapted to new conditions could arise and become common. Meanwhile, local populations of the species could be in danger of becoming very reduced in numbers or even extinct. Although objections have been raised to some of Dobzhansky's ideas, more and more experimental and mathematical evidence seems to support them.

Other important work of Dobzhansky dealt with speciation: the process by which a species does not merely change its characteristics over time but actually splits into two or more species, which are no longer capable of interbreeding with the parent stock or one another and, consequently, go their separate ways. When speciation occurs, then a period of time must exist during which the process is partly completed, so that there are a number of races of a species that are not yet completely separate. He studied such a species, *Drosophila paulistorum*, in parts of South and Central America, that clearly demonstrates this process.

Speciation

In extension of his work in human genetics and in human paleontology, Dobzhansky wrote on the "descent of man." His *Mankind Evolving* (1962) has proved to be almost as influential among anthropologists and other students interested in human evolution as his *Genetics and the Origin of Species* was for those primarily interested in the evolution of nonhuman organisms. Finally, his interest in the direction that human evolution might take in the future, added to a natural philosophical inclination, led him into thought on the nature of man and the purpose of life and death, as evidenced in his works *The Biological Basis of Human Freedom* (1956) and *The Biology of Ultimate Concern* (1967). *Genetics of the Evolutionary Process* (1970) reflects 33 years of scientific progress in the study of evolution, largely by Dobzhansky or under his influence.

Although pre-eminently a laboratory biologist and writer, Dobzhansky never lost his liking for fieldwork; he boasted of having collected from Alaska to Tierra del Fuego and in every continent except Antarctica. An inspiring teacher and lecturer, he received over the years a steady stream of scientists from other countries, who came to spend time in his laboratory to learn his approach to research.

Beginning in 1918, Dobzhansky published well over 400 research papers that provide an important part of the factual evidence for modern evolutionary theory. His pre-eminence, however, lay even more in the rare talent for synthesizing the masses of experimental and theoretical data in the literature into a broad, comprehensive view of the subject. Dobzhansky died at Davis on December 18, 1975.

BIBLIOGRAPHY. HOWARD LEVENE, LEE EHRMAN, and ROLIN RICHMOND, "Theodosius Dobzhansky Up to Now," in *Essays in Evolution and Genetics in Honor of Theodosius Dobzhansky* (1970), includes a detailed description of his scientific work and his bibliography through 1969. See also THEODOSIUS DOBZHANSKY and L.C. DUNN, *Heredity, Race, and Society*, 3rd ed. (1956), an elementary exposition of genetics with respect to man.

(H.Le.)

Doctrine and Dogma

The development of doctrines and dogmas—*i.e.*, the explications and officially acceptable versions of religious teachings—has significantly affected the traditions, institutions, and practices of the religions of the world. Doctrines and dogmas also have influenced and been influenced by the ongoing development of secular history, science, and philosophy.

Nature and significance. Doctrine in theology (Latin *doctrina*, Greek *didaskalia*, *didachē*) is a generic term for the theoretical component of religious experience. It signifies the process of conceptualizing the primal—often experiential or intuitive—insights of the faith of a religious community in support of rationally understood belief. Doctrines seek to provide religion with intellectual systems for guidance in the processes of instruction, discipline, propaganda, and controversy. Dogma (Latin *decretum*, Greek *dogma*) has come to have a more specific reference to the distillate of doctrines: those first (basic or axiomatic) principles at the heart of doctrinal reflection, professed as essential by all the faithful.

This distinction appears in Christianity in the New Testament, in which *didaskalia* means "basic teachings" (as in I and II Tim.) whereas *dogma* is used only in the sense of an official judgment or decree (as in Acts 16:4). Later, however, many theologians of the early church (including, for example, Origen, St. Cyril of Jerusalem, and St. Jerome) use the term dogma in the sense of doctrine. In Eastern Christianity, the theologian St. John of Damascus popularized the term "orthodoxy" (literally "correct views") to connote the sum of Christian truth. In Western Christianity, the great medieval theologian Thomas Aquinas chose the phrase "articles of faith" to denote those doctrines solemnly defined by the church and obligatory for faith. As late as the Roman Catholic reformatory Council of Trent (1545–63), "doctrine" and "dogma" were still roughly synonymous.

Most modern historians, however, have stressed their difference. According to J.K.L. Gieseler, a 19th-century German church historian, in *Dogmengeschichte*,

> Dogma is not doctrinal opinion, not the pronouncement of any given teacher, but doctrinal statute (*decretum*). The dogmas of a church are those doctrines which it declares to be the most essential contents of Christianity.

A modern church historian, Adolf von Harnack, sought to explain the rise of dogma in Christianity as the specific consequence of an alien blend of Greek metaphysics and Christian thought that had been rendered obsolete by Protestantism's appeal to Scripture and history. The German Roman Catholic dogmatician Karl Rahner's contrasting definition, in *Sacramentum Mundi*, points to a perennial process:

> Dogma is a *form* of the abiding vitality of the deposit of faith in the church which itself remains always the same.

Functions of doctrines and dogmas. The functions of doctrines and dogmas vary in the several religious traditions according to the stress each puts on the importance of the rational conceptualization of religious truth first glimpsed in images, symbols, and parables. In what are viewed by some scholars as the more mystical religions of the East, doctrines are usually designed to serve as catalytic clues to religious insight (*e.g.*, the notions of Nirvāṇa, or the goal of the religious life, in Hinduism, Jainism, and Buddhism). In what are regarded as the more personalistic religions of the West, doctrines and dogmas tend to function as aids to theological reflection (*e.g.*, the concept of God's unity in Judaism, Christianity, and Islām). In all the higher religions, doctrines and dogmas emerge and develop in the service of instruction for the faithful: interpreting their sacred Scriptures, understanding their obligations and duties, and safeguarding the lines between allowable diversity and actual error—all of which help to chart the religious pathway to wisdom, rectitude, and fulfillment. Theology (which utilizes doctrines and dogmas) is, according to the medieval Christian theologian and churchman St. Anselm of Canterbury, "faith seeking rational self-understanding."

The rational conceptualization of religious truth

The normative function of doctrinal formulation is a typically vain effort to fix and conserve an interpretation of the original dogmas of a given tradition. The themes of *saṃsāra* (the process of reincarnation) and *karman* (the law of cause and effect) are shared by Hinduism, Jainism, and Buddhism, though with quite different doctrinal explicitations and consequences. Analogous developments are evident in other traditions.

A third function of doctrine is polemical: the defense of the faith against misinterpretation and error, within or without a religious tradition. Given the invariably pluralistic character of theological reflection, there is a constant tension between the concern for identity and continuity of the tradition, on the one hand, and for deeper and richer comprehension of truth itself, on the other. Over against this there is in most cultures a concurrent rivalry with other religions, with their contrary doctrinal claims, and beyond that, the challenges of secular wisdom and unbelief. This calls forth a special sort of doctrinal formulation: apologetics, the vindication of the true faith against its detractors or disbelievers.

The problem of primal authority

At the heart of all efforts to support religious faith lies the problem of primal authority. It is required of a doctrinal statement that it be clear and cogent, but doctrines always point past their logical surface to some primitive revelation or deposit of faith. The appeal may be to any one of a number of primary authoritative positions: to the memory of a founder (as in Zoroastrianism), or a prophet (Moses in Judaism), or to ancient Scriptures (*e.g.*, the Veda and *Upaniṣads* in Hinduism), or an exemplary event (as in Gautama, the Buddha's "enlightenment"), or to God's self-disclosure (as in the Torah, or Law, for Judaism, or in Jesus Christ in Christianity, or Muḥammad's revelations to Islām). Here again, the diversity between doctrines ("allowable interpretations") and the stability of dogmas ("essential teaching") points to the vexed problem of doctrinal development in history that is apparent in all the traditions.

The development of doctrines and dogmas. Every religion has a history of doctrine that is more than a replication of the deposit of faith. Doctrine, as a mode of pedagogy, is conservative of its tradition; as a mode of inquiry, it may be innovative, generating new insights that alter the rhetoric of conventional teaching and, sometimes, its substance as well. There are, of course, wide variations. The persistent continuities between ancient Zoroastrianism and its modern form, Parsiism, or in Jainism, are clearer than those between primitive Hinduism and modern Vedānta (a Hindu philosophical system). All forms and sects of Buddhism appeal jointly to the Three Jewels (the Buddha; the *dharma*, or law; and the *saṅgha*, or monastic order) but are irreconcilable in their differences of interpretation and practice. In each case, the question as to what constitutes legitimate development (*e.g.*, the rival claims of Theravāda, or "Way of the Elders," and Mahāyāna, or "Greater Vehicle," in Buddhism) is left undetermined.

All Jews profess devotion to Torah, even in their disagreements over its authentic observance. Christians profess a common loyalty to the Bible and a common acceptance of the twin dogmas of the Trinity (that the one God is three Persons—Father, Son, and Holy Spirit) and the God-Manhood of Jesus (that Christ is both divine and human) but then divide in their doctrinal systems as they have developed historically. Later dogmas (*e.g.*, transubstantiation, the teaching that the substance of the bread and wine in the Lord's Supper is changed into the substance of the body and blood of Christ, with the properties of the bread and wine remaining unchanged) were defined by the Latin Church without concurrence from Eastern Orthodoxy; the modern dogmas of the Roman Catholic Church (*i.e.*, the immaculate conception of the Virgin Mary, the bodily assumption into heaven of the Virgin Mary, and papal infallibility) were defined in separation from both the Eastern and Protestant consensus. Protestantism has continued an emphasis on its distinctive dogmas of "grace alone" (*sola gratia*), "faith alone" (*sola fide*), and "Scripture alone" (*sola Scriptura*) but has nevertheless undergone immense change and proliferation.

Islām lays great stress on doctrinal stability focussed in the Qur'ān, the *sunnah* (custom or tradition), and the consensus (*ijmāʿ*) of its jurists (*ʿulamāʾ*). Even so, it has produced doctrinal variants—especially in the medieval period—as disparate as the mysticism of the Iranian-born philosopher al-Ghazālī and the rationalism of the Spanish philosopher Averroës and the Persian philosopher Avicenna.

Views of doctrinal development

The process of doctrinal development has been explained variously as a process of logical unfolding or of organic growth, or else as a process of purgations of error and restorations of the original deposit. The notion of a logical unfolding assumes that all that has developed in a religious tradition over the course of its history was already implicit in its original foundation and subsequently had only to become more fully understood. In the case of the doctrine of the Trinity in Christianity, for example, it is argued that the abundant references in the New Testament and the earliest liturgies to God as Father, Son, and Holy Spirit required the development of the dogma as the explicitation of essential Christian conviction. Similarly, the dogma on the nature of Christ is understood as the logical outcome of sustained reflection on the testimony about Jesus as the Christ in the Bible and in the apostolic tradition. In the notion of logical unfolding, even in its development, truth remains forever unchanged.

Theories of organic development stress the fact that the history of doctrine includes more than explicit formulation of implicit revelation. Such theories take into account the ways in which religious thought is affected by "contemporary" science, philosophy, and historical crises (*e.g.*, the "Copernican revolution" in astronomy, the Renaissance, and other such events). The holders of this view are convinced, however, that all such historical supplementations have been integrated into the original deposit and thus exhibit the power of the religious organization (*e.g.*, the church) to grow and change without substantial alteration of its identity. Thus the 19th-century Roman Catholic cardinal J.H. Newman, in his *Essay on the Development of Christian Doctrine* (1845), argued that

> . . . the highest and most wonderful truths, though communicated to the world once for all by inspired teachers could not be comprehended all at once by the recipients, but, . . . have required only the longer time and deeper thought for their full elucidation (*Introduction*, pp. 29–30).

Newman also believed that this process was safeguarded by the authority of the teaching that would even allow for revisions and occasional corrections of antecedent.

Protestants, by and large, have been more impressed by the lapses and deviations they see in church history and doctrine and thus have tended to construe authentic "development" in terms of a perennial recourse to Scripture and apostolic tradition. Such a view takes historical flux for granted and is less sensitive to the problem of historical continuity.

In all traditions, the course of doctrinal development is crucially affected by the occasional emergence of profound and powerful thinkers who have gathered up scattered elements in their tradition in freshly relevant syntheses, altering thereby the subsequent history of that tradition. This can be seen, for example, in the North African theologian Augustine's contributions to the making of Latin Christianity and in the matching services of John of Damascus in Eastern Orthodoxy. Such also was the role and contribution of Moses Maimonides in medieval Judaism (*e.g.*, the Thirteen Articles of Faith in his *Mishne Torah*) and of Thomas Aquinas in medieval Christianity (*e.g.*, *Summa theologiae*). The 16th-century Reformers Martin Luther and John Calvin gave Protestantism its classical form, to be followed by yet other and different system builders (*e.g.*, Friedrich Schleiermacher in the 19th century and Karl Barth in the 20th century).

Each theory of development has had its own distinctive prescription for doctrinal stability and doctrinal change. In Christianity, Eastern Orthodoxy locates its authority in "Holy Tradition," fixed and guided by the dogmas of the ecumenical councils. Roman Catholicism relies on the magisterium (teaching authority) of the church, directed by the bishops as a "college" (*collegium episcoporum*) and supremely by the bishop of Rome as their collegial head. Protestantism has sought to bind both tradition and the church to the authority of Holy Scripture, with the resulting problem of specifying what is to be regarded as truly authoritative interpretations of Scripture.

The relation of faith, reason, and religious insight to doctrine and dogma. Insofar as doctrines and dogmas represent conceptualizations of the human encounter with the divine mystery, they are bound to reflect the interplay of faith and reason in religious experience and to imply some notion of levels and stages in the progress of believers from the threshold of faith toward its fulfillment. Doctrine is concerned with communication and consensus, with the exposure of the religious vision to rational probes and queries. There is, therefore, a tension in all religions between mystical intuition and logical articulation, between insight and dialogue. Most traditions agree that perfect understanding is a goal that lies beyond a "simple faith" and the routine observance of rites and duties. Most of them also agree that the utmost pinnacle of religious insight is ineffable. One mode of differentiation between doctrinal traditions, therefore, is their relative openness or resistance to the auxiliary services of philosophy and science of faith's fulfillment.

The interplay of faith and reason in religious experience

In the religions of the East, very broadly, reason's chief role is the purgation of illusion and self-deception so that souls may follow the ways of wisdom and right conduct to their true fulfillment in Nirvāṇa. The Hindu passes from the initiatory level of "the student" (dependent on a teacher, guru) to the ambivalent freedom of "the householder," to the great freedom of "the forest dweller," to the fullest freedom of the *sannyāsin* (enlightened ascetic). Reason, chiefly reflective, assists at every stage in perfecting faith's self-understanding. In Buddhism, one follows the *dharma* from *saddha* (practical knowledge of one's religious obligations), to *jñāna* (rational insight), to *vijñāna* (mystical illumination).

In the religions of the West, again very broadly, the primary function of reason has been that of rendering the mysteries of faith as intelligible as possible, in support of the intellectual love of God. In Judaism, progress in the knowledge of Torah is focussed in the Bible and the Talmud (commentaries on the Law), guided by the twin hermeneutical (critical interpretive) principles of Halakha (the oral precepts and decisions of the rabbis) and Haggada (instructive stories, parables, and other such devices).

Variations in Islām range from the rigid orthodoxy of the Ḥanbalites (a conservative school of law following the teachings of Ibn Ḥanbal), to the rational liberalism of the Muʿtazilites (a school of law utilizing Greek philosophical methods), to the dialectical doctrines (*kalām*) of the Arabian theologian al-Ashʿarī and the Turkish philosopher al-Fārābī. All of these, however, are anchored in the twin dogmas of the unity of God and the prophetic office of Muḥammad. Spiritual progress is measured by the believer's faithfulness in obedience to the "Five Pillars," or religious duties, including prayer, fasting, and the pilgrimage to Mecca.

In Christianity, the dialectic between faith and reason has ranged from the fideism (emphasis on faith) of the 2nd-century North African theologian Tertullian to the intellectualism of Thomas Aquinas. An ancient distinction between faith as bare assent to orthodox doctrine (*fides informis*) and faith as existential trust in God's grace (*fides formata*) gave rise to the further distinction between faith as a set of doctrines to be believed (*fides quae creditur*) and faith as personal involvement (*fides qua creditur*). Philipp Melanchthon, a 16th-century Lutheran Reformer, stressed the point that even the devils are "orthodox" (having "dead faith") but to no avail, since only those who have embraced God's reconciling love (*fiducia*) receive the benefits of salvation ("living faith"). In general, this distinction has become standard in Protestantism.

Faith as assent and as trust

Conclusion. In all the great religious traditions, and between them, the clash of doctrines and dogmas has,

more often than not, been polemical. The *odium theologorum* ("bitterness of the theologians") of which Melanchthon once complained so plaintively has been notorious. Within the several traditions, doctrinal disputes have sometimes led to division or else have accompanied divisions caused otherwise. In relationships between the great world religions, dogmas and doctrines have usually been regarded as mutually exclusive. There are, however, significant signs of change in this attitude. The rise and spread of the ecumenical movement in the 20th century and notable advances in the comparative study of world religions reflect an enlarged commitment to the widest possible community of mutual religious interests. The "Decree on Ecumenism" and its "Declaration on the Relationship of the Church to Non-Christian Religions" of the Roman Catholic second Vatican Council (1962–65) are signal instances of this new disposition.

BIBLIOGRAPHY. HEINRICH EMIL BRUNNER, *Dogmatik*, vol. 1 (1946; Eng. trans., *Dogmatics*, vol. 1, *The Christian Doctrine of God*, 1949), advocates the primacy of Scripture over tradition; OWEN CHADWICK, *From Bossuet to Newman: The Idea of Doctrinal Development* (1957), an excellent survey of the gradual shift from "the classical consciousness" of identity in doctrine (Bossuet) to a "historical consciousness" of growth and continuity (Newman); ADOLF VON HARNACK, *Lehrbuch der Dogmengeschichte*, 3rd ed., 3 vol. (1893; Eng. trans., *History of Dogma*, 7 vol., 1900, reprinted 1961), a massive exposition of the thesis that Christian dogma represents the process of Hellenization of the original Gospel, hence a deviation; JOHN HENRY NEWMAN, *An Essay on the Development of Christian Doctrine*, new ed. (1878), a classical statement of the emergence of the historical consciousness within the Catholic tradition; JOHN B. NOSS, *Man's Religions*, 4th ed. (1969), a general survey; J. PELIKAN, *Development of Christian Doctrine: Some Historical Prolegomena* (1969), an important current statement of the interaction of Scripture and tradition in the formation of Christian doctrines and dogmas; KARL RAHNER, "Dogma," in *Sacramentum Mundi*, vol. 1, pp. 909–917 (1968), reflects the new perspectives of Vatican II; FREDERICK J. STRENG, *Understanding Religous Man* (1969), an excellent summary of the common elements in religious experience, including those relating to doctrine and dogma; MARTIN WERNER, *Die Entstehung des christlichen Dogmas* (1941; abr. Eng. trans., *The Formation of Christian Dogma*, 1957), argues that Christian dogma displaced the original eschatological message of the early church; R.C. ZAEHNER, *Concordant Discord: The Interdependence of Faiths* (1970), helpful insights as to the various ideas of authority in the major religions of the world.

(A.C.O.)

Dog

Dogs, along with wolves, jackals, and foxes, are members of the family Canidae. As such, they share certain features that suit them admirably for a life of active hunting; powerful jaws with teeth adapted to seizing, keen senses of smell and hearing, and a social instinct that maintains and coordinates the efforts of the pack. The family of dogs and their relatives belongs to the order Carnivora, the "flesh eaters"; for an account of the family and its relationships, see CARNIVORA. Regardless of superficial differences, all dogs belong to a single species, *Canis familiaris*, and its more than 100 breeds depend entirely on selective breeding and maintenance by man for their continued existence. The dog is an extremely social animal, whose well-being and normal psychological development are products of association with other dogs in a pack. Unlike man's other favourite domesticate, the cat, a dog adjusts with difficulty, if at all, to an independent and wild existence and draws heavily on the mutual exchange of pack members and the guidance of the pack leader, or a human master, who is in fact a surrogate pack leader.

ORIGIN AND HISTORY

The dog is the oldest domesticated animal, having been in the company of man for at least 10,000 years and having originated probably somewhere in Eurasia 12,000 to 14,000 years ago. It belongs to the same genus (*Canis*) as do the coyote, jackal, and wolf, all of which have been considered to be his ancestors. Much farther back in time, about 40,000,000 years ago, lived a tree-climbing carnivorous mammal called *Miacis*, from which the lineage of the canines is traced through *Cynodictis*, from whom the African Cape hunting dog, the Indian wild dog (or dhole), and the South American bush dog ultimately sprang, through *Cynodesmus*, which gave rise to the hyenas, and finally through *Tomarctus*, the progenitor of the fox, wolf, jackal, and dog. The coyote, which lives only in North America, can be eliminated as an ancestor of the early European dog.

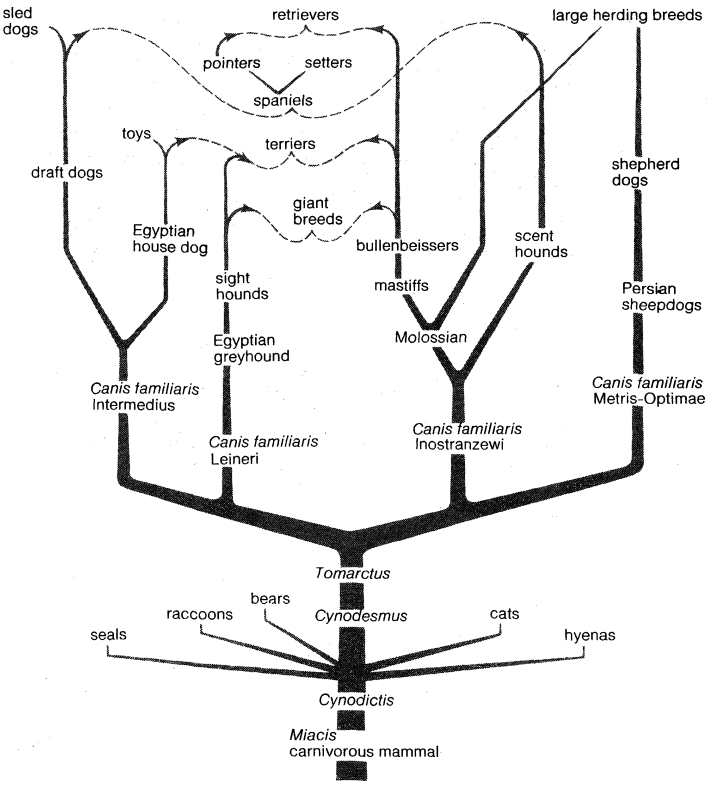


Figure 1: Possible genealogy of the dog.

Probable wolf ancestry

The most likely candidate is the gray wolf, which was originally found over all of Europe, Asia, and North America, with a great many highly variable subspecies and local varieties. The wolf is so variable in colour that even individual members of a pack can be easily distinguished. Northern wolves are much larger than southern ones, which makes it probable that dogs came from a smaller central or southern variety. Such a race of small wolves lives in India, and another was formerly found in China. The other possible candidate, the jackal, is primarily an African animal, although its range extends into Mesopotamia, southeastern Europe, and India. The jackal is less social than the dog; it howls in a manner unlike any domestic dog and has a narrow, foxlike head—factors that make it unlikely that it is the dog's ancestor.

Domestication. The wild dog, like the wolf, roams in packs in search of food, and it would seem not unlikely that some primitive dogs found food more readily available near man's encampments, where animal hides and picked-over carcasses were left to rot. The dog apparently was less wary than its canine relatives and may have ventured more boldly into man's settlements to feed. Its association with man developed gradually; the more docile and tractable dogs were tolerated; the others were run off or killed.

Man and dog as mutually dependent

Man may have grown to depend on the dog to warn him of approaching threat, and the dog in turn grew to depend on man for food and shelter. Mutual bonds of benefit and affection probably strengthened gradually over the centuries until by selection for traits and appearance the domesticated dog became man's creation. In time, further attention to breeding resulted in modifications that have led to widely variant breeds.

After dogs became domesticated, certain traits became established that distinguish dogs from their wild relatives. One such trait is the upturned tail, ranging from a sickle shape to a tight curl. This feature probably goes back to the original stock of domestic dog, pointing to a common ancestry for all types of dog. Another characteristic that distinguishes the dog from the wolf is its smaller, less powerful dentition. There must have been an early artificial selection for those animals that were smaller, less toothy, and also more easily tamed and controlled.

Distribution

The domesticated dog apparently spread very rapidly all over the world, through both hemispheres and from tropical to Arctic climates. When the Europeans arrived in North America they brought their own dogs with them, but every Indian tribe already had them. At that time there were at least 20 distinct breeds in North and South America. Most of these have disappeared except for the Mexican hairless and the Eskimo dogs.

In Australia there is the dingo, a species separate from the domestic dog. Typically a wild animal, the dingo is sometimes found semidomesticated in the camps of the Aborigines. Its ancestors must have been brought as domestic dogs to Australia by the first immigrants several thousand years ago and later allowed to run wild.

Since the beginning of history dogs have been found all over Africa. One of the surviving native breeds is the African basenji, still used by the pygmy tribes in equatorial regions. The basenji is probably descended from an early breed adapted to tropical living, which spread through southern Asia and the East Indies and eventually to Australia, where it became the dingo. Other varieties of dogs lived throughout Asia and on most of the oceanic islands.

Most is known about the history of European dogs. From the earliest times traders and travellers not only took their favourite dogs on long journeys but often returned with new and exotic varieties. Dogs were nowhere more cultivated than in England.

By the time John Caius (the founder of Caius College at Cambridge) wrote a description of English dogs for Konrad von Gesner, a 16th-century naturalist, the English had collected at least six main varieties of dogs—greyhounds, true hounds, bird dogs, terriers, mastiffs, and shepherd dogs. A basic group of dogs not mentioned by Caius includes the sled dogs of the Eskimo, found in the Arctic in both America and Eurasia. These large curly-tailed dogs, reputedly crossed with wolves, and similar smaller dogs of northern Eurasia are sometimes called the polar, or spitz, group.

Although the same general types of dogs were found all over the world many less distinct in their physical and behavioral traits have since disappeared. Many ancient breeds have greatly changed or entirely disappeared, but in Iraq two ancient types of dog, the saluki and the Kurdish herding dog, are still found. The latter of these is a large breed somewhat resembling the mastiffs and war dogs pictured in Babylonian art in 2200 BC.

Historical records indicate that dog breeds have frequently been crossed with each other, so that it is difficult in many cases to determine the ancestors. While conscious human selection has doubtless played a part in producing genetic changes from the earliest times, the dog is also an evolving species. Just as an animal that enters a new physical habitat undergoes rapid change and differentiation, the dog on domestication entered a new biological habitat in association with the human species and underwent a similar rapid evolution assisted by man.

Associations with man. The dog figures prominently in many tales of courage and selfless devotion in the service of man, of steadfastness and perseverance, of attentiveness and seeming concern for his master. The romantic stories of Albert Payson Terhune are filled with the heroic deeds of the dog, as are Jack London's *White Fang* (1906) and other novels. The dog has been bred for many special tasks—hunting, guarding, herding, drafting, guiding (for the blind)—but most popularly for companionship and as a household pet.

Because the close social relationship between dogs and man appears to many to be similar to the human parent–child relationship, dogs have been used to test various theories of child training. Dogs also have important uses in medical research, and the resulting discoveries have helped to improve both animal and human health.

GENERAL FEATURES AND SPECIAL ADAPTATIONS

The wolf-like heritage

The dog, in many of its breeds, is basically a wolf-like hunter, with physical features usually identified with such a mode of life: musculature fitting a coursing or running animal, teeth suitable for seizing and holding prey, internal adaptations to a carnivorous habit (short gut and other digestive tract features), and keen senses of smell and hearing.

Coordination and musculature. While the dog is no rival for the cat in fluid movement and balance, it has retained some of the wolflike aspects of deliberate tracking and cautious cunning. It moves the way a horse does when walking but prefers the trot, in which the right foreleg and left hindleg advance together, followed by the other legs; at top speed it breaks into a gallop. The dog is not as agile and foot sure as the cat, nor as flexible in body construction, but its speed and perseverance are greater than the cat's. The wolflike musculature is still found, though greatly modified in mass and development, in the different breeds. The sheetlike musculature covering the head and main body mass is particularly noticeable by its action in raising the hackles, or back hair, baring the teeth, and cocking the ears. Muscles of the hind back work the tail, which is generally wagged briskly and held high in contentment and waved slowly and in a horizontal position in attentive approach or before attack. The tail aids in balance, as when the dog dashes around a curve.

Teeth. A dog's first, temporary, set of teeth, or milk teeth, is replaced by a permanent set at about five months. The 42 permanent teeth include incisors, which are used to nip and bite; canines, used to tear and shred flesh; and premolars and molars, which shear and crush. The canines are upper and lower fangs for which the dog family was named. As in most carnivores, the teeth are high crowned and pointed, unlike the broad, grinding teeth of many herbivorous animals. Having less manipulative ability than the cat, the dog uses its teeth to catch and hold items, such as food or a toy, as well as to prepare food for digestion.

Senses. *Smell.* The two senses most supremely developed in the dog are hearing and, pre-eminently, smell. The dog's world is primarily one of scents, as man's is one of sights. The dog's nasal passages are so arranged that a greater volume of air can be drawn over the sensitive lining than is the case with man. Rapid sniffs carry messages to the enlarged olfactory centre in the brain where the scents are analyzed and cataloged. Practical scents seem to be of greatest interest—the smells that identify other animals: blood, sweat, excrement, urine, and the smells associated with the sex organs. Grassy smells, soil, and ripe, putrefactive odours are likewise attractive.

Hearing. Hearing is an acute sense in the dog; frequencies up to 35,000 vibrations per second can be detected, compared with 25,000 in the cat and 20,000 in man.

Vision. Sight is a relatively poorly developed faculty in the dog. Moving forms are readily discerned but a still object with no particular odour may go unnoticed. The dog is colour-blind (even guide dogs, which respond to traffic signals, do so by the position of the lighted signal rather than by the colour).

BEHAVIOUR

The wolf–dog pattern. Many peculiarities of dog behaviour can be understood only in reference to the lives of their wild ancestors, the wolves. Within recent years the behaviour of wolves in undisturbed wild conditions has been studied scientifically, reducing much of the mythology surrounding them to scientific fact.

Territory and range. Dogs recognize a central headquarters, or den, which, with the immediate area sur-

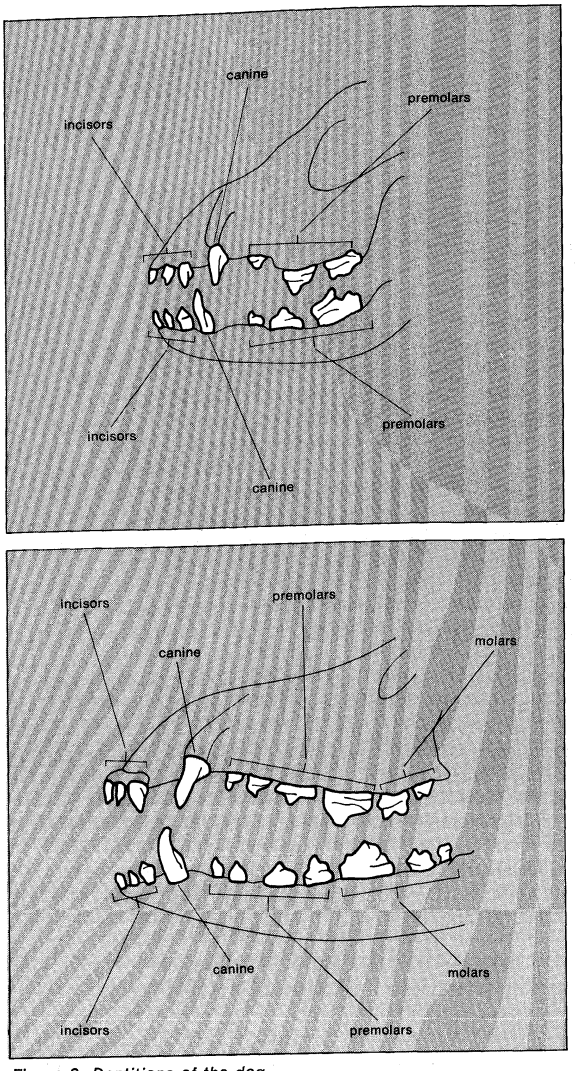


Figure 2: *Dentitions of the dog.*
(Top) Temporary (milk) teeth of a puppy. (Bottom) Adult (permanent) teeth.

"Scent posts"

rounding, is defended as a territory from which other dogs are excluded. At various points beyond this central territory dogs will establish "scent posts," where they stop to urinate and defecate. Males lift the leg to urinate and scratch the ground after defecation. Strange dogs passing by will also mark the place in the same manner. These wolflike habits are seen especially in dogs allowed to run loose in town or country: the tendency to defend the immediate areas around the houses of their owners, but to wander much more widely, making certain "scent posts" as they go.

Group activities. Wolves are highly social animals. Within the pack, wolves are peaceable and highly cooperative. While feeding, they observe a dominance order, in which the most dominant animal feeds first. Parts of prey animals are often buried, either at the site of the kill or in the vicinity of the den, presumably to save the food from scavengers. Domestic dogs show the same habits, even attempting to bury such things as dry dog food.

Wolves show the same general types of sounds and communication as dogs—barking when a strange animal approaches the den, yelping in fear or distress, growling when threatening another animal, and howling, either alone or, when the pack is together, in unison.

Dominance and subordination behaviour

Combative behaviour. Two strange male dogs usually approach each other stiffly, with tails held erect, and appear to identify each other by sniffing in the tail region. If a fight starts, they rush at each other, snapping and snarling, and the fight will last until one runs away or submits. The beaten animal will roll on its back, extending its paws, yelping, and protecting its throat by snapping. The winning animal stands over it, growling and threatening. If dominance has already been established, one dog may indicate this by placing his paws on the other's back and growling while the subordinate animal keeps his tail low. Similar behaviour is seen in wolves.

Courtship and care of young. In courtship dogs show a characteristic pattern of play. They crouch extending their forepaws and cocking their heads to one side, then they throw their forelegs around each other's necks and wrestle. This is followed by running and chasing and eventually mating. Like wolves, dogs show the peculiarity of the sexual tie; after copulation the two animals remain locked together for many minutes.

At first, young puppies are fed exclusively by nursing. After three weeks or so, they begin to take solid food, with final weaning at about seven weeks of age. Wolves feed the young by vomiting. Domestic dogs show this pattern more rarely but vomit readily and, like wolves, frequently eat the vomitus.

Special traits. *Intelligence.* Judging how intelligent a dog is, either in comparison to other dogs or to other animals, is difficult. Obviously, a dog will be unable to perform as well as a monkey any problem involving complex manipulation or colour vision. Intelligence is measured primarily by the ease with which a dog can be trained. This may depend in a large measure, however, on the degree of motivation. In general, when individual dogs are sufficiently motivated there appear to be no wide differences in intelligence between breeds.

The smartest breeds

There are big differences, however, in the ease with which breeds and individual dogs will accept complex training. Certain breeds, particularly the shepherd dogs and poodles, have reputations for intelligence of this sort. These animals have a high capacity for developing motivation and attention toward their handlers and a similarly high capacity for accepting inhibitory training. Most of the "wonder dogs" famous for their intelligence belong to these groups. A highly trained dog is capable of mastering many different commands.

At the same time there are definite limitations on canine intelligence. Like other mammals (excluding man), dogs cannot be taught to talk and indeed learn to bark on command only with great difficulty. There is no evidence that dogs are able to recognize the meaning of words when they are used in new combinations (*i.e.*, to understand new sentences); thus the amount of information that actually can be conveyed to a dog is quite limited. A well-trained dog, however, is capable of such attentiveness to his owner's slightest movement and mannerisms that he almost seems to read the owner's mind.

Emotions. Most of a dog's emotional reactions are readily understandable. Growls and snarls are threats, and barking is usually an alarm signal. Rapid horizontal tail wagging indicates a friendly approach and is roughly the social equivalent of the human smile. High-pitched yelping occurs in situations involving pain or terror. The rapid yelps and whines of puppies indicate distress, which may have many causes. Solitary howling usually indicates loneliness or may be a reply to another howling dog.

Some other reactions are less easy to interpret and can be understood only in terms of the behaviour of dogs toward each other. Jumping up and offering the forefeet is usually an attempt to initiate playful fighting so often seen between dogs. A stiff-legged approach with erect and slowly wagging tail indicates aggressiveness and may be followed by an attack. A "worried" dog holds his ears down, so that his forehead is smooth. An attentive dog often displays a wrinkled forehead, usually the result of the erection of the ears.

The external expression, however, is not always a true guide to internal emotional state. A serious prolonged internal emotional disturbance will often show up in depressed activity and loss of appetite.

Abnormal behaviour; neuroses. It is difficult to interpret any behaviour of a dog as abnormal, because the animal will adapt to a situation in any way that gives him satisfaction, however bizarre this may appear to the on-

looker. Unusual behaviour may also result from organic disease. Any marked changes in behaviour, such as convulsions (often called fits), staggering gait, and the like, are usually indications of serious disease.

A neurosis may be defined as a kind of behaviour that gives relief from tension without adaptation to the cause of tension. Dogs confined to small pens for long periods, for example, will often develop the habit of running in circles or jumping from side to side whenever a person approaches. The dog is stimulated to respond to the person but because of confinement can only make what appear to be useless and inappropriate movements. Similarly, a dog left by itself in a house for long periods will often chew furniture and rugs. This behaviour can be understood as a thwarted attempt to escape.

The effect of complete isolation

Psychotic behaviour, defined as a general disruption of adaptation to social situations, is seldom seen in dogs. A puppy reared in complete isolation during the period of socialization, however, will avoid all human and canine contacts, showing more severe symptoms than the kennel-dog syndrome that arises from being reared in a limited environment. When dogs closely attached to their masters are separated from them for long periods, they may show symptoms of depression very similar to those exhibited by human beings.

DOGS AS PETS

The ideal time to acquire an individual puppy is between six and eight weeks of age, in order to permit normal psychological development. Since neither physical nor behavioral development is complete at this time, the best guide to the puppy's future behaviour is that of its parents.

As adults, males are usually larger than females, more active, and tend to wander more.

A mongrel is a dog of unknown ancestry. First-generation hybrids between two pure breeds show a great deal of hybrid vigour and are usually more healthy and hardy than either parent breed. Their appearance and behaviour traits may be like one or the other parent, or neither. Some dog breeders regularly produce such hybrids for special purposes when the result can be predicted.

Developmental periods. The early development of a dog is divided into several distinct periods: (1) neonatal, (2) transition, (3) socialization, and (4) juvenile. In the neonatal (newborn) period, the puppy's activity is largely confined to nursing and sleeping.

The newborn puppy is blind and deaf and is consequently largely isolated from the external world. In about 14 days its eyes will open, marking the beginning of the transition period. During the third week the puppy undergoes a rapid change in behaviour and in sensory and motor abilities. Toward the end of the week, when its ears open, the puppy begins to react to sound. At the same time its first teeth appear, and it will attempt to eat solid food if it is offered.

Socialization

Meanwhile it has begun to walk instead of crawl and to show social responses to humans and to other dogs. This marks the beginning of the period of socialization, at approximately three weeks after birth. The puppy will now slowly approach a strange person, nosing and wagging its tail. Another social response is playful fighting with its littermates. At any time during this period, which lasts up to 12 weeks of age, it is easy to form a close social relationship between a puppy and its owner, the maximum favourable response being obtained between six and eight weeks of age.

By removing a puppy from its litter early in the socialization period all social relationships are transferred to human beings. Emotional disturbance and prolonged yelping is a normal reaction to removal from the litter and can be relieved by fondling and companionship. The puppy soon associates its owner with relief from distress. Adopting a puppy late in the period will throw the balance in the opposite direction, so that the puppy's strongest relationships are with dogs rather than with human beings. Puppies raised in large fields apart from human beings will become almost completely wild by 14 weeks of age. This period is therefore a critical one for determining the nature of social relationships and adjustment to the environment. Puppies left in a kennel environment much beyond 12 weeks of age are likely to be permanently shy and timid when brought into the outside world.

Following the period of socialization, the juvenile period lasts up to sexual maturity, usually sometime after six months of age. During this time the puppy is still physically undeveloped and relatively unskillful, and complex training is not advisable.

Trainability. Development of the nervous system is largely completed by eight weeks of age.

Punishment versus reward

Successful training methods vary with the breed of dog and the type of activity desired. Methods that work well with one breed may work only poorly with another. Mild punishment, for example, can easily inhibit undesirable behaviour in many of the shepherd dogs, whereas similar punishment may only stimulate resistance and fighting in the aggressive terrier breeds.

The basis of most training is reward. To be effective the reward must come immediately after the action, so that the dog forms a connection between the two. Good habits formed in this way also help prevent the formation of bad habits, since one interferes with the other. The rewards used vary with breeds and individuals, and the trainer must study the animal to determine which type of prize is most effective. For the hunting breeds food is usually a very strong reward. For many breeds praise and fondling is reward enough.

Punishment is chiefly useful to inhibit activity and is ineffective otherwise. For example, if an owner repeatedly calls a dog to punish it for misdeeds, the dog will soon learn not to come. A general rule is to use punishment sparingly and only in situations that the dog understands. At the same time dominance must be established; it is best accomplished over very young puppies, and by restraint and handling rather than outright punishment.

Breed psychology

Specializations of breeds. The psychology of each breed can be best understood in terms of the work for which it has been selected. Hunting breeds such as hounds have been selected to work independently of human handlers and are therefore somewhat difficult to keep under strict control. They have also been selected to work peaceably in packs or groups; aggressive behaviour is seldom a problem. Shepherd dogs, on the other hand, have been selected for their ability to learn to work under direction and to form firm habits. It is therefore easy to teach them restraint.

Hunting dogs. The hound breeds of dogs have been selected to hunt primarily by sight or by scent and are used to find and run down various sorts of land mammals. Dogs such as the greyhound, which were originally adapted to desert and plains, are representative of the sight hounds, or gazehounds. The bloodhound and related breeds are examples of the scent hounds. Both types track down their quarry.

A different type of hunting is found among bird dogs. When the bird is detected on the ground, the dog, if it is a pointer or setter, is trained to stop and wait; other breeds, such as the spaniels, simply flush the birds so that they can be shot. Another problem in bird hunting is to find the shot and wounded birds and return them to the hunter. For this job several special breeds of retrievers have been developed, some of which are used for both finding and retrieving.

The spaniels appear to have been the original bird dogs, and according to tradition they originated in Spain. They were used for various types of hunting in the Middle Ages. Water spaniels were used for retrieving birds from the water. Land spaniels were used for falconry, being trained to find and flush birds off the ground so that the falcons could attack them. Others were used in setting birds for the net; *i.e.*, on finding birds, they lie flat on the ground while the net is thrown over them and the birds. This was the original meaning of "setter." Later, when shotguns were developed for use in bird hunting, this behaviour trait was useless, and setter breeds were then selected for pointing instead. The original pointers were developed about the same time, probably from a mixture of hound and spaniel breeds.

Hunting game on open plains or deserts is not so much a problem of finding the game as of catching it. Various greyhounds have been selected and bred for great speed and running down game of different sorts by sight.

The final part of hunting is the attack on the prey, and terrier breeds have been bred as attack dogs. The fox terriers were used to drive foxes out of their dens, and the larger Airedale terriers were, and still are, sometimes used to attack mountain lions and bring them to bay.

Guard dogs and watchdogs. Although guard dogs and watchdogs have had limited uses under modern civilized conditions because their owners are responsible for the damage they do, almost any breed of dog will still sound an alarm when strangers approach. Breeds such as the German shepherd dog (or Alsatian) and the Doberman pinscher are sometimes used in police work and by night watchmen. Dogs find certain use in modern warfare, chiefly in night patrols and in scouting, mine detection, and sentry duty.

Herding dogs. Under primitive conditions herding dogs were used to protect flocks of sheep or goats from predators, and they consequently had to be large and aggressive animals. Many of the modern shepherd breeds, such as the border collies, are now small or medium-size animals that have been selected for the ability to learn commands and obey them from a distance. Dogs are still used for herding cattle.

Sled dogs. One use of dogs that has no counterpart in the behaviour of wolves is as draft animals. With the domestication of larger animals and the development of modern systems of transport, this use has largely disappeared, surviving chiefly in Arctic sled dogs. Before the European carried the horse to North America, Indians used dogs to pull their wheelless vehicles (*travois*).

Companions and pets. An important modern use of dogs is as companions and household pets.

The popularity of dog breeds as pets seems to vary from year to year reflecting a current taste. The most popular are usually small or medium sized. Within the last half-century fox terriers, cocker spaniels, and beagles have been favourites, with some of the working, toy, and hound breeds such as collies, boxers, German shepherds, chihuahuas, Pekingese, miniature schnauzers, dachshunds, and basset hounds being runners-up. Among non-sporting breeds, poodles were registered in numbers more than twice as great as those of the nearest competitor, the German shepherd. Guide dogs for the blind are particularly valuable and devoted companions. Several breeds have been used successfully as guide dogs, but the German shepherd dog is most commonly employed.

Guide dogs

NUTRITION AND GROWTH

Adult dogs. Dogs are basically carnivorous hunting animals and as such are physiologically adapted to going for long periods without food or water. Many adult dogs can go without eating for a week without serious harm. When food is available, a hungry dog will gulp down large quantities as rapidly as possible.

The amount of food a dog needs depends upon the amount of exercise it gets, but a rough guide is one-half pound of dry dog food or its equivalent for a 20-pound dog. Smaller breeds generally require more food per pound of body weight than do larger breeds, and puppies need more food than mature dogs, especially during the period of rapid growth between six weeks and six months of age.

A well-balanced diet for a dog is not too different from that for a human being, except that the dog's intestine is not well equipped for handling roughage, so foods like bran and certain vegetables containing an excess of fibre should be avoided. Dogs digest bones easily and can live almost exclusively upon fresh ones that contain marrow. Splintery bones, such as those of poultry or chops are, however, dangerous. Dogs, because they can produce their own vitamin C, have no need for vegetables and fruits but can eat them if there is nothing else available. They can also readily digest cooked starchy foods.

Puppies. Puppies naturally have special nutritional requirements. In normal development, the young puppy gets all of his food from the mother's milk during the first few weeks of life. The mother then begins to supplement this with vomited food and completely weans the puppies at seven to ten weeks of age. Most mothers are unable to feed their puppies adequately by natural means throughout this whole period, and food supplementation is usually necessary at about three or four weeks of age.

The best supplement is some sort of mash containing milk and meat with large quantities of high-grade protein and also iron. A puppy's diet may be supplemented with vitamin D to prevent rickets. This is particularly important in the large breeds such as the Great Dane.

A normal rate of growth is the best indication of good health in a young puppy. For the first three weeks puppies gain between 50 and 100 percent of their birth weight each week. Once supplementary feeding is begun they gain weight very rapidly up to four months of age, then more slowly, reaching nearly adult size by six months of age. Growth thereafter is quite slow, and the dog reaches full development at approximately two years. The prime of adult life extends through the fifth year, though many dogs live more than twice that long, and some last well into the teens.

Rate of growth

REPRODUCTION

Sexual maturity. The ancestral wolves do not become sexually mature until nearly two years of age. However, females of most breeds of domestic dogs will show their first heat (or estrus) period before they are a year old and sometimes before six months. There is considerable variation both between breeds and individuals. The African basenji has a seasonal cycle, in which the females come into heat in the autumn of each year. A similar cycle is found in the Australian dingo when taken into northern latitudes. Most domestic breeds come into heat at any season of the year, at approximately six-month intervals. The pattern of any individual dog is usually fairly consistent, but longer and shorter cycles are common. The reproductive powers of females diminish after five years of age and reproductive cycles usually cease entirely by the age of eight. Males usually remain capable of breeding to a more advanced age, but a male of six years is entering middle age.

Heat, or estrus, periods

Spaying the female (removing the ovaries) or castrating a male (removing the testes) before maturity affects the normal pattern of growth; such animals usually become taller and more obese than the average. Spaying terminates the sexual cycle in the female, but castration may have little effect on the sexual behaviour of an experienced male.

Receptivity and gestation. The first sign of estrus in the female is a gradual enlargement of the external genitalia, followed after several days by the discharge of a small amount of blood. At the same time an odorous substance highly exciting to males is secreted in the urine. If the female is not allowed to urinate where males can find it, there will be little trouble, but if she is allowed to run freely, males will gather from miles around. Bleeding may continue for a week or more and at its end the female will accept the male. She may be receptive for a few days or as much as two weeks. Ovulation occurs 72 hours before the last point of receptivity. Conception may occur from matings at any time in the cycle. Many breeders make a practice of repeating matings every other day throughout the receptive period.

The period of gestation is approximately nine weeks counting from the time when the animal is first receptive. The embryos develop quite slowly at first, not becoming implanted until about 21 days after fertilization. The greatest intra-uterine growth of the puppies occurs in the last half of pregnancy, during which the female requires an increased diet. Occasionally a female that has not been bred will show a pseudopregnancy, with swelling of the abdomen and enlargement of the mammary glands.

Litter size varies roughly with the size of breed, but there is great individual variation. Some toy breeds rarely have more than one or two puppies, whereas the setters and larger breeds may have eight to ten, with some rec-

Litter size

ord litters going much higher. Four to six is a good average; in very large litters the pups are frequently small and weak.

AILMENTS

Parasites. *Worms.* There are many intestinal worms that attack dogs. The most serious one is a variety of *Ascaris*, a roundworm that can cause a high percentage of fatalities in young puppies. Most puppies acquire considerable resistance to them after three months of age, but the majority of dogs have at least light infections. Females about to be bred are usually wormed, bathed, and transferred to thoroughly clean quarters. They can also be wormed as late as seven weeks in pregnancy. If this is not done, the puppies can become infected from the mother, even before birth. Puppies can be given a mild worm expellant (vermifuge) if symptoms develop. Other parasitic roundworms that attack dogs include heartworms, hookworms, and whipworms.

With the exception of rabies, most diseases of dogs are not transmissible to human beings. A rare exception is a tapeworm, *Echinococcus granulosus*, normally transmitted from dogs to sheep through fecal matter but which can accidentally be transmitted to human beings, particularly children. The parasite grows as a cyst and may enter the brain or other vital organs and cause severe symptoms. It is found only where dogs have access to the bodies or entrails of sheep and cattle. Other less harmful varieties of tapeworms occur commonly in dogs and yield to treatment with appropriate worm expellents, or vermifuges. Dogs may become infected with these tapeworms by eating fleas and the uncooked bodies of food animals in which the tapeworms spend part of their life cycle.

Mange. Mange is caused by two varieties of mites that live in the hair follicles. An afflicted dog loses his hair, and the affected area is itchy and inflamed. The mites are difficult to kill, successful treatment requiring many weeks.

Lice. Lice spend their entire life cycle on the dog and are transferred to new hosts by direct contact. Successful treatment usually requires two or more applications of some oily substance that smothers the lice, or of an insecticide. The developing lice (or nits) are highly resistant, but adult lice are controlled readily, even with soap and water.

Fleas. Fleas present a different problem, since they leave the dog during part of their life cycle in order to breed. Those on the animal are easily eliminated with commercial preparations (sprays and powders), but their breeding areas must also be eliminated for control. Larval fleas live on filth and commonly breed around barnyards or other places where animals deposit feces.

Ticks. Ticks, which breed in grassy or bushy areas, may attach themselves to a dog as it brushes against the vegetation. Individual ticks are easily removed by applying kerosene or some other oily substance to the ticks' bodies, thus suffocating them and causing them to release their hold.

Diseases. *Canine distemper.* One of the most important diseases of domestic dogs is canine distemper, an airborne, highly infectious virus disease that attacks the nervous system. It runs a long course, with a high proportion of fatalities. The disease can be prevented by inoculation with an attenuated virus.

Puppies nursing on immune mothers are protected by the antibodies obtained in the mother's milk for a few weeks after birth and can be successfully vaccinated only after these antibodies have disappeared.

Infectious canine hepatitis. This disease attacks the liver primarily. The early symptom of high fever resembles that of distemper but the disease, which is often fatal, runs a much shorter course. Transmitted by contact through urine, it is less infectious than distemper. Many dogs are carriers, however, and spread the disease long after they have recovered. It can be prevented by inoculation.

Rabies (hydrophobia). Rabies (*q.v.*) is an invariably fatal disease that is highly dangerous because it can be spread to human beings as well as to other dogs and other mammals. The disease attacks the nervous system, chiefly, causing animals to become highly irritable. Infected dogs are so fearless that they bite anything they come across, spreading the disease through the saliva entering the wounds. Many wild animals, such as foxes and rodents, provide a reservoir for rabies. Dogs can readily be inoculated (see INFECTIOUS DISEASES).

Other ailments. Finally, dogs are susceptible to many other less common bacterial and parasitic diseases, which vary according to the life of the animal and the climate in which it lives. Dogs are also subject to constitutional ailments such as heart disease and cancer. Congenital defects such as crooked legs and cleft palates are common. As in other traits, dogs show hereditary differences in their resistance to disease, but this trait is greatly aided by good nutrition and proper exercise.

GENETICS

Not only are there wide differences between breeds but there is also great individual variability within each breed. All traits are genetically determined, but their expression is more or less modified by the environmental circumstances; behavioral traits especially are influenced by training and experience.

Colour. There are certain physical characters the inheritance of which is well-known and highly predictable in breeding experiments. Most of these are based on genes affecting colour. Only a few of the major colour-influencing genes are described below.

The wild-type coat

The basic wolf-gray or wild-type coat consists of long guard hairs, banded with black and red, and a lighter coloured undercoat, which grows heavily in winter and is shed in summer. The colour is distributed in a basic pattern of "countershading": more black hair on the top of the body and more red underneath, becoming almost white on the belly. Thus there are two basic pigments, black and red, with the absence of pigment producing white.

In domestic dogs the original coat has been modified in length, texture, pattern, and colour, in a wide variety of combinations. The A series is the major series of genes that modify colour. Given in order of relative dominance, these genes are: coal black, a^s; red, varying from clear red to red with some dark hair (the so-called sable colouring), a^y; the wild type or wolf gray, a^w; and bicolour (black-and-tan), having a clear black on the upper part of the body and clear red below, with red dots over the eyes, a^t.

Other genes may influence the A series: the recessive *ee* changes all black hair to red, and *bb* changes black hair to brown or liver colour. These pigments may be modified still further by other genes to produce all shades from dark to very light.

Another important series of genes produces various degrees of white spotting, ranging from small spots on the tail, feet, and belly through piebald (heavy mottling over the entire body) to almost completely white animals, as in white bull terriers. Hair length and coat texture are also influenced genetically, but less precisely than coat colour. In general, short hair is dominant over long, coarse over fine, straight over curly, wire coat is incompletely dominant over smooth, sparse over dense. There are an enormous number of possible combinations, each breed having a limited but often confusing number.

Shape and weight. In body form, there is an ancient mutation for the upcurved tail. This is a physiological trait, since the dog can usually straighten his tail; its inheritance is not well-known. There is major mutation for short legs, seen in such breeds as dachshunds and basset hounds; the first-generation hybrids of a mating of a short-legged dog with a normal-legged one have legs that are intermediate in length. The bulldog mutation chiefly affects the head, producing a short and flattened snout accompanied by an undershot jaw. Again, the first generation of a cross between a bulldog and a normal-faced dog has an intermediate appearance. Still another inherited trait is ear carriage, varying from lop ears to erect ones, the lop-eared condition being dominant.

One of the outstandingly variable characteristics of dogs is body weight, ranging from as little as 0.9 kilograms (two pounds) in dwarf breeds such as the chihuahua, to 68 kilograms (150 pounds) in some of the large breeds such as the mastiff and St. Bernard. Offspring from crosses between large and small breeds tend to be intermediate in size, and the trait is affected by large numbers of genes.

Other traits. The inheritance of behavioral and temperamental characteristics is highly complicated. The behaviour characteristic of a particular breed consists of a combination of several independently inherited traits, each of which is affected by one or two major genes and perhaps other minor ones. The tendency to crouch or sit in cocker spaniels, for example, depends on two independent traits: the crouching posture itself and the tendency to remain quiet.

The danger of preserving unfavourable traits

It is very likely that more than one combination of genes will produce a desirable trait in a dog; therefore, two excellent dogs bred together may produce somewhat inferior offspring, and vice versa. Consequently, most successful dog breeders experiment with matings that give the highest proportion of desirable progeny. Since dogs are long-lived and often fertile, this method of progeny testing can be highly successful. At the same time there is in most pure breeds a large number of undesirable recessive traits of form and behaviour that crop out in certain matings. Many of these are preserved and may be spread throughout a breed if, for instance, a champion male carrying an undesirable recessive trait is widely bred to numerous females. Furthermore, many of the traits desired by dog fanciers, such as the bulldog head, which would be a defect in a wild animal and promptly "selected out" in nature, are perpetuated by man through special care and attention.

BREEDS

The British classification, revised periodically by the Kennel Club of England, set the standard that other countries have followed, with some modifications. The Kennel Club recognizes two major classes of breeds: sporting and nonsporting. The sporting division includes the hound, gundog, and terrier groups; the nonsporting division includes the working, utility, and toy groups.

Variations in breed classification

Classifications of recognized breeds vary widely in other countries. The French list recognizes, in addition to hunting dogs, watchdogs, running dogs, 17 kinds of shepherd dogs, and 24 "ladies' dogs," including toy dogs and lapdogs. The German list emphasizes utility dogs and watchdogs, whereas the Swedish list includes nine different spitz breeds.

The current American classification, devised by the American Kennel Club, is somewhat different from the British. It lists six groups of breeds and embraces most of the breeds of other countries. Some breeds on the British list are not recognized in America and vice versa. The American Kennel Club provides registry service for the more than 100 breeds listed below, with the number still increasing.

Sporting breeds. These are primarily bird dogs or gundogs. The basic breeds were the medieval spaniels, from which the modern setters and pointers were developed—the latter with some admixture of hound ancestry.

Griffon (wirehaired pointing)
Pointer
Pointer (German shorthaired)
Pointer (German wirehaired)
Retriever (Chesapeake Bay)
Retriever (curly-coated)
Retriever (flat-coated)
Retriever (golden)
Retriever (Labrador)
Setter (English)
Setter (Gordon)
Setter (Irish)
Spaniel (American water)
Spaniel (Brittany)
Spaniel (clumber)
Spaniel (cocker)
Spaniel (English cocker)
Spaniel (English springer)
Spaniel (field)
Spaniel (Irish water)
Spaniel (Sussex)
Spaniel (Welsh springer)
Vizsla
Weimaraner

Hound breeds. This group includes two main types: (1) sight hounds and (2) scent hounds. The greyhounds are of the first sort, and the closely related saluki and Afghan hound are probably nearest to the original type. The borzoi and Irish wolfhound, as well as the Scottish deerhound, are related animals. The second main type comprises the scent hounds, of which the foxhounds and beagles are excellent examples. The basset hound, otter hound, bloodhound, and harrier are closely related. Other animals in this group are more miscellaneous.

Afghan hound
Basenji
Basset hound
Beagle
Bloodhound
Borzoi (or Russian wolfhound)
Coonhound (black and tan)
Dachshund
Deerhound (Scottish)
Foxhound (American)
Foxhound (English)
Greyhound
Harrier
Norwegian elkhound
Otter hound
Rhodesian ridgeback
Saluki
Whippet
Wolfhound (Irish)

Working breeds. The largest number of these breeds are derived from various sorts of herding and farm dogs. The second largest group are the guard dogs, again of miscellaneous origin. Finally there are the sled dogs from various Arctic regions.

Alaskan Malamute
Belgian Malinois
Belgian sheepdog
Belgian Tervuren
Bernese mountain dog
Bouvier des Flandres
Boxer
Briard
Bullmastiff
Collie
Doberman pinscher
German shepherd dog
Great Dane
Great Pyrenees
Komondor
Kuvasz
Mastiff
Newfoundland
Old English sheepdog
Puli
Rottweiler
St. Bernard
Samoyed
Schnauzer (giant)
Schnauzer (standard)
Shetland sheepdog
Siberian husky
Welsh corgi (Cardigan)
Welsh corgi (Pembroke)

Terrier breeds. These are typically dogs developed to attack vermin living in the terre, or earth.

Airedale terrier
Australian terrier
Bedlington terrier
Border terrier
Bull terrier
Cairn terrier
Dandie Dinmont terrier
Fox terrier
Irish terrier
Kerry Blue terrier
Lakeland terrier
Manchester terrier
Norwich terrier
Schnauzer (miniature)
Scottish terrier
Sealyham terrier
Skye terrier
Staffordshire terrier
Welsh terrier
West Highland White terrier

Toy breeds. These are very small dogs of various origins. Some are merely dwarf editions of larger breeds, in fairly normal form and proportion. Others have heads of the bulldog type, with short flat noses.

Affenpinscher
Chihuahua
English toy spaniel
Griffon (Brussels)
Italian greyhound
Japanese spaniel
Maltese
Papillon
Pekingese
Pinscher (miniature)
Pomeranian
Poodle (toy)
Pug
Silky terrier
Toy Manchester terrier
Yorkshire terrier

Nonsporting breeds. This is a miscellaneous group used entirely for companions and show dogs.

Boston terrier
Bulldog
Chow chow
Dalmatian
French bulldog
Keeshond
Lhasa Apso
Poodle (miniature)
Poodle (standard)
Schipperke

Other breeds. The Kennel Club (Great Britain) greatly expanded its list in the 1960s and recognized the following breeds not found in the American classification. Hounds: dachsbracke, Finnish spitz, Ibiza hound, Pharaoh hound. Gundogs: German longhaired pointer, Italian spinone, kleine Munsterlander. Terriers: Glen of Imaal, miniature bull, Norfolk, soft-coated wheaten terrier. Utility dogs: Iceland dog, Japanese akita, Leonberger, Mexican hairless, Shih Tzu, Tibetan spaniel, Tibetan terrier. Working dogs: Anatolian sheep-dog, Australian kelpy, bearded collie, beauceron, Maremma Italian sheepdog, Norwegian buhund, Polish sheepdog, Portuguese water dog, Tibetan mastiff. Toys: bichon frise, Chinese crested dog, lowchen. Many American breeders have not sought recognition for certain pure-

bred varieties that they have developed. Others have registered their distinct breeds in organizations other than the American Kennel Club. There are many special strains of hounds, often originating from certain famous packs and some named after their owners. Among American foxhounds, the Walker, Trigg, July, Trumbo, and Birdsong strains are well-known, and coonhounds include the Bluetick, Redbone, Plotthound, and Treeing Walker varieties as well as the black and tan variety recognized by the American Kennel Club.

Other breeds found in various parts of the world include the border collie, Australian cattle dog, Drahthaar, Drentsche partrijshond, Catalan sheep dog, Istrian pointer, lurcher, Portuguese pointer, Rumanian sheepdog, Sealydale, Svensk vallhund, and many more.

BREEDERS' ASSOCIATIONS AND DOG SHOWS

In western Europe and North America, dog breeding has been highly developed as a pastime and business. Dog shows and systematic attempts to improve and maintain dog breeds originated in the latter half of the 19th century. The first recorded dog show was held in Newcastle, England, in 1859, and the first large show was held in Chelsea in 1863. About the same time, the showing of dogs became popular in the United States, and by 1880 an annual show in New York included about 29 breeds.

As dog shows grew in number, a need was felt for some kind of regulating body. The Kennel Club of England filled this need in Great Britain; it was founded in 1873 and became the supreme governing body of dog breeders' associations in that country. A few years later, in 1884, the American Kennel Club was formed, becoming the ruling body of breeders' associations in the United States. Soon thereafter similar organizations were formed in many other countries.

Dog shows are organized by local or national dog clubs. Some dog clubs are devoted to only one breed, whereas others include any breed. Rules for holding shows in the United States are made by the American Kennel Club, and in Britain by the Kennel Club of England.

Types of shows

Show dogs are judged on the basis of breed standards, the various physical characters considered desirable and those considered faults in any particular breed. These standards are not directly concerned with health, vigour, or ability to reproduce; in fact, the quality of the show breeds sometimes suffers in these respects. A popular feature of many shows is the obedience trial, in which dogs are judged on performance rather than appearance.

Field trials, held for hunting breeds, give an opportunity for dogs to compete against members of their own breeds in the performance of hunting duties such as trailing, pointing, and retrieving. There are also standard field trials for shepherd dogs, in which the dog has to herd a small flock of sheep along a prescribed course, cut sheep out of a flock and drive them in a small pen, relying only on signals given by the shepherd. Other working trials are popular in many parts of the world.

BIBLIOGRAPHY. The rules applying to registration and dog shows, lists of registered breeds, and information about specialty breeding clubs may be obtained from the American Kennel Club, New York, and the Kennel Club, London.

AMERICAN KENNEL CLUB, *The Complete Dog Book*, rev. ed. (1969); E.C. ASH, *Dogs: Their History and Development*, 2 vol. (1927); M. BURNS and M.N. FRASER, *Genetics of the Dog: The Basis of Successful Breeding*, 2nd ed. (1966); H.P. DAVIS (ed.), *The New Dog Encyclopedia* (1970); SPORTS ILLUSTRATED, *Book of Dog Training* (1960); M.W. FOX, *Canine Behavior* (1965); C.C. LITTLE, *The Inheritance of Coat Color in Dogs* (1957); C.M. MCCAY, *Nutrition of the Dog*, 2nd ed. (1949); J.Z. RINE, *The World of Dogs* (1965); J.P. SCOTT and J.L. FULLER, *Genetics and the Social Behavior of the Dog* (1965); RICHARD and ALICE FIENNES, *The Natural History of the Dog* (1968); R.F. WALL, *Keeping a Dog*, 5th ed. (1957); L.F. WHITNEY, *The Complete Book of Dog Care* (1953) and *Dog Psychology: The Basis of Dog Training*, 2nd ed. (1971).

Domestication, Plant and Animal

Domestication is the process of hereditary reorganization of wild animals and plants into domestic and cultivated forms according to the interests of man. In its strictest sense it refers to the initial stage of man's mastery of wild animals and plants. The fundamental distinction of domesticated animals and plants from their wild ancestors is that they are created by man's labour to meet his specific requirements or whims and are adapted to the conditions he alone maintains for them. Without man's continuous care and solicitude, domesticated animals and plants could not exist.

GENERAL FEATURES

The significance of domestication

Domestication has played an enormous part in the development of mankind and its material culture. It has provided man with improved sources of food, materials for clothes and shelter, and means of conveyance and physical work. These benefits contributed enormously to the liberation of man from elemental natural forces and, together with the mastery of fire, were the decisive factors in the passing of mankind from wildness to civilization.

Domestication has resulted in the appearance of agriculture as a special form of animal and plant production. It is precisely those animals and plants that became objects of man's agricultural activity that have undergone the greatest changes when compared with their wild ancestors. Other animal and plant species are bred by man merely to satisfy his whims or aesthetic desires, such as, for example, the pigeons, aquarium fishes, and garden and house plants. While these organisms played no role in the development of the material culture of mankind, they are still considered as domesticated because of the profound changes they have undergone in comparison with their wild ancestors. Animals and plants that are bred by man as objects for scientific studies, however, are not considered as domesticated.

The domestication of most animals and plants is rooted in the remotest past. The history of domestication, which remains in many points obscure, disputable, and contradictory, is gradually being reconstructed from evidence obtained from archaeological studies and from modern systematic, morphological, cytological, and genetic data. The study of ancient literatures and ancient languages sometimes also provides material for the understanding of domestication.

The first attempts at domestication of animals and plants apparently were made in the Old World by peoples of the Mesolithic Period. The tribes that engaged in hunting and in gathering wild edible plants made attempts to domesticate dogs, goats, and possibly sheep as early as 9000 BC. It was not until the Neolithic Period, however, that primitive agriculture appeared as a form of social activity, and domestication was well under way. Although the great majority of domesticated animals and plants that still serve man were selected and developed during the Neolithic Period, a few notable examples appeared later. The rabbit, for example, was not domesticated until the Middle Ages; the sugar beet came under cultivation as a sugar-yielding agricultural plant only in the 19th century; and mint became an object of agricultural production as recently as the 20th century. Also in the 20th century, a new branch of animal breeding was developed to obtain high-quality fur. The objects of fur breeding—the silver fox, polar fox, mink, and coypu, to name a few—are now at the first stage of domestication, as are also some deer bred by man for the alleged curative properties of their antlers. New plant species have also been introduced to cultivation relatively recently, among them the ginseng, which is intensively cultivated in Eurasia as a medicinal plant. Thus, the process of domestication of animals and plants continues to meet the many different material and aesthetic requirements of man.

THE CIRCUMSTANCES SURROUNDING DOMESTICATION

Human settlements as a precondition. Plant and animal breeding, even under conditions of primitive agriculture, required a settled way of life and a certain number of people in settlements, which, in turn, would be impossible without sufficient food resources. Settlements,

therefore, were located near rivers and other freshwater reservoirs, which offered possibilities for fishing. Mountain valleys and their plateaus, which contain a great variety of plant and animal forms, offered many candidates for domestication. It is, therefore, supposed that the first domestication centres were forested valleys with clearings that could be used for primitive agriculture.

The beginnings of domestication

The reasons that impelled man to domesticate the first plants and animals and the circumstances that contributed to it are unknown. The idea that the beginnings of domestication were imposed on many by hunger does not seem plausible. A man living under permanent pressure of hunger would very likely not engage in domestication, which involves the storage of seeds and the breeding of animals. It is far more likely that a hungry man would eat the means of continued production—the seeds and the stock. More plausible is the hypothesis that domestication was initiated by people already organized in settled communities that had access to the necessary minimum of food, probably from nearby lakes or rivers. Decisive in the development of the domestication process was the gradually acquired understanding that domesticated plants and animals are the most reliable sources of many of man's vital requirements.

The impulsion to domesticate. It is difficult to tell precisely what were the direct reasons that impelled primitive settlers to undertake the first domestication experiments. Perhaps having already experienced the taste of certain wild plants, especially those that reproduce themselves by tubers or rootstocks, man tried to grow them near his dwelling place. The very first success of this experiment convinced the settler of the usefulness of cultivation and naturally led to repetitions and, subsequently, to improvement of its methods. In this initial stage of plant domestication, using mainly vegetatively reproducing plants, particularly the tuberous sorts, man first assured himself of a reasonably reliable source of carbohydrates.

Next came domestication of the seed plants—cereals, legumes, and other vegetables. Some plants were domesticated for the strong fibres in their stalks, which man used for such purposes as making fishing nets. Hemp, one of the most ancient plants domesticated in India, is an example of a multipurpose plant: oil is obtained from its seeds, fibres from its stalk, and the narcotic hashish from its flowers and leaves.

Some plants were domesticated especially for the production of narcotics; such a plant is tobacco, which was probably first used by American Indian tribes for the preparation of a narcotic drink and only later for smoking. The opium poppy is another example of a plant domesticated solely for a narcotic. Beverage plants of many kinds were discovered and cultivated, including tea, coffee, and cola. Only when man reached a sufficiently high level of culture did he begin to domesticate to fulfill his aesthetic requirements for the beautiful and the bizarre in both plants and animals. Thus, plant domestication and, later, plant breeding became possible because man, at the dawn of his culture, could appreciate the properties and qualities of many wild plants and use those that met his special needs in the best way.

The direct reasons that impelled man to domesticate the first animal are less understandable than the reasons for plant domestication. The process may have begun as the result of sporadic cases of man rearing an animal. There is a widespread and quite understandable idea that hunters who killed an adult animal would take along any young ones, which would become children's and women's pets. Lending credence to this notion is the fact that even today the women of some peoples of tropical America and Polynesia nurse certain animals who may have lost their mothers.

Economic application of domestication

The specific economic application of domesticated animals did not appear at once. Dogs probably accompanied hunters and helped them hunt wild animals; they probably also guarded human settlements and warned the inhabitants of possible danger. At the same time, they were eaten by man, which was probably their main importance for man during the first stages of domestication. Sheep and goats were also eaten in the initial stages of domestication but later acquired those qualities of milk production and of wool and down that made them valuable for those commodities.

One point of view concerning the domestication of cattle is that such animals were first used as sacrificial animals, and thus religion was the impelling force for domestication. Despite documentation of such ritual use of cattle, the principal aim of cattle breeding in ancient times was to obtain meat and skin and to produce work animals, which greatly contributed to the development of agriculture. Cattle, at the initial stages of domestication, produced a small amount of milk, sufficient only to rear their calves. The development of high milk yield in cows with their breeding especially for milk production is a later event in the history of domestication.

The first domesticated horses were also used for meat and skin. Later the horse played an enormous role in the waging of war. Peoples inhabiting the Middle East in the 2nd millennium (2000–1000 BC) used horses in chariot battles. With time the horse began to be used as transportation. In the 1st millennium (1000–1 BC), carts appeared, and the horses were harnessed to them; other riding equipment, including the saddle and the bit, seems to have appeared in later centuries.

The donkey and the camel were used only for load transport and as means of conveyance; their unpalatability ruled out their use as a preferred food.

The first domesticated hens perhaps were used by man for sport; cockfighting was instrumental in bringing about the selection of these birds for larger size. Cocks later acquired religious significance. In Zoroastrianism the cock was associated with protection of good against evil and was a symbol of light. In ancient Greece it was also an object of sacrifice to gods. It is probable that egg production of the first domesticated hens was no more than five to ten eggs a year; high egg yield and improved meat qualities of hens developed at later stages of domestication.

Early domestication of the cat was probably the result of the pleasure experienced from keeping this animal. The cat's ability to catch mice and rats was surely another reason that impelled man to keep cats at his home. In ancient Egypt the cat was considered a sacred animal.

Some animals were domesticated for utilitarian purposes from the very beginning. Here belongs, first of all, the rabbit, whose real domestication was carried out from the 6th to the 10th centuries AD by French monks. The monks considered newborn rabbits fish and ate them when the church calendar indicated abstinence from meat.

For the sake of honey, the bee was domesticated at the end of the Neolithic Period. Honey has played an enormous role in man's nutrition since ancient times; it ceased being man's sole sweetening agent only about 200 years ago. Bees also provided man with wax and with bee venom, which was used as medicine. Bees were used also, to a limited extent, in warfare, hives being thrown among enemy troops to rout them.

To obtain silk, the silkworm was domesticated in China no later than 3000 BC, and by 1000 BC the technology of silkworm breeding and raising had been thoroughly documented.

Thus, domestication, which appeared under the influence of different circumstances as a natural event in the social and cultural evolution of mankind, has involved increasingly larger numbers of wild animals and plants and has resulted today in the creation of many thousands of varieties of domesticated and cultivated forms without which man's existence is unthinkable.

THE MAIN CENTRES OF DOMESTICATION

Domestication of plants and animals was not a onetime event, and it was not a prerogative of any one people. The existence of centres of origin of plants and animals, however, does not mean that domestication was realized within narrow geographical zones.

Centres of origin of cultivated plants. Contributions to the modern knowledge of the main centres of origin of

cultivated plants were made by the Soviet biogeographer N.I. Vavilov. His works, coupled with later archaeological findings, indicate the main areas of domestication and the formation of primary plant-breeding cultures.

Vavilov's "gene-centres"

As the main criterion for localization of the centres of origin of cultivated plants, Vavilov investigated the geographical locales of wild ancestors of modern cultivated plants. He called these locales "genecentres" and distinguished between primary genecentres, regions where the ancestors of modern cultivated plants are still present, and secondary genecentres, regions where certain plant species found their greatest expansion and where the greatest number of their varieties appeared. Many plants and animals appear from the evidence to have arisen in more than one genecentre. Accumulated evidence since Vavilov's time has suggested the following eight principal genecentres of plant origin and domestication: Indonesia–Indochina, China–Japan, middle Asia, Middle East, Mediterranean, Africa, South America, and Central America.

Indonesia–Indochina. This genecentre occupies vast areas of Southeast Asia and is the area of origin mainly of vegetatively propagating plants, such as the banana, sugar palm, sago palm, coconut palm, breadfruit, bamboo, sugarcane, taro, yam, and durian.

According to the U.S. biogeographer C.O. Sauer, the most ancient domestication appeared within this genecentre, perhaps near what is now India–Indochina, probably in the early Neolithic Period. After many generations of purely vegetative propagation, many plants that had been able to propagate themselves by seeds lost this ability under continued selection by man. They changed in other respects, as well, and formed what are called clones of cultigens, plants vegetatively related and maintained by man. Among man's cultigens are the banana, taro, yam, sugarcane, sago and sugar palms, durian, and mangosteen.

China–Japan. The China–Japan genecentre (east China, Korea, and Japan) produced such plants as rice, millet, soybean, perilla, tea, tung, and mulberry. Some of the above-mentioned plants have important secondary genecentres; such is the soybean, which, after entering North America in the 19th century, began a diversification into a great number of varieties.

Middle Asia. This is the southwest Asia genecentre of Vavilov. It includes Afghanistan, Tajikistan, Uzbekistan, and west Tien Shan and is the genecentre for peas, flax, carrot, onion, almond, walnut, grape, alfalfa, and a number of fruit trees and berries.

Middle East. This area includes Turkmeniya, Iraq, Transcaucasia, Asia Minor, and Arabia. It is the primary centre of origin of wheat, rye, barley, oats, chick-pea, lentil, vetch, melon, chestnut, pear, quince, and hazelnut. Independently of the middle Asia genecentre, peas and onions were also domesticated in the Middle East genecentre.

Mediterranean. This region is the primary centre of origin of oats, olive, cabbage, rutabaga, flax, lupines, oak, and lavender.

In addition, the Mediterranean area is a secondary genecentre for many plants primarily domesticated in the Orient, including large-seeded pod-bearing plants and improved wheats in a large number of varieties.

Africa. Africa is the primary genecentre for sorghum, African rice, sesame, castor bean, cotton, watermelon, cowpea, coffee, oil palm, and kola nut, and a secondary genecentre for many varieties of wheat, barley, and oats.

South America. The South American genecentre, extending over Brazil, Argentina, Peru, and Chile, seems to be the most ancient focus of plant domestication and primitive agriculture of the New World. Here, in the 3rd and 2nd millennia (3000–1000 BC), there already existed settled tribes engaged in plant cultivation. It is possible that some plants had been cultivated still earlier. As in the Old World, the first domesticated plants were vegetatively propagated tuberous species rich in starch: manioc, sweet potato, yautia (*Xanthosoma*), sorrel (*Molina*), ullucu (*Ullucus tuberosus*), and tuberous nasturtium (*Tropaeolum*).

Later, seed plants, such as the tomato, peanut (groundnut), and pineapple, were also domesticated. A large number of potato varieties appeared, but the one that was cultivated became the domestic potato now used over most of the world.

Central America. This genecentre spreads over the area of Mexico, Guatemala, Costa Rica, Honduras, and Panama. Settled agriculture in this area developed in the 2nd millennium, but as early as the 6th millennium (6000–5000 BC) some domesticated plants (*e.g.*, pumpkin and peppers) were known and used.

This region is the origin of such plants as corn (maize), kidney bean, pumpkin, and cotton. In the same region a series of potato species was domesticated, some of them still being cultivated in Central and South America. This genecentre is the homeland also of red pepper, avocado, sunflower, agave, cocoa, and tobacco.

Plant–animal complexes. According to Sauer, animals that lived near man's dwellings (such as dogs and pigs and, later, chickens, ducks, and geese) were domesticated in the regions of domestication of vegetatively propagating plants. Thus, in Southeast Asia there appeared a peculiar complex of domesticated animals and plants that spread in three main directions: northward to the territory of modern China and Japan; westward and southwestward to Equatorial and South Africa; northwestward to Central Asia and Asia Minor and farther on, across the Mediterranean coast toward the Atlantic.

The spread of domestication

The pig spread especially far northward—to China, where it gave origin to many domestic breeds. The distribution of the pig to northwest Africa and to middle Asia and Asia Minor, however, met with serious obstruction. The obstacles consisted, on the one hand, of unfavourable climate and disease organisms and, on the other hand, of the religious laws of certain faiths that prohibited the eating of pork. In this connection, the Asian pig in Africa either died out or became wild; nevertheless, it penetrated into the Middle East and farther on into the Mediterranean area and the Caucasus.

The chicken, goose, and duck, domesticated at first in Southeast Asia, spread all over the world; however, the goose and duck, which were bred in great numbers in Egypt, Greece, ancient Rome, and China, were probably domesticated in those countries during the 1st millennium BC independently of the Southeast Asian centre.

In southwest Asia, in the regions of domestication of seed plants, the domestication of herd animals, such as cattle, sheep, goats, and horses, took place.

It is difficult to tell which of the centres of domestication and of primary agriculture is more ancient, the east Asian or the west Asian. Sauer claimed that the east Asian was the more ancient. Later investigations suggest that the west Asian agriculture began earlier, in the Mesolithic or early Neolithic Period. Possibly, both of these centres appeared and developed simultaneously and independently of each other; in any case, there is no recorded evidence that the east Asian domestication centre is more ancient than the west Asian.

Shepherdy and nomadic animal breeding, which determined the social and economic organization and the way of life of some peoples to a great extent, appeared at later stages of human development, after the accumulation of a large number of domestic animals. Rudiments of nomadic animal breeding in Eurasia appeared no earlier than 1000 BC, considerably after the domestication of animals took place. The only exception seems to be the deer, whose domestication was carried out by nomadic tribes of central and north Europe.

The process of domestication and formation of primary agricultural and animal-breeding cultures in the New World took place somewhat later than in the Old World and independently of the latter. This is connected with the fact that man first appeared in the New World only during the Pleistocene, long after he had settled vast areas in the Old World.

A SURVEY OF SELECTED DOMESTICATED ORGANISMS

In the following section a few prominent plants and animals will be considered to exemplify the method of in-

vestigation of origin, the special biological event that gave rise to the domesticated animal or plant when evidence is clear enough to warrant it, and the movements of organisms from their points of origin.

Criteria for distinguishing domestic from wild organisms

Traditionally, the main criteria for judging relationships between domestic or cultivated organisms and wild ancestors are similarities of structure and function, but cytogenetical examinations, particularly comparisons of chromosomes and chromosome sets, also are adding to the knowledge of the origins of domesticated organisms.

The rise of cultivated crop plants. *Wheat.* Cytogenetics helped to establish the origin of wheat, one of the world's most important food crops. Species of wheat differ in the degree of ploidy—*i.e.*, in the number of chromosome sets in their cell nuclei. The einkorn wheat, for example, which seems to have been involved in cultivation earlier than all other wheats, has 14 chromosomes (the common diploid set). There then developed tetraploid wheat, with 28 chromosomes, such as durum wheat, which is used mainly for macaroni products, and hexaploid wheat, with 42 chromosomes, to which belong all the varieties of common wheat.

The einkorn, which is now cultivated only in some parts of Spain and Transcaucasia, takes its origin from a diploid 14-chromosome wild species. The tetraploid wheats appeared as the result of crossing the einkorn with a wild, 14-chromosome cereal, *Aegilops* (or *Triticum*) *speltoides*. Some of the tetraploid species seem to have been domesticated between 8000 and 7000 BC and certainly have been domesticated in the Orient for a few thousand years. Subsequent natural crossing of tetraploid wheats with another wild cereal (*Aegilops squarrosa*, or *Triticum tauschii*) resulted in the appearance of 42-chromosome, or hexaploid, wheats. These crossings all took place directly in nature in the zone of the primary origin of wheat, and man simply domesticated already existing, though still primitive, forms of wheat. With subsequent selection of wheat in many countries for many centuries, great variety and high perfection of forms of this plant have resulted.

Rice. Rice is another most important crop, consumed by almost half the world's population. Of about 20 wild rice species known in Asia, Africa, America, and Australia, only two have been cultivated: the Asian species *Oryza sativa* and the African species *O. glaberrima*, both of which spring from perennial rice, *O. perennis*. In Southeast Asia the cultivation of rice has been known for 4,000 to 5,000 years and seems to have begun still earlier.

Corn. Corn (maize), a New World plant, belongs to a genus having only one species, *Zea mays*, with a series of subspecies differing in morphological peculiarities of the grain. In the process of domestication, man selected corn for well-developed husks, which protected the ear from breaking, and for ears surrounded by leaf sheath. During domestication, corn hybridization seems to have taken place with such cereals as teosinte and gama grass, with which some forms of corn still show similarities. Corn spread into Europe through Spain after Columbus' second voyage (1494) and then spread all over Eurasia and Africa.

Barley. Barley belongs to the genus *Hordeum*, which includes more than 20 wild species, mostly perennial; the domesticated forms, however, are annual. As shown by cytogenetical studies and archaeological findings, only one species seems to have been domesticated (as early as the period 7000 to 6000 BC), namely, *H. spontaneum*, which bore grain in two vertical ranks. Modern barley, with six vertical ranks, appeared in the process of further domestication as a result of genetic mutations.

Oats and rye. Oats and rye were domesticated much later, in the epoch when man had already mastered metals. Before that, these grains existed as wild plants contaminating wheat sowings. As the wheats spread to the north, the volunteer oats and rye, being cold resistant, survived, while the wheat perished. In such a manner, natural selection helped man involve oats and rye in his cultivation.

Modern cultivated oats (*Avena sativa*) seem to take their origin from *A. byzantina*, and cultivated rye, which is now represented only by *Secale cereale*, springs probably from the *S. segetale*.

Potato. The origin of the potato is a more complex matter. Among the many species of this plant found in Central and South America, cultivated varieties and forms are represented primarily by the tetraploid, 48-chromosome species *Solanum tuberosum*. This species, which has not been found wild, has not yet been related clearly to an ancestral form. In South America another 48-chromosome potato, *S. indigenum*, is also cultivated; it differs from *S. tuberosum* in certain structural and functional features but so slightly that some experts consider them a single species.

Soya. Soya, one of the most ancient leguminous plants, belongs to the genus *Glycine*, which numbers more than 40 wild species. The ancestor of cultivated soya is probably *G. ussuriensis*, which was domesticated in China in about 3000 BC and then penetrated, already as a cultivated form, into Korea and Japan and, in the 19th century AD, into other countries.

Sunflower. One of the main oil-producing plants is the sunflower, whose genus, *Helianthus*, numbers about 50 wild species in its homeland, North America; more than 15 more species grow in South America. The cultivated sunflower is considered to have sprung from the wild species *H. lenticularis*, inhabiting the dry plains of North America. Much breeding and selection of the sunflower to yield varieties containing seeds with more than 50 percent oil have been carried on in the Soviet Union.

Pumpkin. The origin of the pumpkin, one of the most important ancient vegetables, is obscure. In Central and South America there are a few pumpkin species; two have been cultivated, *Cucurbita pepo* and *C. moschata*.

Cotton and flax. Among fibrous plants, the most important is cotton (*Gossypium* species), wild species of which inhabit all the continents except Europe. Four species have been cultivated: two 26-chromosome species in the Old World (*G. herbaceum* and *G. arboreum*) and two in America (*G. raimondii* and *G. thurberi*). Currently, two tetraploid species (*G. hirsutum* and *G. barbadense*), which originated in South America, are cultivated. Cytogenetical analysis has shown that both species originated from a cross between diploid cottons of the Old World and the New World, with subsequent duplication of chromosome number. Where and when this hybridization took place, however, remains obscure. In Mexico, products of cotton have been found in excavations dating from about 2400 BC. The seeds of tetraploid cotton were carried to Africa and Asia after the discovery of America by Columbus.

The second most important fibrous plant is flax (*Linum* species). Cultivated flax (*L. usitatissimum*), which has differentiated into oil-producing and fibre-producing forms, is considered to take its origin from the wild flax, *L. angustifolium*. Flax was first cultivated in mountainous regions of India and China in about 5000 BC; it spread westward, where, in the mountains of middle Asia, there appeared the oily-crown flax, and, farther on, in the Mediterranean area, fibre flax.

Tobacco. Tobacco belongs to the genus *Nicotiana*, which includes more than 60 species spread over North America and Australia. The homeland of the cultivated tetraploid tobacco, *N. tabacum*, is Mexico, where it originated from a cross between *N. silvestris* and *N. otophora* (or *N. tomentosiformis*), with subsequent duplication of chromosome number. It reached the Old World after Columbus.

Tea. Tea is a shrub of the genus *Thea*. According to some experts, at least four species of cultivated tea occur in east Asia. In south China the ancestor of cultured species, *Thea monticola*, is widespread. Tea was taken into culture in south China and, perhaps independently, in Indochina.

Coffee. Coffee belongs to the genus *Coffea*, the 50 or 60 species of which inhabit Africa. Most of the species are diploid and have 22 chromosomes. The cultivated species *C. arabica* also includes forms with chromosome numbers 44 and 66. It grows in Ethiopia in a wild or es-

caped state and is cultivated in the tropical countries of Africa, Asia, and America. It was first taken into culture in the 15th century AD, in Arabia. From there coffee passed to Europe and, in the 18th century, to America.

Apple tree. One of the oldest fruit trees is the apple (*Malus* species), with a few dozen wild species in east and west Asia, in Europe, and in North America. The great majority of current varieties are hybrids; some developed as the result of bud mutations, or sports, on an otherwise uniform tree.

Grape. Grape belongs to the genus *Vitis*, which includes about 50 species inhabiting the Northern Hemisphere, mainly North America, where some wild species have been cultivated since ancient times for both food and wine production.

In Eurasia the grape was domesticated independently and was introduced into cultivation 7,000 to 9,000 years ago. The cultivated grape of the Old World *V. vinifera* seems to have originated from a wild species, *V. sylvestris*, which spread from Gibraltar to middle Asia. From different local forms of this species originated various protected cultivated forms.

The rise of domestic animals. The questions of the origin of domestic animals are also worked out on the basis of comparative morphological, physiological, and partly genetical and cytogenetical studies. Biochemical methods, including studies on blood composition, are also used.

Dog. The origin of the dog has provoked much dispute, which is understandable when the enormous variety of modern breeds is considered. It has been supposed that man domesticated wolves and jackals in different locales or that there existed a special ancestor of the modern dog (see also DOG). Currently, another point of view holds that the wolf (*Canis lupus*) was the only wild ancestor of the domestic dog. It was domesticated in different places by different tribes. Then the domesticated wolf dog (*C. familiaris*) seems to have backcrossed repeatedly with the wolf in different places.

Sheep. The domestic sheep originated from different subspecies (or races) of the Eurasian wild sheep *Ovis ammon*. Another wild sheep, *O. canadensis*, uniting a series of subspecies in northeast Asia and America, seems not to have played an important role in the origin of domestic sheep. There seem to have been two main domestication centres of wild sheep: the Middle East and middle Asia. The earlier prevailing idea about a south European centre of sheep domestication is now doubtful. Under intensive domestication a great variety of specialized breeds appeared for the production of different kinds of wool, of meat, and sometimes of milk.

Goat. The wild ancestor of the domesticated goat is *Capra hircus*, belonging to the same subfamily (Caprovinae) as the sheep. It is supposed that the only wild subspecies from which the domestic goats took their origin was *C. hircus aegagrus*, inhabiting the mountainous regions of Asia Minor, which seem to have been the main focus of its domestication.

Cattle. Modern cattle appear to have descended from the aurochs (*Bos primigenius primigenius*). There seem to have been several centres of domestication: eastern Europe, middle Asia, and Southeast Asia.

The hump characteristic of zebu-like cattle appeared early during domestication. It is thought that India was the homeland of primitive cattle and that they then expanded to Eurasia and North Africa. These primitive cattle were characterized by a great variety of forms, from which, by subsequent selection, man obtained many breeds with high degrees of specialization and uniformity as meat or milk producers or as beasts of burden.

Some species allied to the aurochs also underwent domestication; *e.g.*, the yak (*Poephagus grunniens*), ranging in the mountains of Tibet, in Pamir, in some regions of middle Asia, and in south Siberia. In India, Indochina, and the islands of the Malay Archipelago, species belonging to the genus *Bibos* were domesticated, including the banteng (*B. banteng*), also called Bali cattle, and the gayal (*B. frontalis*), especially selected and bred in India and Burma.

Horse. The origin of the domestic horse (*Equus caballus*) is extremely problematic. The ancient ancestors of modern horses possibly originated in Asia, flourished during the Oligocene Epoch (about 38,000,000 to 26,000,000 years ago) in North America, and then died out, but not before their representatives migrated to the Old World. They quickly spread all over Eurasia, some even reaching Africa. The direct ancestor of the modern domestic horse is thought to have inhabited a wide area from western Europe to northern Asia. One of the races of this ancestor, Przewalski's horse (*E. caballus przewalskii*), is believed to be the form that was domesticated by man. Przewalski's horse ranged the steppes of south Siberia, Mongolia, and northeast China, where small herds are now protected. On the basis of chromosome-number studies, some experts believe that Przewalski's horse is a different species, though they do not deny the origin of the domestic horse from Przewalski's horse. Other experts suggest that both Przewalski's and the domestic horse originated from a common ancestor and have evolved independently of each other.

Camel. There are few conclusive data on the origin of the domestic camel. Although the two-humped camel (Bactrian) was domesticated in the Middle East and the one-humped camel (Arabian) was domesticated in middle Asia, they are both assigned to the genus *Camelus*. An allied genus, *Lama*, includes the guanaco, llama, alpaca, and vicuña. Not much is certain about their origins except that both of these genera are believed to have originated from North American ancestral forms.

Pig. Until recently it was supposed that the domestic pig originated from different species of wild pigs belonging to the genus *Sus*. Now, however, it is believed that the only ancestor of the domestic pig is the wild species *Sus scrofa*, whose geographic range is very wide throughout Eurasia. In Southeast Asia and Indochina a few local forms of the wild pig were domesticated; of them, *Sus vittatus* provided the greatest contribution in the formation of domestic breeds. Independently in Europe, the local race of *Sus scrofa* was domesticated and gave origin to a great number of European breeds. Many modern breeds are the result of hybridization between breeds of Asian and European origins.

Cat. The domestic cat, *Felis catus*, originated from the African wildcat, *Felis lybica*, whose area included the Balearic Islands, Corsica, Sardinia, Africa, Arabia, and India. The domesticated cat seems to have been crossed repeatedly with the European wildcat, *F. silvestris*, which itself has never been domesticated.

Chicken. The main ancestor of domestic chickens is the red jungle fowl *Gallus gallus*, whose original home is chiefly India. It is possible that, at early stages of domestication, *G. gallus* crossed with a kindred species, *G. sonnerati*, and that the resulting form gave rise to later breeds.

Rabbit. The only ancestor of the domestic rabbit, *Oryctolagus cuniculus*, is a wild species whose area, after the last glaciation, occupied the Iberian Peninsula but then spread southward to Morocco, Algeria and northward over Europe.

Bee and silkworm. Two bee species are involved with man: *Apis mellifera*, whose natural range is Africa, Europe, and the Middle East, and *A. cerana*, native to India, Indochina, and Japan. Of the greatest economic importance all over the world is *A. mellifera*, and several distinct bee breeds have been created from geographical varieties.

Out of about ten silk-producing lepidopteran species, only one has been domesticated, the silkworm moth (*Bombyx mori*), whose caterpillars feed on mulberry leaves. Although other moths spin silk cocoons, these species have not been truly domesticated and have no economic importance.

THE BIOLOGICAL CHANGES ACCOMPANYING DOMESTICATION

Consequences of domestication

Expressed alterations. During the 10,000 or 11,000 years that have passed since the beginning of domestication, the animals and plants that man has selected as useful to him have undergone profound changes. The con-

sequences of domestication are so great that the differences between breeds of animals or varieties of plants of the same species often exceed those between different species under natural conditions. Even the most superficial comparison of the tremendous range of shapes and sizes of modern dog breeds or of garden plants is sufficient to establish the extent of the alterations brought about by domestication and the rate at which evolution accelerated under domestication.

Domesticated animals differ from their wild ancestors and from each other in a number of physical and biochemical features and, most importantly, in behavioral traits. Especially changed are the characters for which man originally chose to domesticate the animals. The wild ancestors of cattle gave no more than a few hundred grams of milk; the best milk cow now can yield 12,000 to 15,000 litres of milk during its lactation period.

The character of sheep wool has thoroughly changed since sheep underwent domestication. In the ancestors of domestic sheep, wool (which served as protection for the skin and as a means of insulation) consisted mainly of thick, rough hairs and a small amount of down; the total weight of wool grown per year never reached one kilogram (2.2 pounds). The wool of present-day fine-fleeced sheep consists of uniform, thin down fibres; the yearly total weight may reach 20 kilograms (44 pounds).

The behaviour of domesticated animals differs from that of wild animals. Wild animals tend to avoid man and, as a rule, fear him; domestic animals not only live with man and submit to him but, in some cases, actively seek his approval and affection.

The most important consequence, and perhaps the most demonstrative manifestation, of domestication of animals consists of a sharp change in their seasonal biology. The wild ancestors of domesticated animals are characterized, for example, by strict seasonal reproduction and molting rhythms. Most domesticated species, on the contrary, can reproduce themselves at almost any season of the year and molt little or not at all, according to a seasonal pattern.

No less characteristic are the changes that occur in plants as a result of domestication. Their structure and general appearance may be drastically changed. The enormous variety of cultivated cabbages—with respect to form and colour of leaves, size and form of the heads and stalks, and degree of development and form of the flowers—is one of the many examples of the kinds of changes that can occur when a single wild species is domesticated.

Still more demonstrative are the dramatic changes that have occurred in decorative plants, whose qualities of flower form and colour have been developed to great lengths under intensive selection by man. Not so apparent but still significant are the many physiological and biochemical changes that plants undergo with domestication. Some of these changes are most notable in cultivated cereals. Wild and primitive cereal forms, particularly wheat, are characterized by fragility of the spike, which is a natural adaptation for the spread of seeds. Cultivated cereals, however, are characterized by the strength of the spike and the clinging of the grain, without which harvesting would be all but impossible. In all cultivated legumes and oilseed plants, covered and indehiscent (closed) fruits are characteristic, whereas the wild ancestors showed dehiscent fruits, the splitting open of which ensured the dispersal of seeds. Among seed-crop plants generally, large seed size is characteristic, as well as simultaneous germination of seeds and simultaneous ripening of fruits. The chemical composition of domesticated plants is frequently different from the wild ancestors. Primitive sugar beets, for example, contained only 3–4 percent of sugar in their roots, whereas modern varieties contain as much as 17 percent.

Even at the initial stages of domestication, some morphological changes in animals and plants are apparent. In mink, for example, which became objects of breeding for fur in about 1920, there have already appeared more than 20 different variations of fur colour and several variations in fur texture.

The underlying genetic alterations. Genetic analysis has shown that domestic forms develop on the basis of mutations of certain genes existing in a hidden (or recessive) state in the wild species; these genes assert themselves when matings are restricted and when selection by man is exerted on small groups of animals at the very first stages of domestication.

Results of genetic analysis

The elementary genetic mechanism that draws the recessive genes out from the cover of the wild genotype of the natural species also brings about the first domestication-dependent changes and the initial differentiation of a wild species into types that can serve as the basis for superior breed formation. This mechanism is characteristic in both animals and plants. In rye, for example, cultivation has uncovered hidden genes that result in plants that differ from wild rye in seed colour, in having short and strong straw, and in lacking a waxy film on plant parts.

Nature, in effect, has a store of various types and forms hidden as recessive mutations in every natural population of wild animals and plants. It is this accumulated mutation pool that man exploits when he selects and breeds organisms. Such interference by man, called artificial selection, plays a truly creative role in the formation of modern animal breeds and plant varieties to suit the needs of man. In selecting, first without any definite aim and then consciously, the forms that have best met his requirements, man has gradually created, by emphasizing features he has desired and de-emphasizing features he has disliked, a long list of plants and animals he cannot now live without. The material for natural and artificial selection is constantly being replenished in all populations of organisms.

Important in domestication-induced changes is the role played by a special category of chromosomal variability called polyploidy; *i.e.*, the multiplying of chromosome sets. A great portion of cultivated plants are allopolyploids, forms that appeared as the result of hybridization of different species, with subsequent duplication of chromosome number in hybrids; such are wheat, oats, cotton, potato, rutabaga, cabbage, plum, and strawberry.

As indicated above, domestication has sped up the evolutionary process. Important factors in this acceleration are the protection that man afforded his domesticates, thus allowing them to survive, and the specific consequences of artificial selection, which, especially at the first stages of domestication, seems to have been particularly influential on the character of individual development and variability of organisms. In this sense it differs considerably from natural selection, which creates stabilized biological systems that ensure the development of a normal, or so-called wild, phenotype; *i.e.*, an organism containing a wealth of properties that preadapt it to a wide variety of environmental conditions and ensure continuation of the species. Artificial selection breaks down precisely these stabilized systems, thereby creating gene combinations that could not survive in nature and providing a range of new possibilities.

Destabilization phenomena are especially clear in studies of the role of selection for behaviour in domestication. There can be no doubt that taming was one of the very first conditions of animal domestication. Docility and the capacity to be tamed seem to be genetically determined, like other aspects of behaviour; however, they are associated also with levels of regulatory substances called hormones, including those that regulate seasonal activity. A shift in hormonal levels as a result of selection for docile behaviour during the first stages of domestication brought about destabilization and rearrangement of some of the most vital functions of animals.

Since hormones influence all the vital processes in animals and in plants, selection involving the hormonal system cannot help affecting the basic mechanisms of individual development. It is possible that on this basis and by the accumulation of different mutations, selection has operated to provide the necessary raw material for domestication.

BIBLIOGRAPHY. E. ANDERSON, "The Evolution of Domestication," in SOL TAX (ed.), *The Evolution of Man*, pp. 67–84

(1960), a short review of the origin and history of the domestication of wheat, cotton, corn, and other plants; D.K. BELYAEV, "Domestication of Animals," *Sci. J.*, vol. 5 (1969), a paper on the genetical basis and mechanisms of domestications through the author's view of destabilizing selection and its role in the domestication process; CHARLES DARWIN, *The Variation of Animals and Plants Under Domestication*, 2 vol. (1868, reprinted many times), a basic work illustrating the changes animals and plants have undergone under domestication, and the role of selection in this process; E.B. HALE, "Domestication and the Evolution of Behaviour," in E.S.E. HAFEZ (ed.), *The Behaviour of Domestic Animals*, 2nd ed., pp. 22–42 (1969); A. MUNTZING, "Darwin's Views on Variation Under Domestication in the Light of Present-Day Knowledge," *Proc. Am. Phil. Soc.*, 103:190–220 (1959), an account of the role of environment and heredity in the domestication phenomena in view of the achievements of modern genetics; C.A. REED, "Animal Domestication in the Prehistoric Near East," *Science*, 130:1629–1639 (1959); C.O. SAUER, *Agricultural Origins and Dispersals*, 2nd ed. (1969), a moderately difficult book elucidating the first appearances of plant and animal domestication in the Old and New Worlds; N.I. VAVILOV, "The Origin, Variation, Immunity, and Breeding of Cultivated Plants," *Chronica Bot.*, vol. 13 (1949–50), a paper presenting the basis of the author's theory of genecentres of cultivated plants; F.E. ZEUNER, *A History of Domesticated Animals* (1963), a fundamental book treating, in a lively manner, the problems of origin and domestication of main domestic animal species; P.M. ZHUKOVSKY, *Cultivated Plants and Their Wild Relatives* (1962; orig. pub. in Russian, 1950), a difficult and detailed book, the main thrust of which has been retained in the abridged English translation.

(D.K.B./V.V.K.)

Dominic, Saint

St. Dominic, the founder of the Order of Friars Preachers, known also as the Dominicans, introduced in the 13th century a dramatically new form of religious order that had a universal mission of preaching (to pagans as well as to Christians), a centralized organization and government that included democratic elements, and a great emphasis upon scholarship. Dominic's innovative leadership, together with that of St. Francis of Assisi, the founder of the Franciscan Order, greatly influenced the rapid development of the orders of mendicant friars, whose members were not attached to any particular monastery and thus were free to be sent wherever their services were needed; the friars renounced all property through a vow of poverty and depended upon active work, such as preaching and teaching, and charity to subsist.

Domingo de Guzmán was born at Caleruega in Castile, possibly a year or two later than 1170, the traditional date. His father was lord of the manor in the village, and his mother was also from the local nobility. He studied at Palencia and then joined the canons regular (a religious community attached to the cathedral of a diocese) of Osma about 1196 and he became subprior, or assistant to the superior, a few years later. In 1203, Diego, bishop of Osma, was sent on a royal mission abroad and took Dominic with him.

Anderson—Alinari



St. Dominic, panel by the school of Messina(?), 15th century. In the Museo Archeologico Nazionale, Palermo, Italy.

Early life and career

This journey first made Dominic aware of the threat posed to the church in the south of France by the Albigensian heretics, or Cathari, who were reviving and developing the Manichaean teaching that two supreme beings, Good and Evil, dominate spirit and matter respectively, so that whatever concerns the body—eating, drinking, marriage and procreation, and the possession of worldly goods—is essentially evil, and the ideal is the renunciation of these things and even of life itself. Thus, there arose among them a caste of the "perfect," who led a life of great austerity, while ordinary people were regarded as reprobates. A regularized Albigensian hierarchy had come into existence, and local feudal lords, especially the count of Toulouse, supported the Albigenses. Pope Innocent III had launched a mission to preach against the heresy.

On a second journey Dominic and the Bishop visited the Pope, who refused their request to preach to the pagans, so they returned to France. In 1206 the papal legates and preachers, depressed at the failure of their mission, consulted the Bishop and Dominic, who reasoned that the heretics would be regained only by an austerity equal to their own; the preachers must tramp the roads barefoot and in poverty. This was the birth of Dominic's "evangelical preaching." An important part of his campaign was the establishment of a convent of nuns at Prouille, formed in 1206 from a group of women converted from the heresy.

In 1208 the papal legate, Peter de Castelnau, was murdered by an emissary of the Count of Toulouse. The Pope called upon the Christian princes to take up arms. The leader on the papal side was Simon de Montfort, earl of Leicester and a subject of the King of France. The Albigensian leader was Raymond VI, count of Toulouse, an opponent of the King of France and brother-in-law of King John of England, lord of neighbouring Aquitaine. Dominic's work, though confined to the Prouille area, continued, and six others eventually joined him. Meanwhile, the civil war dragged on until Simon's victory at Muret in 1213. The Catholic party entered Toulouse, and Dominic and his friends were welcomed by the bishop, Foulques, and established as "diocesan preachers" in 1215.

Foundation of the Dominicans

From Foulques's charter in that year Dominic's design for an order devoted to preaching developed rapidly. A characteristic concern was for the theological formation of his men, whom he therefore took to lectures given at Toulouse by an Englishman, Alexander Stavensby. Still in 1215, he went to Rome with Foulques (bound for the Fourth Lateran Council) to lay his plans before the Pope, who, however, recommended adoption of the rule of one of the existing orders. It was, perhaps, at this time that Dominic met Francis of Assisi (though the meeting may not have taken place until 1221), and the friendship of the two saints is a strong tradition in both the Franciscan and the Dominican orders. In the summer of 1216 Dominic was back at Toulouse conferring with his companions, now 16 in number. This meeting has been called the *capitulum fundationis* ("chapter, or meeting, of foundation"). The rule of St. Augustine was adopted, as well as a set of *consuetudines* ("customs"), partly based on those of the canons regular, concerning the divine office, monastic life, and religious poverty; these are still the core of Dominican legislation. In July, Innocent III died, and it was from his successor, Honorius III, that Dominic, once more in Rome, finally received on December 22, 1216, formal sanction of his order.

The order was now an established body within the church, and Dominic returned to Toulouse. On August 15, 1217, he sent his men to Paris and to Spain, leaving

two each at Toulouse and Prouille, while he and another went to Bologna and Rome. He placed his two principal houses near the universities of Paris and Bologna and decided that each of his houses should form a school of theology. This at once determined the capital role that the Dominicans would play in university studies. In setting up his houses in the larger cities, especially in those that were teaching centres, he involved his order in the destiny of the medieval urban movement.

Dominic was gifted in being able to conceive his ideal, to form his men to that ideal, and then to trust them completely. His leadership had great clarity of vision (even to the geographical distribution of his forces and precise details of legislation), firmness of command, and certainty of execution. At the same time it was said of him that his gentleness was such that anyone who came to speak to him, even for reproof, went away happier.

The rest of Dominic's life was spent either in Rome, where he was given the Church of S. Sisto, or travelling. In 1218–19 he made a great tour (3,380 miles entirely on foot) from Rome to Toulouse and Spain and back, via Paris and Milan, and in 1220 a tour of Lombardy. Everywhere his communities were growing, and he planned many new foundations covering the key points of France and northern Italy. In Rome the Pope gave him the delicate task of reforming various groups of nuns, whom he finally gathered at S. Sisto in 1221, when the men moved to Sta. Sabina, which is still the residence of the master general of the order.

At Pentecost in 1220 the first general chapter of the order was held at Bologna, and a system of democratic, representative government was devised. At the second general chapter, held on Pentecost in 1221, also at Bologna, the order was divided geographically into provinces. After a visit to Venice, Dominic died at Bologna on August 6, 1221. He was canonized in 1234.

BIBLIOGRAPHY. The fully documented modern biography is that of M.H. VICAIRE, *Histoire de Saint Dominique*, 2 vol. (1957; Eng. trans., *Saint Dominic and His Times*, 1964); the companion volume, *Saint Dominique de Calervega d'après les documents du XIIIe siècle* (1955), provides the documents edited by Vicaire. *Saint Dominic: Biographical Documents*, ed. by F.C. LEHNER (1964), is a translation from 13th-century sources. The best biography in English is BEDE JARRETT, *Life of Saint Dominic, 1170–1221*, 2nd ed. (1934). W.A. HINNEBUSCH, *The History of the Dominican Order*, vol. 1 (1966), provides a good presentation of Dominic's life and the development of the order.

(S.Bh./Ed.)

Dominican Republic

The Dominican Republic (República Dominicana) occupies the eastern two-thirds of Hispaniola, the second largest island of the Greater Antilles archipelago in the Caribbean Sea. Haiti (*q.v.*), also an independent republic, occupies the western third of the island. Hispaniola lies between the islands of Cuba to the west and Puerto Rico to the east and is situated about 600 miles southeast of Florida and 310 miles north of Colombia and Venezuela. The northern shores of the Dominican Republic are washed by the Atlantic Ocean, while the southern shore is bordered by the Caribbean Sea. Between the eastern tip of the island and Puerto Rico runs a channel called the Mona Passage. The republic has an area of 18,704 square miles (48,442 square kilometres; including 46 square miles of adjacent islands) and a population that in 1970 numbered a little more than 4,000,000. The capital is Santo Domingo.

The country, although small and underdeveloped, occupies a strategic position on major sea routes leading from both Europe and the United States to the Panama Canal. Between 1930 and 1961 the republic's history was dominated by the repressive dictatorship of Rafael Trujillo—a ruler who nevertheless maintained internal stability, managed to pay off the national debt, and introduced a measure of prosperity and modernization. Yet, the human costs were excessive; today, large numbers of Dominicans remain in misery, and many feel their situation worsening.

With the rise of Socialism in neighbouring Cuba, political affairs in the Dominican Republic have assumed an even greater international importance (see the city article SANTO DOMINGO; for historical aspects, see DOMINICAN REPUBLIC, HISTORY OF THE).

THE LANDSCAPE

The natural environment. *Relief, drainage, and soils.* Topography is complicated and varied and includes five distinct highland or upland areas running along a northwest to southeast axis. The Pico Duarte in the major Cordillera Central (Central Highlands) is the highest mountain in the West Indies, rising to a height of 10,417 feet (3,175 metres). There is also a smaller range, the Cordillera Septentrional, running parallel to the northwest coast, and two lesser ranges in the southwest. Another minor upland area, the Cordillera Oriental (Eastern Highlands), lies in the northeastern portion of the country. The extreme northwest and the extreme southwest are dry, low, and desertlike. The southeastern region consists of rolling lowlands. Topography

The Río Yaque del Sur empties into the Bahía de Neiba (Bay of Neiba), draining the Cordillera Central to the south, while the Río Yaque del Norte drains the northern slopes, flowing into the Bahía de Montecristi. The eastern part of the island is drained by the Río Yuna, which flows into the Bahía de Samaná, and by the Río Ozama, the mouth of which is near Santo Domingo, on the south coast. The salt lake of Enriquillo, 23 miles long and about 11 miles wide at its widest point, located near the Haitian border in the southwest, is a distinctive landmark.

Soils vary, but those of the upland areas are mostly of residual origin, deriving from metamorphic and sedimentary rocks. Soils in the lowlands are of recent alluvial origin, except in the southeastern savanna (grassland) area, where they consist of sedimentary deposits of recent marine origin. In general all the soils are quite fertile except in the far southwest, in the Pedernales region, where the sedimentary soils are relatively barren.

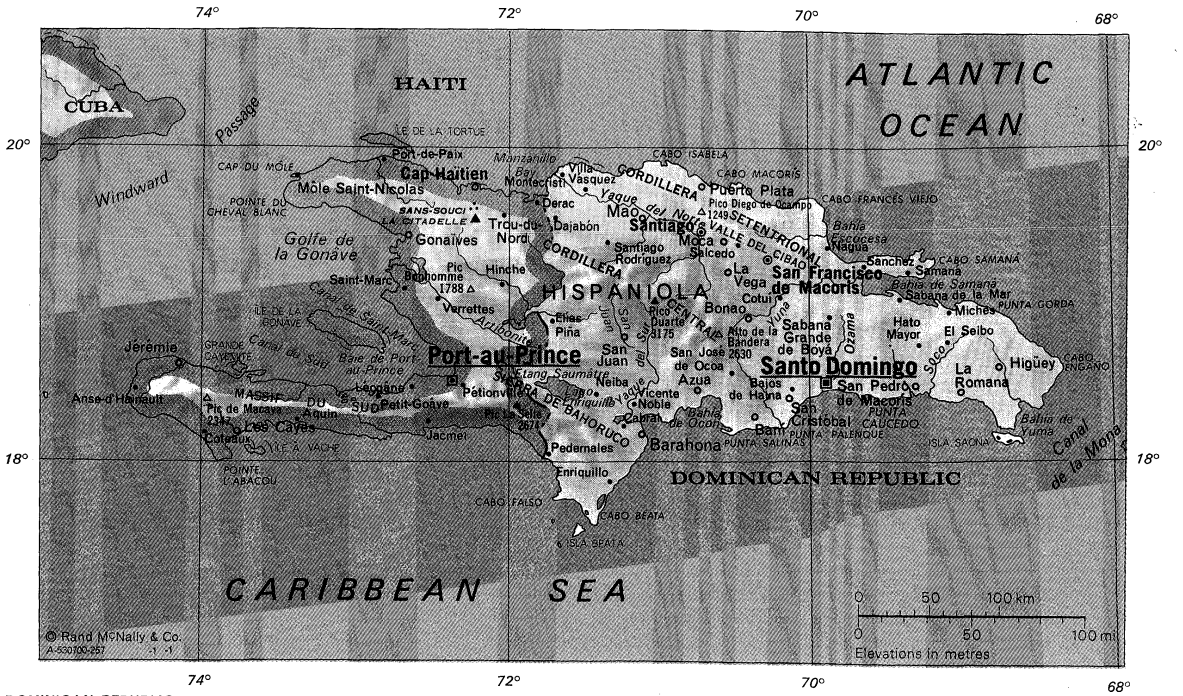
Climate, vegetation, and animal life. The Dominican Republic lies well within the tropic zone, but the hot, moist climate typical of this zone is tempered in many areas by the altitude and in other areas by the insular character of the republic and by the Northeast Trade Winds that blow steadily from the Atlantic all year long. The heaviest precipitation is in the northeast, where the average rainfall is 100 inches a year. As the trade winds pass over the various mountain ranges, they lose their moisture until, in the far western part of the country, along the Haitian border, only a little more than 30 inches of rainfall is received annually. The island is constantly in danger from tropical storms and hurricanes, which originate in the mid-Atlantic from August until October each year.

In spite of local variations, the country as a whole enjoys a relatively mild and pleasant climate. The national mean temperature is 77° F (25° C). Temperatures rarely rise above 90° F (32° C), and, even in the heart of the highest central mountain range, the overall mean is only about 69° F (21° C). Very cold temperatures are literally unknown. Temperatures

Vegetation varies considerably. The mountains are still largely forested with pines and hardwoods, although during the past century the lower and more accessible slopes were practically denuded of trees by commercial lumbering operations. In the drier regions, low shrubs and scrub by trees predominate, but, as rainfall increases—for example on the Samaná Peninsula—grasslands and dense rain forests occur.

Cultivation of a wide variety of crops has largely replaced the natural vegetation in many areas—particularly in the more fertile upland valleys and on the lower mountain slopes. Mangrove swamps line some coastal areas, while elsewhere, particularly along the northern shore, sandy beaches of great beauty are to be found.

Wild-animal life is not abundant, although for several centuries cattle and goats, introduced by the early Spanish colonists, ran wild on the grasslands and in the desert areas. Alligators are found near the mouths of the Yaque del Norte and the Yaque del Sur and in the waters of


74°
72°
70°
68°
20°
18°
HAITI
CUBA
ATLANTIC OCEAN
CARIBBEAN SEA
DOMINICAN REPUBLIC
Passage
Windward
Golfe de la Gonâve
Port-de-Paix
Cap-Haitien
Môle Saint-Nicolas
Gonaïves
Saint-Marc
Hinche
Verrettes
Jérémie
Anse-d'Hainault
Les Cayes
Jacmel
Port-au-Prince
Pétionville
Pic de Macaya
Pic la Selle
Trou-du-Nord
Dajabón
Manzanillo Bay
Montecristi
Villa Vásquez
Derac
Mao
Santiago
Santiago Rodríguez
Monte Cristi
Puerto Plata
Pico Diego de Ocampo
Moca
Salcedo
La Vega
San Francisco de Macorís
Cotuí
Bonao
Pico Duarte
Elías Piña
San Juan
San José de Ocoa
Azua
Neiba
Barahona
Pedernales
Enriquillo
Sabana Grande de Boyá
Bajos de Haina
San Cristóbal
Baní
Santo Domingo
San Pedro de Macorís
La Romana
Higüey
Hato Mayor
El Seibo
Miches
Sánchez
Samaná
Nagua
Cabo Isabela
Cabo Macorís
Cabo Frances Viejo
Cabo Samaná
Cabo Engaño
Punta Caucedo
Isla Saona
Bahía de Yuma
Canal de la Mona
Cordillera Central
Cordillera Septentrional
Valle del Cibao
Sierra de Neiba
Sierra de Bahoruco
Isla Beata
Cabo Falso
Cabo Beata
Punta Salinas
Punta Palenque
Punta Caucedo
Punta Gorda
Isla Vache
Pointe Abacou
Pointe du Cheval Blanc
Cap du Môle
Sans-Souci
La Citadelle
Ile de la Tortue
Ile de la Gonâve
Grande Cayemite
Massif du Sud
Pic de Macaya 2347
Pic la Selle 2674
Pic Chasseur 2680
Pic Pelée 2632
Pic Macaya 2347
Pic Formon 3175
Pic Duarte 3175
Pic Diego de Ocampo 1249
Pic Pomme 1788
Etang Saumâtre
Lac de Péthionville
Lac de Azuéi
Lac de Enriquillo
Lac de Rio Limpio
Lac de Laguna Redonda
Lac de Laguna Salada
Lac de Laguna Limón
Lac de Laguna Cabral
Lac de Laguna Rincón
Lac de Laguna Bávaro
Lac de Laguna Redonda
Lac de Laguna Salada
Lac de Laguna Limón
Lac de Laguna Cabral
Lac de Laguna Rincón
Lac de Laguna Bávaro
Yaque del Norte
Yaque del Sur
Yuna
Ozama
Soco
Nizao
Haina
Nigua
Bahía Escocesa
Bahía de Samaná
Bahía de Ocoa
Bahía de Neiba
Bahía de Yuma

has exhibited one of the highest urbanization rates of any country in the world.

The capital city, Santo Domingo, is an increasingly cosmopolitan city in which manufacturing and other industry vie in importance with commerce, education, and governmental business. The second largest city is Santiago, located inland in the heart of the Cibao.

THE PEOPLE

The Spanish language has always been predominant, although other tongues can also be heard. Various European languages are often spoken; among those who have migrated from Haiti, a French patois is common.

The racial composition is predominantly mestizo (*i.e.*, of mixed blood) with a small number of persons claiming to be of European descent and an even smaller black minority group. The racial categories are roughly associated with economic and social class status, with more blacks at the bottom and more whites at the top. In addition, there is a small colony of German Jews, now beginning to break down as a distinctive unit, and a small group of Japanese in the Constanza Valley, most of whom are engaged in agriculture. During the 19th century, large numbers of persons immigrated from the Mediterranean area, especially from Lebanon; this group has today blended culturally and socially with the remainder of the population.

Although it is generally supposed that the original Taino (Arawak) Indians were annihilated during the early Spanish occupation, it is probable that the present-day Dominican population includes some genetic components deriving from this group, although no identifiable Amerindian ethnic group remains.

Religious groups

Approximately 95 percent of the population is Roman Catholic. The Catholic Church is highly esteemed and exerts a marked influence on cultural life at national, local, and family levels. Approximately 2 percent of the population is Protestant; the remaining 3 percent is Jewish, practices other religions, or professes no religion.

In 1969 the crude birth rate was about 39 per thousand, having fallen from a high of 49 per thousand in 1960. The crude death rate, however, also fell during the same period from about 14 to about 7 per thousand. This gave the country a continuing high rate of natural increase, the effects of which are reflected in a marked preoccupation with emigration, the actual rate of which is unknown because of repeated trips, and in the low standard of living of most of the people.

One of the most significant features of the population is its extremely youthful character. In 1970, about 47 percent of the population was under 15 years of age.

Dominican Republic, Area and Population

	area*		population	
	sq mi	sq km	1960 census	1970 census†
National district				
Distrito Nacional	533	1,380	465,000	817,000
Provinces				
Azua	983	2,547	74,000	92,000
Bahoruco	516	1,338	53,000	67,000
Barahona	661	1,711	80,000	113,000
Dajabón	343	889	42,000	51,000
Duarte	553	1,431	162,000	201,000
El Seibo	1,178	3,052	122,000	133,000
Espaillat	337	872	126,000	140,000
Independencia	775	2,008	28,000	33,000
La Altagracia	1,124	2,911	70,000	87,000
La Estrelleta	601	1,155	44,000	53,000
La Romana	257	665	37,000	57,000
La Vega	1,288	3,337	234,000	294,000
María Trinidad Sánchez	506	1,310	90,000	97,000
Montecristi	768	1,989	60,000	69,000
Pedernales	708	1,833	9,000	13,000
Peravia	628	1,627	108,000	128,000
Puerto Plata	671	1,739	164,000	186,000
Salcedo	168	435	79,000	90,000
Samaná	382	989	43,000	54,000
Sánchez Ramírez	448	1,159	90,000	106,000
San Cristóbal	1,498	3,879	252,000	324,000
San Juan	1,313	3,401	152,000	191,000
San Pedro de Macorís	486	1,259	70,000	105,000
Santiago	1,188	3,077	292,000	386,000
Santiago Rodríguez	487	1,262	41,000	50,000
Valverde	243	629	60,000	77,000
Total Dominican Republic	18,658*‡ 18,704	48,323*‡ 48,442	3,047,000	4,012,000§

*Excludes adjacent islands. †Preliminary. ‡Converted area figures do not add to total given because of rounding. §Figures do not add to total given because of rounding.
Source: Official government figures.

THE NATIONAL ECONOMY

Resources and resource exploitation. *Agriculture.* The economy is based primarily on agriculture. About 90 percent of all foreign exchange comes from the country's farms and pastures. Sugar, tobacco, cacao, coffee, hides, fruits, and tomato paste are the most important exports.

Agricultural products

The country imports most manufactured goods and some foodstuffs. The gross national product (GNP) and the per capita income remain low. Agricultural production constitutes about 22 percent of the GNP, commerce 18 percent, manufacturing 17 percent, government 12 percent, and other sectors 31 percent. In the early 1970s the economy was only beginning to recover from the political upheavals following the fall of Trujillo in 1961 and the civil war of 1965.

Mineral resources. A number of minerals are known to be present but—with the exception of bauxite, gypsum, iron ore, and nickel—have not yet been highly developed commercially. Salt, largely from salt deposits near Lago Enriquillo, is also produced in commercial quantities. A smaller salt-producing enterprise, based on the evaporation of sea water, has also been of some importance at Monte Cristi. Other minerals that might be developed in the future include sulfur, coal, molybdenum, cobalt, tin, oil, and zinc.

The Dominican Republic is one of the relatively few sources of quality amber in the Western Hemisphere. Although it has not yet been extensively exploited, craftsmen produced distinctive jewelry for the local and tourist trade.

Biological resources. Unexploited biological resources are few, but proper management of agricultural lands, forests, and grazing areas could result in far greater productivity. Although the possibilities for a large-scale fishing industry seem limited by the relative scarcity of marketable fish in nearby waters, the potential for the development of tourism and big-game fishing in several of the southern bays was being explored in the early 1970s.

Hydroelectric resources. In order to utilize fully the agricultural lands now available, as well as to open new lands, the irrigation system needs much improvement. Two multipurpose hydroelectric dams were being built in the early 1970s. The Presa (dam) de Tavera will provide electrical power and increase the availability of water for irrigation in the Cibao. The Presa de Valdesia, to be installed on the Río Nizao, will serve the dry southwestern area and provide power for Santo Domingo and its environs.

Management of the economy. *The public and private sectors.* During the Trujillo regime, from 1930 to 1961, the government (which was, in effect, the Trujillo family) largely controlled both agriculture and industry and, thus, the economy. Since 1961, most Trujillo enterprises have remained under governmental control. Two holding corporations were formed to manage them. The first, the Industrial Development Corporation (Corporación de Fomento Industrial), was created in 1962 and handles all the Trujillo industrial holdings other than sugar mills. This was replaced in 1966 by the creation of the Dominican State Enterprise Corporation (Corporación Dominicana de Empresas Industriales, CORDE). There is also the State Sugar Council (Consejo Estatal del Azúcar, CEA), which in 1966 replaced the former Dominican Sugar Corporation (Corporación Azucarera Dominicana). It manages the national sugar mills. In spite of managerial failure and despite a study that suggested they be returned to the private sector, in the early 1970s the government was apparently determined to retain control over these public companies.

Control of the sugar industry

In addition to the government, several foreign com-

panies and one local family dominate the sugar industry. Since 1965 the United States has made loans to increase industrial activity throughout the country. The processing of foods and beverages is especially well developed. Textiles and finished clothing—particularly shoes, shirts, and hats—are beginning to be produced locally and to replace some imports.

Many industries are characterized by their small size. These include, among others, manufacturers of furniture, soap, candles, rope, some food products, and paper clips, as well as building materials, such as concrete blocks and tiles.

The monetary system. The monetary system is based on a managed gold-bullion standard. The peso has been maintained at par with the U. S. dollar since the 1940s. The banking system is well developed, and there are several loan companies and more than 20 insurance agencies, most of the latter foreign owned.

Taxation. In the early 1970s, import duties had, for some years, comprised as much as 40 percent of the entire public revenue. Other government income derives from sales tax, vehicle licenses, and road taxes. Income taxes are levied but are unpopular and relatively unimportant.

Trade unions and employer associations. Trade unionism has not been important historically in the Dominican Republic. During Trujillo's rule, there was one so-called confederation of workers, which, in fact, was no more than a company union. After 1961 the labour movement developed rapidly, with unions tending to organize on the basis of political affiliation. Many affiliated with international labour organizations. As the agricultural sector continues to decline, trade unionism may be expected to become increasingly important.

There are also a number of employer associations, including the U.S. and Dominican chambers of commerce, as well as associations of cattle growers, sugar producers, and tomato growers. A number of factories and industries have associations that include both the workers and management.

The road system

Transportation. Santo Domingo is the hub of a transport system that is generally more than adequate to assure the flow of both people and goods to virtually all parts of the republic. Although certain routes are time consuming and indirect because of the inadequacy of the linkage through the central mountain system, the road network throughout the country is, nevertheless, generally good. Major highways of concrete or asphalt extend in three directions from Santo Domingo, and secondary roads branch out from these major arteries.

Although buses are rare, a system of private taxicabs provides rapid and relatively inexpensive transportation both within and between cities. Most goods are transported by truck to the important market centres.

A government-owned railroad line runs through the eastern half of the Cibao from La Vega to the port of Sánchez on Bahía de Samaná. This carries both passengers and freight, but, with the decline of Sánchez as a commercial port, the railroad has practically ceased functioning.

The only international airport is at Punta Caucedo, about 15 miles east of Santo Domingo. It is capable of handling all types of modern jet craft. A secondary airport in Santiago handles small jets and propeller planes. Fifteen other airfields around the country are open to small civilian craft. The government-owned airline, Compañía Dominicana de Aviación (CDA), operates between Santo Domingo and San Juan, Puerto Rico; Curaçao, New York City, and Miami, and also makes local flights.

Airlines now handle most passenger traffic to and from the Dominican Republic, but goods are exported and imported primarily by sea. The Bahía de Samaná is perhaps the finest and largest natural harbour in the entire Caribbean area. Until the 20th century, the primary commercial ports lay along the northern coast, but, with the rise of the sugar plantations in the south, the ports of Santo Domingo, San Pedro de Macorís and La Romana have increased in importance. Most general goods pass through Santo Domingo, but sugar is largely exported through the ports of San Pedro de Macorís and La Romana. Once important historically, the ports of Monte Cristi and Sánchez in the north are now almost defunct. Only Puerto Plata retains its commercial importance, and it, too, has declined since the 1930s. It is still viable largely because of the tobacco, coffee, and cacao interests in the Cibao region. Barahona ships bauxite, gypsum, and salt but receives few imports.

ADMINISTRATION, SOCIAL CONDITIONS, AND CULTURAL LIFE

The 1966 constitution

The structure of government. *The constitutional framework.* There is a Senate and a Chamber of Representatives. The Senate is composed of one representative from each province, and one from the National District. The Chamber of Representatives reflects the size of the population, but has no fewer than two representatives from each province, and two from the National District. The present (1966) constitution, like its 28 predecessors, guarantees human rights, prescribes the division of governmental powers, and provides for popular sovereignty. It also accords suffrage to all Dominicans of either sex over 18 years of age, unless they are members of the armed forces or the police.

Historically, the various constitutions have provided special emergency powers for the president that have made it possible for the executive to supersede the legislative and judicial branches of the government should the president deem it necessary. While retaining provisions for emergency executive powers, each successive constitution has, nevertheless, expanded the social and economic rights guaranteed in earlier documents. A formal relationship between the Catholic Church and the government, incorporated in earlier constitutions, has now been eliminated. The right to private property is guaranteed but is limited by the right of the state to expropriate for the general good. Terms for national elected offices are four years, and incumbents may seek re-election for an additional term.

Local government. The nation is divided into 26 provinces and one National District (Distrito Nacional). The central government administers the outlying provinces through governors appointed by the president. Each province elects representatives to the bicameral national congress. Internally, each province is subdivided into municipalities that elect their own councils and enjoy considerable local autonomy.

Elections and de facto political developments. Even though recent elections have been run democratically (as has been observed by international representatives), the nature of Dominican social and economic organization imposes conditions inimical to the development of democratic organization. Many less educated persons, for instance, vote as their employers (*patrones*) tell them, having few sources of information upon which to base decisions. Many also fear losing jobs or patronage if they do not follow the wishes of their mentors. Most lower-income Dominicans are convinced that their votes will be identifiable. The bulk of the current adult population has never known freedom in the electoral process, and it may take considerable time before truly representative elections can be expected. On the other hand, the elections of 1966 and 1970 showed that lower-income urban dwellers, especially in Santo Domingo, had developed considerable political consciousness and organization.

The re-election in 1970 of the essentially conservative regime headed by President Joaquín Balaguer, meanwhile, illustrated the power of the business, commercial, industrial, and agricultural establishment. It seemed clear in the 1970s that the support, both economic and political, of the United States and its allies would remain an important consideration in Dominican politics for some time, as has been evident since the military intervention of 1965 to 1966. Even though many liberal Dominicans chafe under the lack of strict autonomy, most seem to believe that economic prosperity can only be achieved through alignments with the United States.

The legal system

Justice. The legal system is based upon the Code Napoléon. There is a series of regular courts, the judges of which are appointed by the Senate and may have no other

public employment. These courts, with the exception of the land and commercial courts, have jurisdiction over both criminal and civil matters.

In criminal cases the judicial process begins with an investigation, generally conducted by an investigating judge. This is followed by the trial proper, conducted by the appropriate court. Both fact and law are taken into account in all cases. Appeals are made to the appropriate superior court and may finally be considered by the Supreme Court, composed of nine justices.

The armed forces. It was not until the United States occupation of 1916 to 1924 that a modern military body was formed in the Dominican Republic. Today, the army, navy, and air force are organized separately, each being headed by its own chief of staff.

During the Trujillo regime, the armed forces were used to preserve the dictatorship, and their members enjoyed a status and power unusual in Latin American states of the 20th century. Even today, it is clear that no government can remain in power long if it does not have the support of at least a portion of the armed forces.

Administrative services. *Education.* The more isolated and rural the population, the less accessible are educational institutions. Primary schooling ordinarily lasts six years, although in rural areas only three may be offered; this is followed by a two-year intermediate school and a four-year secondary course, after which a diploma called the *bachillerato* is awarded. Relatively few lower-income-group students, however, succeed in reaching this level, since the system is designed to encourage middle- and upper-income-group students to prepare for admittance to a university. Most richer students attend private schools, which are frequently sponsored by religious orders. Some public and private vocational education is available, particularly training in the field of agriculture, but this too reaches only a very small percentage of the population.

The universities

There are three universities in the country, two in Santo Domingo and one in Santiago. The Universidad Autónoma de Santo Domingo (UASD) is proud of being the first university in the New World. It is also the largest of the three and is autonomous, being free of both governmental and religious control, although most of its funds are from the national budget. Costs are low and even poor students may attend, if they have been fortunate enough to have secured the requisite primary and secondary preparation.

The Universidad Nacional "Pedro Henríquez Ureña" (UNPHU), located in Santo Domingo, enjoys the support of not only the Catholic Church but also of some private endowments, as well as receiving funds from the national government. The Universidad Católica "Madre e Maestra" (UCMM) in Santiago is similarly supported both by the Catholic Church and through private and public endowments.

The traditional tendency of Latin American university students to engage in political activity has been discouraged in the two newer universities, though with little success. Student political organizations, most of them affiliated with national political parties, are large and active, especially at UASD.

Health. Health conditions among the poorer classes in both rural and urban zones are characterized by a generally unsanitary environment, inadequate health services, and poor nutrition. As a result of these conditions, both infectious and parasitic diseases are rampant, and infant mortality is high.

Hospital and trained medical personnel are available only in the larger cities and towns. In the rural areas, home remedies and the professional services of practitioners of traditional local medicine are the only means of preserving or restoring health. Cases of severe illness may be taken to the nearest urban centre, where hospitalization is free. The tendency, however, is to take this measure only in extreme cases, often when death is already imminent.

Medical care for the more affluent is available through private physicians, many of whom maintain clinics, small nursing homes, and maternity units.

Housing. Housing is considered by Dominican planners to constitute one of the most serious problems in the country. In 1963 it was estimated that more than 200,000 new units—about 145,000 in rural areas alone—were needed to take care of the population at that time. Sixty-five percent of existing residences were considered substandard in 1962, meaning that they were either unsafe or inadequate in size.

Rural housing

On the sugar plantations in the south, barracklike housing is provided for temporary workers, but more permanent employees frequently have their own small huts, or *bohios*, often on company-owned land. These may be little more than a lean-to of palm leaves and bamboo. Others, more sturdy, may have double walls filled with rubble and plastered with mud.

In the Cibao, a relatively prosperous zone, houses are built solidly of palm board or pine and are frequently painted and decorated, with shutters and lintels in contrasting colours. Roofs are most often covered with sheets of zinc or tin but, in poorer households, may be thatched. A prosperous family will have a concrete floor, but most are of packed earth.

In the cities are to be found the squatter settlements and poverty-stricken inner-city ghettos characteristic of most rapidly urbanizing underdeveloped countries of the world. In some cases these may be built of cardboard, tin cans, discarded inner tubes, and any other materials the inhabitants may scavenge. Most houses have a small religious shrine with holy pictures, candles, and other objects on it which often shares a place of honour with a portrait of the current president.

Also in the cities are districts with well-appointed modern houses, occupied by members of the new commercial and industrial elite, as well as by the more traditional land-based oligarchy. Government programs, often funded with international loans, have financed housing construction for lower- and middle-income families.

Police services. The police are frequently associated in the minds of the people with the army and, indeed, have served similar functions. Civil and political offenses frequently merge into one, and arrests for criminal offenses are often politically motivated. Most of the populace, even the middle- and upper-income groups, consequently fear the police, in spite of efforts to improve their image.

Police are poorly paid, but, since most are drawn from the lower-income groups, for them the position represents security and prestige and may well serve to improve their social and economic status. Through graft, which may be considered a systematic mechanism of distribution, their standards of living rise in spite of low wages.

Social conditions. Social conditions in the Dominican Republic resemble those found in other underdeveloped American nations. The small farmers rarely eke out more than a subsistence crop and most often must supplement this by the sale of handicrafts; products include baskets, pottery, rocking chairs, straw hats, and foodstuffs. These items are either sold to middlemen, who market them in towns, or else are displayed and sold along the roads and highways, frequently by children.

Wage rates

Wages and cost of living. Wages in the Dominican Republic tend to be low. In the sugar industry, which is one of the largest employers in the country, wages range from 90 cents per day for unskilled cane cutters to U.S. $5 per day for semiskilled workers in the mills. Skilled laborers, such as bricklayers, may earn up to U.S. $10 a day. White-collar workers with professional degrees, generally paid by the month, may earn U.S. $600 or $700 per month.

The cost of living is high, and in the early 1970s prices were rising steadily.

Economic and social divisions. Although most Dominicans, as well as some social scientists, have insisted that there is no traditional oligarchy, there nevertheless exists an intellectual and economic elite that constitutes less than 5 percent of the total population. Today there also appears to be an emergent middle-income group.

By far the largest proportion of the population belongs to the lower-income group. This category includes small farmers, landless agricultural workers, itinerant mer-

chants, and unskilled manual labourers in domestic and maintenance occupations, as well as in construction, industry, and shipping. This group is underpaid, poor in health, and apathetic concerning its future. Many of them see emigration as the only way out of their poverty.

Cultural life and institutions. *The folk arts.* It is difficult to define any particular and unique cultural tradition that may be labelled Dominican. There are, nevertheless, some cultural items worthy of special mention. Music, especially when accompanied by dancing, is important at all social levels and in all regions. The most typical forms are those with clear African antecedents, especially in their rhythms. There are also folk songs and tunes deriving from Spain and the Middle East. The merengue is a particularly popular dance, followed closely by the bolero. The guitar is probably the most popular instrument, but in some rural areas flutes and homemade marimbas are also found. There is no such thing as an indigenous national costume, although an elaborate fiesta dress reminiscent of Spanish flamenco styles is sometimes worn by wealthier women and proclaimed to be "Dominican."

During the month preceding Lent, a Carnival season is celebrated, during which parades are held in the streets; the somewhat distinctive costumes worn probably derive from medieval Europe.

Recreation. Organized sports are popular, especially baseball. In the more rural areas, cock fighting remains a traditional and popular spectator sport. Soccer is also played but seems to have been eclipsed by baseball in recent years.

Literary clubs are popular, and some of them have remained in existence for nearly a century. There are many fine painters whose work is, although little known outside of the country, well appreciated there.

Press and broadcasting. Bookstores abound in every major city. In the rural zones and among poorly educated urban dwellers, newspapers and paperback novels are popular.

In 1969 there were six daily newspapers, and in 1972 several more were being published, most of them in Santo Domingo.

Radio and, to a lesser extent, television have, in recent years, also contributed to the cultural life of the country. The number of licensed transmitters climbed from 25 in 1950 to more than 80 in 1963 and over 120 in 1971. Most of these were located in Santo Domingo or in Santiago but were readily heard throughout the country.

Prospects for the future. The modern Dominican nation is complex and often difficult for the stranger to understand. It presents many faces, based not only upon regional variations but also upon social and ethnic differences. Though small, it has no single traditional culture, and its people frequently display greater loyalties to other nation-states, either in Europe or in the Americas, than to their own country. For some, the greatest goal in life is emigration; yet, this may be coupled with a fervent patriotism such as that which caused many to lay down their lives in the civil war of the mid-1960s.

The primary obstacles to progress in the early 1970s seemed to lie in the area of social organization, rather than in any lack of natural resources. The close interdependence of the Dominican Republic with larger nations, such as the United States, is a factor that colours nearly every event occurring there.

The long history of despotism under Trujillo has not been forgotten and has clearly made its mark upon the land. Many wish to return to what seemed to them a period of peace and prosperity, even at the cost of freedom and dignity for the individual. On the other hand, it is also clear that, since Trujillo's death, the economic situation has declined. Population pressure is great and becomes more intense as public-health measures reduce mortality. Economic development has so far enabled the rich to become richer but has produced few benefits for the poor. Until now, there has been the safety valve of emigration to the United States. Whether, as in the case of Puerto Rico, this will prove to be a viable temporary solution remains to be seen. Unlike Puerto Ricans, Dominicans are not citizens of the United States nor does the country enjoy other special benefits available to Puerto Rico because of its special political status. On the other hand, it is clear that the Dominican Republic is also economically and socially dependent upon its powerful northern neighbour. So long as leading citizens and governments try to imitate the economic and social structure of the United States, it seems likely that the fate of the masses will not improve. The theory that this will result in the wider distribution of economic benefits has not been justified by results in the Dominican Republic. A new means of economic organization seems to be called for, and whether this can be accomplished under the present political regime remains to be seen.

Relations with the U.S.

BIBLIOGRAPHY. JUAN BOSCH, *Crisis de la democracia de América en la República Dominicana* (1964; Eng. trans., *The Unfinished Experiment: Democracy in the Dominican Republic*, 1965), a view of the Dominican social and political situation by a former president, ousted in 1963; JAMES A. CLARK, *The Church and the Crisis in the Dominican Republic* (1966), a superficial yet singular account of the Catholic Church in the Dominican Republic; ROBERT D. CRASSWELLER, *Trujillo: The Life and Times of a Caribbean Dictator* (1966), the best recent account of its kind; THEODORE DRAPER, *The Dominican Revolt: A Case Study in American Policy* (1968), a well-documented criticism of the North American invasion of 1965; DEBORAH S. HITT *et al.*, *A Selected Bibliography of the Dominican Republic: A Century After the Restoration of Independence* (1968); RAYFORD W. LOGAN, *Haiti and the Dominican Republic* (1967), interesting for the British point of view (but author did not revisit country since 1961)—contains economic summaries; JOHN BARTLOW MARTIN, *Overtaken by Events: The Dominican Crisis from the Fall of Trujillo to the Civil War* (1966), a view of Dominican political events from 1960 to 1965 by an American ambassador; HOWARD J. WIARDA, *The Dominican Republic: Nation in Transition* (1969), a general work on the Dominican Republic by a North American political scientist.

(N.L.G.)

Dominican Republic, History of the

Very little is known of the island of Hispaniola's pre-Columbian history. At the time of Columbus' first landing in 1492, the Caribs, a marauding tribe that had apparently originated on the South American mainland and conquered its way up from the Lesser to the Greater Antilles (and for whom the Caribbean Sea is named), were preying upon the more peaceable Tainos, who had previously settled there. Estimates of the size of the indigenous population range from 300,000 to 3,000,000; other features of pre-Columbian Indian life are similarly shrouded in uncertainty. Part of the uncertainty stems from the fact that the Indians on Hispaniola were primitive and backward; they were not a part of a large-scale and advanced Indian civilization—as in Mexico, Guatemala, or Peru, for example—with a long history and proud tradition of their own.

Hispaniola was discovered by Columbus on his maiden voyage. A colony was established on the north coast, but the first settlers were slaughtered by the Indians. Returning, Columbus established a second colony; but reports of abundant gold farther south quickly led to the abandonment of the northern outpost and to the founding of Santo Domingo city on the Caribbean coast.

Period of Spanish rule

Hispaniola was the first area in the New World to receive the full imprint of Spanish colonial policy. The oldest cathedral, monastery, and hospital in the Americas were established on the island, and the first university chartered. The earliest experiments in Spanish imperial rule were conducted here. Class and caste lines were rigidly drawn; the Roman Catholic Church served as the strong right arm of the temporal authority. A cruel, exploitative, slave-based society and economy came into being. The first "revolution" in the New World was also recorded on Hispaniola.

During the first half-century of Spanish rule, Hispaniola flourished, for its rich mines and lush lands yielded abundant wealth, and it served as the administrative centre for Spain's burgeoning American empire. But the more lucrative conquests of Mexico and Peru soon turned it

into a poor way station. Its Indians were decimated, gold and silver were more easily available elsewhere, and the more ambitious Spaniards emigrated. By 1550 Hispaniola had been all but abandoned.

For the better part of the next three centuries, Hispaniola remained the neglected, poverty-ridden backwater of the Spanish Empire in the Americas. Successive raids by British, Dutch, and French marauders and buccaneers devastated the island still further. Socially and economically, it retrogressed, reverting to a more primitive form of existence. Eventually, French claims to the western third of the island were recognized, and a prosperous, sugar-producing, black-slave-based colony grew up in the area that was later to become the independent nation of Haiti. As a by-product of neighbouring Haiti's prosperity, the Spanish colony also experienced a modest boom in the 18th century.

Slave revolt

In 1795, as a result of the French Revolutionary Wars that had been raging in Europe, Spain also ceded the eastern two-thirds of Hispaniola to France. Meanwhile, inflamed by the revolutionary currents of liberty and equality then sweeping France itself and stirred to revolt by the inhuman conditions under which the slaves were forced to labour, a slave uprising had begun in Haiti. Led by Toussaint-Louverture (*q.v.*), the blacks not only succeeded in throwing off the yoke of French rule but soon overran the previously Spanish eastern end of the island as well, burning the plantations and instilling terror in the white, aristocratic ruling class.

With the aid of the British fleet, however, the Haitians were driven back, and in 1809 the colony was reunited with Spain. Spanish rule was this time short-lived; in 1821, following the lead of the countries on the mainland, the Dominican Republic declared its independence. The new republic comprised approximately the eastern two-thirds of the island.

The Dominican Republic to 1930. Within a matter of weeks, however, Haitian columns under Jean Pierre Boyer (president of Haiti, 1818–43) again overran the entire island. Dominican historians have portrayed this occupation (1822–44) as cruel and barbarous. Haitians held the highest offices, closed the university, severed the church's ties with Rome, forced out the traditional ruling class, and all but obliterated the western European and Hispanic traditions. But Boyer also freed the slaves, and his administration was generally more efficient and his troops less uncontrolled than Dominicans like to admit.

In the 1830s Juan Pablo Duarte—known as the father of Dominican independence—organized a secret society to fight the Haitians, and in 1844, after a long struggle and finally aided by the outbreaks of civil war in Haiti itself, independence was finally achieved. But Duarte and the other young idealists who had led the independence movement were soon exiled, and the new nation quickly fell into other, less noble hands.

19th-century dictators

From 1844 until 1899 the Dominican Republic was dominated by a succession of dictatorial "men on horseback," who served to prevent the growth of a genuine representative democracy, as the constitution proclaimed, and who were not above selling out the country to foreign "protectors" and commercial interests. Pedro Santana and Buenaventura Báez emerged as the two most prominent figures, alternating in the presidency for nearly 30 years and, in the process, almost destroying the infant republic. In order to ward off Haiti's continuous assaults, Santana returned the country to Spain (1861–65) and arranged to have himself named governor general. After a series of battles, Spain withdrew its troops; but the idea of a protectorate remained, and Báez now approached the United States with a plan. President Grant favoured annexation; but after the questionable activities of a United States land-speculating company became public, the Senate failed to ratify the treaty.

During the 1870s the instability continued. Báez returned to the presidency for the fifth time, and the country's first, but short-lived, democratic government came to power. The instability culminated in the emergence of a new strong man, Ulises Heureaux, who dominated the country from 1882 to 1899. Heureaux ruled as a dictator, but he presided over a period of unprecedented stability and national growth. New roads were built, communications improved, production rose, foreign capital entered, population increased. The Dominican Republic began its process of modernization.

Following Heureaux's assassination in 1899 the country returned to the chaotic, unstable politics of the past. New "men on horseback" galloped in and were in turn forced out of the presidential palace with frequent regularity. For a brief time, under Ramón Cáceres (1906–11), the country seemed to be recovering; but he too was assassinated, and political chaos ensued once more. Even the accession of the archbishop Adolfo Nouel to the presidency in 1912 failed to stem the disorder; within four months he, too, was forced to resign and civil war broke out again.

United States influence

Meanwhile, the deteriorating financial situation of the country, plus the expanding interests of the United States in the Caribbean area, had drawn its large North American neighbour ever deeper into Dominican affairs. By this point, the United States had replaced Europe as the major importer of Dominican products, as well as its chief supplier; United States private investments were also rising rapidly. In 1905, threatened with the possibility that European creditors would use force to collect some unpaid Dominican debts, the United States took over the administration of the Dominican customs revenues. Over the next decade the United States presence increased; and in 1916, as the fragile political situation collapsed again, the United States assumed complete control.

The United States occupation (1916–24) saw the building of some new roads, schools, and communications and sanitation facilities, and there were other constructive projects. But the occupation forces had assumed arbitrary control and frequently abused their authority. In addition, the creation by the United States Marines of a modern, unified, military constabulary provided the instrument by which some future Dominican strong man could seize power and use it to dominate the country. That eventuality was not long in coming.

In 1924, in United States-supervised elections, Horacio Vásquez was elected president. His rule eventually proved incompetent and corrupt, and the 1929–30 market crash undermined the Dominican economy. In 1930 a revolution was launched against his rule. The military, now under the firm control of Rafael Trujillo, held the balance of power; but rather than defending the government, it stood by and let the revolution succeed. Then Trujillo moved to take power himself.

Trujillo's dictatorship

The Dominican Republic since 1930. The dictatorship of Rafael Trujillo (1930–61) was one of the longest, cruellest, most absolute dictatorships the world has ever seen. For over three decades Trujillo ruled his country with an iron hand; virtually everything in the country he touched had to belong to him. Trujillo dominated the armed forces, the government, the economy, the church, education, intellectual life, sports, everything. Under Trujillo, however, as under Heureaux, the economy prospered, social change was accelerated, and the Dominican Republic in 1961 was a far more complex and more "modern" society than it had been in 1930.

Following Trujillo's assassination in 1961, his heirs and followers attempted to hang on to the reins of power. But they were also driven out, and the country embarked on a more democratic course. In 1963 Juan Bosch took office, establishing the first democratically elected and social-reformist government in the country's history. Conservative forces remained strong, however, and after seven hectic months Bosch was overthrown. As the country returned to conservative, do-nothing rule, the frustrations and pressures began building up; in 1965 the Dominican Republic exploded in a popularly based and democratic social revolution. Fearing a second Cuba, however, the United States again occupied the country militarily and snuffed out the revolution.

After a stormy interim, new elections were held in 1966. The winner was Joaquín Balaguer, a former Tru-

jillo puppet now presenting himself as a moderate and a symbol of orderly change. Under Balaguer some economic gains were registered and a number of social reforms partially implemented, but the level of tension remained high and the problems brought dramatically to the surface in the 1965 revolution continued unresolved. In 1970 Balaguer was elected to a second term. An attempted right-wing coup was crushed in 1971.

BIBLIOGRAPHY. JUAN BOSCH, *Crisis de la democracia de América en la República Dominicana* (1964; Eng. trans., *The Unfinished Experiment: Democracy in the Dominican Republic*, 1965), a perceptive analysis by one of the country's most important and controversial figures; ROBERT D. CRASSWELLER, *Trujillo: The Life and Times of a Caribbean Dictator* (1966), a superb biography of the late Generalissimo; JOHN BARTLOW MARTIN, *Overtaken by Events: The Dominican Crisis from the Fall of Trujillo to the Civil War* (1966), a fascinating account by the former U.S. ambassador to the Dominican Republic; SELDEN RODMAN, *Quisqueya: A History of the Dominican Republic* (1964), a brief, well-written history; JEROME SLATER, *Intervention and Negotiation: The United States and the Dominican Revolution* (1970), the best account of the revolution and U.S. intervention; HOWARD J. WIARDA, *The Dominican Republic: Nation in Transition* (1969), a general overview focussing on the problems of national development.

(H.J.Wi.)

Domitian

Domitian (Titus Flavius Domitianus), Roman emperor AD 81–96, is chiefly remembered for the reign of terror under which prominent members of the Senate lived during his last years.

Alinari



Domitian, marble bust In the Palazzo dei Conservatori, Rome.

Early years

He was born October 24, AD 51, the second son of the future emperor Vespasian and Flavia Domitilla. During the civil war of 69 over the imperial crown, Domitian remained unharmed in Rome, but on December 18 he took refuge in the Capitol with his uncle Flavius Sabinus, escaping into hiding when the Capitol was stormed by supporters of Vitellius. On the entry of his father's supporters into Rome two days later he was saluted as Caesar, and he became praetor next year. He attempted to turn the repressive campaign of Petillius Cerialis in the Rhineland into a triumphal operation of his own; and for this and other excesses he is said to have required his father's pardon when the latter arrived at Rome in autumn AD 70. Domitian, however, was *princeps juventutis* (an imperial prince) and was consul six times in Vespasian's lifetime; moreover, it was recognized that he would eventually succeed his brother Titus, who had no son and was 11 years Domitian's senior.

On Vespasian's death Domitian expected the same position as Titus had received under Vespasian, in particular, tribunician power and some form of *imperium*. These were not granted, and Domitian was evidently antagonistic to his brother and is alleged to have hastened his death, which occurred on September 13, 81.

As emperor, Domitian was hated by the aristocracy. From the Trajanic writers Tacitus and Pliny the Younger (Suetonius is less partisan) it is hard to disentangle stock vituperation from genuine belief, but it seems certain that cruelty and ostentation were the chief grounds of his unpopularity rather than any military or administrative incompetence. Indeed, his strict control over magistrates in Rome and the provinces won Suetonius' praise. In his secretariat he used both freedmen and knights, some of whom retained their posts after his death; and his *consilium* of close advisers, including senators, involved no departure from precedent. In legislation he was severe, and he incurred censure for attempting to curb vices from which he himself was not immune. A fairer criticism might be for undue paternalism. An edict ordaining destruction of half the provincial vineyards was typical: it was designed to encourage corn growing and to limit the importing of wine into Italy (where, meanwhile, no increased output was permitted), but Domitian was unable to carry the matter through.

His military and foreign policy was not uniformly successful. Both in Britain and in Germany advances were made early in the reign, and the construction of the Rhine–Danube *limes* ("fortified line") owes more to Domitian than to any other emperor. But consolidation in Scotland was halted by serious wars on the Danube, where Domitian never achieved an entirely satisfactory settlement and, worse still, lost two legions and many other troops. This, though admitted even by Tacitus to be due to the slackness or rashness of his commanders, was naturally held against Domitian at Rome. But it did not affect his popularity with the army, whose pay he had wisely raised by one-third in AD 84.

The real issue was his own constitutional and ceremonial position. He continued his father's policy of holding frequent consulates (he was consul *ordinarius* every year from 82 to 88): he became censor for life in 85, with consequent control over senatorial membership and general behaviour; he wore triumphal dress in the Senate; and he presided, wearing Greek dress and a golden crown, over four yearly games on the Greek model, with his fellow judges wearing crowns bearing his own effigy among effigies of the gods. A grave source of offense was his insistence on being addressed as *dominus et deus* ("master and god").

The execution of his cousin Flavius Sabinus in 84 was an isolated event, but there are hints of more general trouble about 87. The crisis came with the revolt of Antonius Saturninus, governor of Upper Germany, on January 1, 89. This was suppressed by the Lower German Army, but a number of executions followed, and the law of *majestas* (treason) was later employed freely against senators. The years 93–96 were regarded as a period of terror hitherto unsurpassed.

Opposition to Domitian

Among Domitian's opponents was a group of doctrinaire senators, friends of Tacitus and Pliny and headed by the younger Helvidius Priscus, whose father of the same name had been executed by Vespasian. Their Stoic views were probably the cause of his expulsions of "philosophers" from Rome on two occasions. It is unlikely that he was carrying out any considered policy on the organization of thought any more than that his condemnation in 95 of his cousin Flavius Clemens for *atheotēs* ("irreligion") was part of a deliberate attack on Christianity.

Domitian's financial difficulties are a vexed question. Cruelty came earlier in his reign than rapacity, but eventually he regularly confiscated the property of his victims. His building program had been heavy: Rome received a new forum (later called Forum Nervae) and many other works. Then there were Domitian's new house on the Palatine and his vast villa on the Alban Mount. Meanwhile, the increased army pay was a recurrent cost. Probably only his confiscations averted bankruptcy in the last years.

The conspiracy that caused his murder on September 18, 96, was led by the two praetorian prefects, various palace officials, and the Emperor's wife Domitia Longina (daughter of Gnaeus Domitius Corbulo). Nerva, who took over the government at once, must clearly have been privy. The Senate was overjoyed at Domitian's

death, but the army took it badly; and next year they insisted on the punishment of those responsible.

BIBLIOGRAPHY. M.P. CHARLESWORTH and RONALD SYME in *The Cambridge Ancient History*, vol. 11, ch. 1 and 4 (1936, reprinted 1954), with a full conspectus of ancient sources; EDMUND GROAG and ARTHUR STEIN, *Prosopographia imperii Romani*, 2nd ed., pp. 147–150 (1943); STEPHANE GSELL, *Essai sur le règne de l'empereur Domitien* (1894); J. JANSSEN, edition of Suetonius, *Domitian* (1919); TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome*, vol. 5, esp. pp. 55–56 and 141–142 (1940); R.L.P. MILBURN, "The Persecution of Domitian," *Church Quarterly Review*, 139:154–164 (1945).

(G.E.F.C.)

Donatello

The greatest sculptor of the 15th century and one of the towering figures of Italian Renaissance art, Donatello was among the founders of the new style—ranking second only to the architect Filippo Brunelleschi, who was his senior by a decade and to whom he was linked by friendship during the first half of his long career. Donatello's statues embody the new image of man conceived by the Italian Humanists, the creators of Renaissance scholarship and philosophy. His influence pervaded both sculpture and painting throughout the 15th century and beyond.

Donatello's life and artistic career are known in considerable detail. Unfortunately, however, there is little information on his character and personality, and what is known is not wholly reliable. Some anecdotes of the late 15th century, for instance, reflect his reputation as a "lover of beautiful apprentices." He never married, and he seems to have been a man of simple tastes in his daily habits. Patrons often found him difficult to deal with; apparently he demanded a measure of artistic freedom beyond what most of them were accustomed to grant at a time when artists' conditions of work were still regulated by guild rules. Although on intimate terms with a number of Humanists, Donatello himself was not a man of literary culture. No letters by him exist or are referred to in the correspondence of those who knew him. Yet the inscriptions and signatures on his works are masterly in design and among the earliest examples of the revival of classical Roman lettering. That he was a connoisseur of ancient art is attested to by his Humanist friends. In fact, 20-century research has shown that he had a more detailed and wide-ranging knowledge of ancient sculpture than any other artist of his day, and that his own work was inspired by ancient visual examples (which he often daringly transformed) to a far greater extent than had hitherto been assumed. The traditional view of Donatello as essentially a realist, therefore, has been modified in the light of these findings.

Anderson—Alinari



Equestrian statue of Gattamelata, bronze sculpture by Donatello, 1447–53. In the Piazza del Santo, Padova, Italy. Height about 3.35 m.

Early career in Florence. Donatello (diminutive of Donato) was born about 1386 the son of Niccolò de Betto di Bardo, a Florentine wool carder. Nothing is known about the beginnings of his career, but it seems likely that he learned the craft of stone carving from one of the sculptors working for the cathedral of Florence around 1400. At some time between 1404 and 1407, he was a member of the workshop of Lorenzo Ghiberti, a sculptor in bronze who in 1402 had won the public competition for the doors of the Florentine baptistery. Donatello's earliest work, of which there is certain knowledge, a marble statue of David, betrays an important artistic debt to Ghiberti, who was then the leading exponent in Florence of the so-called International Gothic, a style of graceful, softly curved lines strongly influenced by northern European art. The "David," originally intended for the cathedral, was transferred in 1416 to the Palazzo Vecchio, the city hall of Florence, where it stood for many years as a civic–patriotic symbol, although from the 16th century on it was eclipsed by the gigantic "David" of Michelangelo, which served the same purpose. Other early works by Donatello, still partly Gothic in style, are the impressive seated marble figure of St. John the Evangelist for the facade of the cathedral and a wooden crucifix in the church of Sta. Croce. The latter, according to a persistently repeated though unproved anecdote, was made in friendly competition with Brunelleschi, who was a sculptor as well as an architect.

Association with Ghiberti

The full power of Donatello's genius was first seen in the two marble statues "St. Mark" and "St. George" (both completed about 1415) for niches on the exterior of Or San Michele, the church of Florentine guilds ("St. George" has been replaced by a copy; the original is now in the Bargello). Here, for the first time since classical antiquity and in striking contrast to medieval art, the human body is rendered as a self-activating, functional organism, and the human personality is endowed with a confidence in its own individual worth. These figures can stand on their own legs, physically as well as spiritually. The same qualities came increasingly to the fore in the series of prophet statues that Donatello did beginning in 1416 for the niches of the campanile, the bell tower of the cathedral (all these figures, together with their companions by lesser masters, have been removed to the Museo dell'Opera del Duomo). According to the payment records, Donatello executed five statues altogether: a beardless and a bearded prophet as well as the group of Abraham and Isaac (1416–21) for the niches on the east side, and the so-called "Zuccone" ("pumpkin," because of its bald head) and the so-called "Jeremiah" (actually Habakkuk) for those on the west side. The "Zuccone" well deserves its fame as the finest of the campanile statues and one of the artist's masterpieces. Both it and the "Jeremiah" (1427–35) show highly individual features inspired by ancient Roman portrait busts. Their entire appearance, which suggests classical orators of singular expressive force, has so little in common with the traditional image of Old Testament prophets that by the end of the 15th century they could be mistaken for portrait statues even though they clearly had not been so intended.

First appearance of a Renaissance style

In his narrative relief panels for the north door of the baptistery, Ghiberti had begun to extend the apparent depth of the scene by placing the boldly rounded foreground figures against more delicately modelled settings of landscape and architecture. Taking this pictorial tendency as his point of departure, Donatello in his marble panel "St. George Killing the Dragon" (1416–17, base of the St. George niche at Or San Michele) invented a bold new mode of relief known as *schiacciato* ("flattened out"). Its actual depth of carving was extremely shallow throughout, yet for that very reason the effect of deep atmospheric space was far more striking than before, since the sculptor no longer modelled his shapes in the usual way but rather seemed to "paint" them with his chisel. Tonal values (or shades) were created through subtle modulations of the relief surface that were meant to control the angle at which the light was reflected by the carved forms. A blind man could "read" a Ghiberti relief with his fingertips; a *schiacciato* panel, on the con-

Invention of *schiacciato*

trary, must be seen, for it depends on visual rather than tactile perceptions.

After the pioneer effort of the St. George relief, Donatello continued to explore the possibilities of the new technique in his marble reliefs of the 1420s and early 1430s. The most highly developed of these are "The Ascension, with Christ Giving the Keys to St. Peter," which is so delicately carved that its full beauty can be seen only in a strongly raking light, and the "Feast of Herod" (1433–35; Musée des Beaux-Arts, Lille, France) with its perspective background. The large stucco roundels with scenes from the life of St. John the Evangelist, below the dome of the old sacristy of S. Lorenzo, Florence, show the same technique, but with the addition of colour for better legibility at a distance (about 1434–37).

In the meantime, Donatello had also become a major sculptor in bronze. His earliest work in that material was the over life-size statue of St. Louis of Toulouse, executed about 1423 for a niche at Or San Michele (replaced half a century later by Verrocchio's bronze group of Christ and the doubting Thomas). Toward 1460 the "St. Louis" was transferred to Sta. Croce; and it is now in the museum attached to that church. Contrary to the unfavourable opinion of older scholars, the "St. Louis" is acknowledged today as an achievement of the first rank, artistically as well as technically. Even though the garments completely hide the body of the figure, Donatello succeeded in conveying the impression of harmonious organic structure beneath the drapery. Unlike Ghiberti, Donatello never maintained a bronze foundry of his own but used the casting facilities of others (usually bell founders). This, however, was a matter of convenience rather than evidence of lack of technical competency and does not justify the view that the master's bronzes were the result of collaboration with other sculptors better versed in the requirements of the medium.

If Donatello needed a collaborator at the time he was engaged on the "St. Louis," it was probably a specialist in decorative architecture, since he had been commissioned to do not only the statue but also the niche and its framework. This niche is the earliest of its kind to display the new Renaissance architectural style created by Brunelleschi without residual Gothic forms. It could hardly have been designed by Donatello alone; he may have been assisted by Michelozzo, a sculptor and architect with whom he entered a limited partnership one or two years later. To the joint enterprises of the two masters, Donatello contributed only the sculptural centre: the fine bronze effigy on the tomb of the schismatic pope John XXIII in the baptistery, the relief of the "Assumption of the Virgin" on the Brancacci tomb in S. Angelo a Nilo, Naples, and the balustrade reliefs of dancing angels on the outdoor pulpit of Prato Cathedral (1433–38). Michelozzo was responsible for the architectural framework and the decorative sculpture, although the actual carving was often left to workshop assistants. The architecture of these partnership projects follows the lead of Brunelleschi and differs sharply from that of comparable works undertaken by Donatello alone in the 1430s, all of which show an unorthodox ornamental vocabulary drawn from both classical and medieval sources as well as an un-Brunelleschian tendency to blur the distinction between the sculptural and architectural elements. The graceful and lyrical figures of the Annunciation tabernacle in Sta. Croce are set against a background of the same ornament that occurs on the framework, and there is a frieze of ecstatically dancing angels on the singers' pulpit, or "Cantoria," in the Duomo (now in the Museo dell'Opera del Duomo). Both of these works show a vastly increased repertory of forms derived from ancient art, the harvest of Donatello's long sojourn in Rome between 1430 and 1433. The master must have returned to Florence with countless sketches after ancient monuments and with a new confidence in his ability to design the architectural framework for his own sculpture. His departure from the standards of Brunelleschi in such works as the "Cantoria" produced an estrangement between the two old friends that was never repaired. Brunelleschi even composed epigrams against Donatello.

Partnership with Michelozzo

During his partnership with Michelozzo, Donatello carried out a number of independent commissions of a purely sculptural character, including several works in bronze for the baptismal font of S. Giovanni in Siena. The earliest, and most important, of these was the "Feast of Herod" (1423–27), an intensely dramatic relief with an architectural background that for the first time displayed Donatello's command of scientific linear perspective; this mathematically precise system of foreshortening and space projection had been invented only a few years earlier by Brunelleschi. To the Siena font Donatello also contributed two statuettes of Virtues, austerely beautiful female figures whose style points toward the Virgin and angel of the Sta. Croce Annunciation, and three nude putti or child angels (one of which was stolen and is now in the Berlin museum). These putti, evidently influenced by Etruscan bronze figurines, prepared the way for the bronze David, the first large-scale, free-standing nude statue of the Renaissance. Well proportioned and superbly poised, it was conceived independently of any architectural setting. Its air of harmonious calm makes it the most classical of Donatello's works. The statue was surely done for a private patron, but his identity remains in doubt. Its recorded history begins with the wedding of Lorenzo the Magnificent in 1469, when it occupied the centre of the courtyard of the Medici palace in Florence. It could not have been made for this location, however, since the palace was not begun until 1444, during Donatello's absence in Padua. After the expulsion of the Medici in 1495, the bronze "David" was placed in the courtyard of Palazzo Vecchio.

The first free-standing nude statue since antiquity

Whether or not the bronze "David" was commissioned by the Medici, Donatello did work for them between 1433 and 1443. This was the decade during which he produced sculptural decoration for the Old Sacristy in S. Lorenzo, the Medici church. His works there included ten large reliefs in coloured stucco and two sets of small bronze doors, which show paired saints and apostles disputing with each other in extraordinarily vivid and even violent fashion.

Paduan period. In 1443 Donatello was about to start work on two much more ambitious pairs of bronze doors for the sacristies of the cathedral when he was lured to Padua (modern Padova) by the commission for a bronze equestrian statue of a famous Venetian condottiere (mercenary soldier), Erasmo da Narni—popularly called "Gattamelata" ("the honeyed cat")—the commander of the Venetian army who had died shortly before. Such a project was unprecedented—indeed, scandalous—at that time. Since the days of the Roman Empire, bronze equestrian monuments had been the prerogative of rulers. To extend the same privilege to a mere professional military leader, however meritorious, must have seemed the height of presumption. In any event, the execution of the monument was plagued by delays of many kinds. Donatello did most of the work between 1447 and 1450, yet the statue was not placed on its pedestal until late in 1453, ten years after the master had left Florence. It represents Gattamelata, in pseudoclassical armour, calmly astride his mount, the baton of command in his raised right hand. The head is an idealized portrait with intellectual power and Roman nobility. This statue was the ancestor of all the equestrian monuments erected since the mid-15th century. Its fame, enhanced by the controversy surrounding it, almost at once spread far and wide. Even before it was put on public view, the King of Naples wanted to secure the services of Donatello for a royal equestrian statue of the same kind, a telling commentary on the importance attached to artistic achievements in Renaissance Italy.

The "Gattamelata" and its influence

While he was awaiting the resolution of the difficulties surrounding the Gattamelata monument, Donatello undertook some important works for the Paduan church of S. Antonio: a splendidly expressive bronze crucifix and a new high altar, the most ambitious of its kind, unequalled anywhere in 15th-century Europe. It consisted of a richly decorated architectural framework of marble and limestone containing seven life-size bronze statues, 21 bronze reliefs of various sizes, and a large limestone

relief of the "Entombment of Christ." Unfortunately, the housing was destroyed a century later, and the present arrangement of the sculpture, dating from 1895, is wrong both aesthetically and historically. Recent attempts to reconstruct the original appearance of the altar agree on certain basic features but vary widely in detail. Among the statues, the majestic Madonna and the delicate, sensitive St. Francis command particular notice; the Madonna's austere frontal pose would seem to be a conscious reference to a much earlier venerated image of the Virgin that then existed in the same church. Of the reliefs, the finest are the four miracles of St. Anthony, wonderfully rhythmic compositions of great narrative power. Donatello's mastery in handling large numbers of figures (one relief has more than 100) anticipates the compositional principles of the High Renaissance.

The last three years of Donatello's Paduan sojourn are puzzling because of the master's apparent inactivity. He had delivered the sculptures of the high altar of S. Antonio in mid-1450, apparently in considerable haste (some of the statues are technically unfinished) and without a financial settlement; six years later, he was still trying to collect money he claimed the authorities owed him for the altar. The Gattamelata monument was finished and waiting to be placed on its pedestal. The master had dismissed the large force of sculptors and stone carvers he had employed on these major projects. Offers of other commissions reached him from Mantua, Modena, and Ferrara, perhaps even from Naples, but nothing came of any of them. Clearly, Donatello was passing through a crisis that prevented him from working. In part, his troubles seem to have been physical; an acquaintance quoted him some years later as having said toward the end of his Paduan stay that he did not want to die "among those frogs in Padua, which he almost did," and in 1456 the Florentine physician Giovanni Chellini recorded in his account book that he had successfully treated the master for a protracted illness. Poor health may in turn have set off an emotional crisis. Such, surely, is the impression conveyed by the only two works Donatello completed between 1450 and 1455: the wooden statue "St. John the Baptist" in Sta. Maria dei Frari, Venice, dating shortly before his return to Florence, and the even more extraordinary wooden figure of Mary Magdalen in the Florentine baptistery. Both are marked by a new depth of insight into psychological reality. The powerful bodies found in Donatello's previous work have become withered and spidery, overwhelmed, as it were, by tremendous emotional tensions within. The "Magdalen" was damaged during the flood that inundated Florence in 1966. Subsequent restoration revealed the original painted surface, which includes realistic flesh tones and golden highlights on every strand of the saint's hair.

Physical and emotional crisis

Late Florentine period. The unbridled expressive force of the "Magdalen" must have been a shock to the Florentines. During the decade of Donatello's absence, a new generation of sculptors had arisen—the most talented were Antonio Rossellino and Desiderio da Settignano—that excelled in the sensuous treatment of marble surfaces. Their work, and the public's taste for it, had little in common with the savage intensity of Donatello's late work. Between 1455 and 1460, therefore, Donatello's important commissions all came from outside Florence. They included the dramatic bronze group "Judith and Holofernes" (later acquired by the Medici and now standing in front of the Palazzo Vecchio) and a bronze statue of St. John the Baptist for Siena Cathedral. In the late 1450s, Donatello was living in Siena, where he was at work on a pair of bronze doors for the cathedral. This ambitious project, which would have rivalled Ghiberti's doors for the Florentine baptistery had it been carried to completion, was abandoned by the master about 1460 for reasons unknown (most likely technical or financial). Only two of the reliefs were executed; one of them is probably the "Lamentation" panel now in the Victoria and Albert Museum, London.

Donatello spent the last years of his life in Florence designing twin bronze pulpits for S. Lorenzo, so that at the time of his death on December 13, 1466, he was once again in the service of his old patrons, the Medici. Covered with reliefs showing the passion of Christ, the pulpits are works of tremendous spiritual depth and complexity, even though some parts were left unfinished and had to be completed by lesser artists.

MAJOR WORKS

All the following are sculpture, in marble unless otherwise stated. "David" (1408–09, 1416; Bargello, Florence); "St. John the Evangelist" (1408–15; Museo dell'Opera del Duomo, Florence); "Crucifix" (wood; *c.* 1410–15; Sta. Croce, Florence); "St. Mark" (1411–*c.* 1415; Or San Michele, Florence); "St. George Tabernacle" (*c.* 1415–17; Bargello, Florence); "Pazzi Madonna" (marble relief; *c.* 1422; Staatliche Museen Preussischer Kulturbesitz, Berlin); "St. Louis Tabernacle" (the statue in bronze; 1423; Or San Michele, Florence); "Lo Zuccone" (1423–25; Museo dell'Opera del Duomo, Florence); sculpture for the font, Siena baptistery (bronze; 1423–29; baptistery, Siena); Tomb of John XXIII (the effigy in bronze; 1425–27; baptistery, Florence); "The Ascension with Christ Giving the Keys to St. Peter" (1428–30; Victoria and Albert Museum, London); "David" (bronze; *c.* 1430–35; Bargello, Florence); "Annunciation Tabernacle" (limestone; 1428–33; Sta. Croce, Florence); "Cantoria" (1433–39; Museo dell'Opera del Duomo, Florence); sculptural decoration of the Old Sacristy (stucco, bronze; S. Lorenzo, Florence); "Crucifix" (bronze; 1444–47; S. Antonio, Padova); equestrian monument of Erasmo da Narni called "Gattamelata" (bronze; 1447–53; Piazza del Santo, Padova); high altar, S. Antonio, Padova (bronze, stone; 1446–50; S. Antonio, Padova); "Mary Magdalen" (wood; 1454–55; baptistery, Florence); "Judith and Holofernes" (bronze; 1456–57; Palazzo Vecchio, Florence); "Lamentation" (bronze relief; 1458–59; Victoria and Albert Museum, London); twin pulpits (bronze; *c.* 1460–70; S. Lorenzo, Florence).

BIBLIOGRAPHY. H.W. JANSON, *The Sculpture of Donatello*, 2 vol. (1957; one-volume ed., 1963), an exhaustive critical catalog, with full digest of documents, sources, and scholarly literature up to 1955; JOHN POPE-HENNESSY, *Italian Renaissance Sculpture*, 2nd ed. (1971); and CHARLES SEYMOUR, *Sculpture in Italy, 1400 to 1500* (1966), stimulating discussions of Donatello in the context of the art of his time; *Donatello e il suo tempo, Atti dell'VIII Convegno Internazionale di studi sul Rinascimento* (1968), an important collection of scholarly papers read at an international congress in Florence in 1966 to commemorate the 500th anniversary of Donatello's death; MANFRED WUNDRAM, *Donatello und Nanni di Banco* (1969), a controversial but fruitful analysis of Donatello's early work.

(H.W.J.)

Donizetti, Gaetano

One of the most prolific of 19th-century Italian opera composers, Domenico Gaetano Maria Donizetti played a significant role in the progress of operatic history; his works represent a transitional stage between Rossini and Verdi's further development of Italian opera. His works are both valuable in their own right and interesting for their influence on his younger contemporary, Verdi. In his serious operas he developed considerably the dramatic weight and emotional content of the genre, and his comic operas have a sparkling wit and gaiety all their own.

The youngest of three sons of the caretaker of the *monte di pieta* (the municipal pawnshop), Donizetti was born on November 29, 1797, at Bergamo in Lombardy, Italy. He began his musical studies with Giovanni Simone Mayr, a Bavarian priest who was musical director of Sta. Maria Maggiore, Bergamo's chief church, and also a successful composer of opera. As a choirboy Donizetti did not shine, but Mayr perceived in him a nascent musical ability and secured his entry into the Liceo Filarmonico (the music school) at Bologna, where he had a thorough training in fugue and counterpoint. His father hoped he would become a church composer, but, though he did compose a vast quantity of sacred music, his natural instinct was for the theatre.

Donizetti scored his first success with *Enrico di Borgogna*, which was premiered in 1818 at the Teatro San Luca, in Venice, and during the next 12 years he composed no fewer than 31 operas, most of them produced at Naples and now forgotten. In 1830 his *Anna Bolena*, produced in Milan, carried his fame abroad to all the European capitals and eventually across the Atlantic.

International success



Donizetti, portrait by Giovanni Carnevali, *c.* 1840. In the Museo Donizettiano, Bergamo, Italy.
By courtesy of the Museo Donizettiano, Bergamo, Italy

Two years later he scored another lasting success with *L'elisir d'amore* (*The Elixir of Love*), a comedy full of charm and character with a libretto by Felice Romani, the best theatre poet of the day. *Lucrezia Borgia* (1833), also with a libretto by Romani, consolidated his reputation at La Scala opera house in Milan and elsewhere. Like the opera composers Gioacchino Rossini and Vincenzo Bellini before him, he next gravitated to Paris, where his *Marino Faliero*, though not a failure, suffered from comparison with Bellini's *I Puritani*, produced a few weeks before. Donizetti then returned to Naples for the production of his tragic masterpiece, *Lucia di Lammermoor* on September 26, 1835.

In 1828 Donizetti had married Virginia Vasseli, the sister of one of his closest friends in Rome; they made their home in Naples. He was deeply devoted to her and never really recovered his spirits after her death, soon after the stillbirth of a son, in 1837. His distress was exacerbated by the fact that none of the three children born to them survived birth. It seems clear that syphilis, to which Donizetti himself later succumbed, was already taking its toll of his family.

Donizetti continued to work in Naples until 1838, when municipal censorship objected to the production of his *Poliuto*, which dealt with a Christian martyr, on the ground that the sacred subject was unsuitable for the stage. He thereupon returned to Paris, where the field had been cleared for him by Bellini's early death and Rossini's retirement. There he revived some of his best operas, though *Lucrezia Borgia* had to be withdrawn because of objections by Victor Hugo, on whose drama the libretto was based. *Poliuto* was produced in 1840 as *Les Martyrs* with a French text by Eugène Scribe. It was preceded two months earlier by the opéra comique, *La fille du régiment* (*The Daughter of the Regiment*), which gained enormous popularity over the years through the performances of the leading sopranos of the day, including Jenny Lind, Adelina Patti, Marcella Sembrich, Emma Albani, and other divas of the 19th century. Later in the same year the Paris Opéra produced *La Favorite*, Donizetti's first essay in French grand opera.

Revival of his operas in Paris

Bartolomeo Merelli, a fellow pupil of Donizetti, was now director of La Scala and also of the Kärnthnerthor Theater, in Vienna. He engaged Donizetti to compose an opera for La Scala. The work, *Maria Padilla*, was produced in 1841 only a few weeks before the famous première of Verdi's *Nabucco*. Merelli also commissioned an opera for his Viennese theatre. There, *Linda di Chamounix*, a romantic *opera semiseria*, was produced in 1842 and dedicated to the empress Maria Anna. Donizetti had already been brought to the notice of the emperor Ferdinand I by his chancellor, Prince Metternich, and had conducted Rossini's *Stabat Mater* in his presence. He now received the appointment of official composer to the Emperor, which obliged him to be in Vienna for six months in the year but left him free to work elsewhere during the rest. At the same time Rossini, who had always furthered Donizetti's interests in Paris and entrusted to him the first performance of his *Stabat Mater* at Bologna, urged him to undertake the vacant directorship of the Liceo in that city. But Donizetti felt that he could not undertake this responsibility and preferred to continue his profitable operatic career. Back in Paris, he produced at the Théâtre Italien the delightful and witty comic opera, *Don Pasquale*.

Beginning of his decline

But Donizetti was already in the grip of his fatal disease. He produced his last important opera, *Dom Sébastien*, with a libretto by Scribe, at the Paris Opéra in 1843 under the strain of constant headaches and occasional lapses of mental capacity. He suddenly aged, lost his good looks and his equability of temper, which had hitherto seen him through the trials of operatic production. *Dom Sébastien*, though unfavourably reviewed in the press, was nonetheless a success with the public.

The remaining years were a story of degeneration into hopeless insanity, first in a private asylum near Paris, where, after considerable difficulties with the French police, who were supported by the doctors, he was at last taken home to Bergamo by his devoted nephew Andrea, son of his eldest brother. He lingered on until April 8, 1848, a victim of general paralysis of the insane, deprived of willpower, speech, and physical control. It was a pitiable end for a gay and handsome man who, unlike Bellini, was never envious of the successes of other composers and at all times displayed an openhearted generosity. To the French composer Hector Berlioz, for example, whose criticisms in *Le Journal des Débats* were consistently hostile, he spontaneously sent a letter of introduction to Prince Metternich, when Berlioz was about to leave for Vienna.

Donizetti always won more favour with the public than with the critics. During his lifetime his success was enormous and the rewards considerable. His popularity continued until the end of the century, but by 1914 his operas had almost disappeared from the repertory, overshadowed by the more substantial masterpieces of Verdi and Richard Wagner. In the 1950s there was a revival of interest in his works, after which it seemed unlikely that, at least, *Lucia di Lammermoor*, *L'elisir d'amore*, and *Don Pasquale* would be allowed to pass into oblivion.

MAJOR WORKS

OPERAS: 75 including *Alfredo il Grande* (1823); *Emilia di Liverpool* (1824); *Le convenienze ed inconvenienze teatrali* (1827); *Il borgomastro di Sardaam* (1827); *La regina di Golconda* (1828); *Il giovedì grasso* (1828); *Il castello di Kenilworth* (after Scott, 1829); *Anna Bolena* (1830); *L'elisir d'amore* (1832); *Il furioso all'isola di San Domingo* (1833); *Torquato Tasso* (1833); *Lucrezia Borgia* (1833); *Maria Stuarda* (1834); *Gemma di Vergy* (1834); *Marino Faliero* (after Byron, 1835); *Lucia di Lammermoor* (after Scott, 1835); *Belisario* (1836); *Il campanello di notte* (1836); *Betly* (1836); *Pia de' Tolomei* (1837); *Roberto d'Evereux, Conte d'Essex* (1837); *Poliuto* (1840); *La fille du régiment* (1840); *La favorite* (1840); *Linda di Chamounix* (1842); *Don Pasquale* (1843); *Maria di Rohan* (1843); *Dom Sébastien, roi de Portugal* (1843).

OTHER WORKS: Two oratorios, several cantatas, religious pieces, at least 20 string quartets, three string quintets, and numerous songs.

BIBLIOGRAPHY. GUGLIELMO BARBLAN, *L'opera di Donizetti nell'età romantica* (1948), the standard work on Donizetti's operas; ARNALDO FRACCAROLI, *Donizetti* (1944), the authoritative Italian biography; WILLIAM ASHBROOK, *Donizetti* (1965), a detailed study of the man and his works; HERBERT WEINSTOCK, *Donizetti and the World of Opera in Italy, Paris and Vienna in the First Half of the Nineteenth Century* (1963), the most comprehensive biography in English.

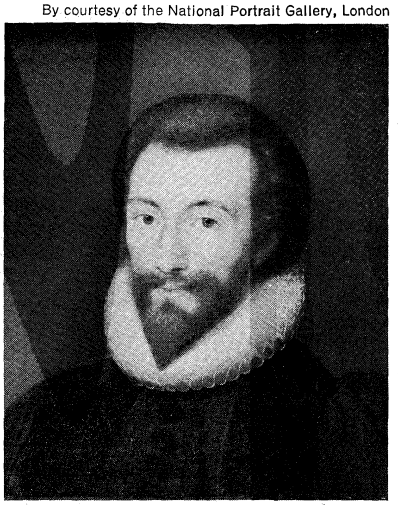
(D.Hus.)

Donne, John

As both poet and prose writer John Donne has had an especially powerful influence on writers of the 17th and the 20th centuries, who have found stimulus in his fusion of witty argument with passion, his dramatic rendering of complex states of mind, his daring and unhackneyed images, and his ability (little if at all inferior to Shake-

speare's) to make common words yield up rich poetic meaning without distorting the essential quality of English idiom. As a leading churchman and preacher, he helped to mold the outlook of the Church of England. He was, furthermore, engaged upon so many activities characteristic of his times that his career has an unusually strong historical interest.

By courtesy of the National Portrait Gallery, London



Donne, oil painting by an unknown artist, after I. Oliver, c. 1616. In the National Portrait Gallery, London.

Early life and career. Donne was born in London at some time between January 24 and June 19, 1572, of Roman Catholic parents. His mother, the daughter of John Heywood, epigrammatist and playwright, was a descendant of Henry VIII's chancellor, Sir Thomas More. Donne rightly claimed that no family had suffered more for its religious beliefs than hers. His father, descendant of an ancient Welsh family whose arms the poet bore, was a successful London merchant. Six months after his death in 1576 his widow married Dr. John Syminges, several times president of the Royal College of Physicians, who brought up the Donne children. According to Izaak Walton (whose charming *Life* of Donne was first published in 1640), Donne was educated at home by a Catholic tutor until, with his brother Henry, he matriculated at Hart Hall (later Hertford College), Oxford, in October 1584. After three years, according to Walton, Donne transferred to the University of Cambridge, though no surviving records confirm this. He did not take a degree from either university, however, for this would have required his accepting the Act of Supremacy and recognizing the Thirty-nine Articles of religion, to neither of which conditions, as a Catholic, he could subscribe. Extensive travels in Europe probably followed, but by May 1591 he was enrolled as a student of the law at Thavies Inn, whence he proceeded to Lincoln's Inn (May 6, 1592). In 1593 his brother Henry was arrested for harbouring a Catholic priest and died in prison of the plague. In June of the same year Donne, being now of age, received his portion of his father's estate. He remained at Lincoln's Inn until at least the end of 1594, having fulfilled the conditions necessary to proceed to final training for the profession of law. Like many others at that time, however, he regarded the Inns of Court only as a gateway to a state office and as a means of making contacts with men of affairs.

In 1596 Donne joined the gentlemen adventurers (many of them, like himself, aspirants to a position in public service) as a volunteer in the successful naval and military expedition against Cádiz. In the following year he took part in the Islands expedition, hunting for Spanish treasure ships in the Azores. One of his companions was Thomas Egerton, who recommended Donne to his father, Sir Thomas Egerton, lord keeper of the great seal. By the beginning of 1598 Donne had become a secretary to Sir Thomas in an office that was a recognized path to high public appointments. Sir Thomas thought well enough of Donne to send him into Parliament in 1601. He took no part in the debates or committees there, however, his function apparently being to keep his finger on the pulse of the House and report to his master. At this time began Donne's rejection of Catholicism (see below).

Early public career

Donne was well on the way to achieving his ambitions. He had also by this time written much of his poetry, most of it in imitation of ancient Latin poets: there were verse letters to male friends (the earliest dating from about 1592); five formal verse satires in the modes of Horace and Juvenal (*c.* 1593–98) and one canto of a satirical epic (*Metempsychosis*, 1601); a "book" of love-elegies in part modelled on Ovid's; and epigrams that to some extent imitated those of Martial. The classical models, however, are so transformed by the wit and daring of Donne's imagination that the verses are thoroughly original. By 1601 he had also composed numerous love lyrics in various moods, which are among his most famous poems. Such are the vigour, realism, and passion of his work at this period that they have sometimes been thought to reflect Donne's own youthful outlook and conduct. His contemporary, Sir Richard Baker, wrote of him as "not dissolute (*i.e.*, careless), but very neat; a great visitor of Ladies, a great frequenter of Plays, a great writer of conceited Verses." The moral fastidiousness and earnestness that Donne displayed throughout his life, however, make it difficult to believe that he was a youthful rake. His visits to the theatre in that great age of English drama reflected but one of the many aspects of literature and learning in which he interested himself. Walton emphasizes Donne's unusual studiousness throughout his life; and he was engaged, at least as early as 1593, in a deep consideration of the points at issue between the English and Roman churches. The "Ladies" he visited were doubtless noblewomen like those who were later to become his patrons.

His wit and charm made him a favourite with the ladies in the lord keeper's household. He was particularly attracted to Anne More, daughter of Sir George More of Loseley Park and niece and protégé of Egerton's second wife. Anne's father had brought her to London for the assembling of Parliament in 1601. She and Donne met secretly, and since there was little chance of winning the consent of so ambitious and important a man as Sir George, they agreed to marry without it, apparently early in December 1601. This was an offense against both civil and canon law, and Donne's risking so much for love was truly what Walton called "the great error of his life." When in February 1602 Donne told Sir George of the marriage, his irate father-in-law had the poet imprisoned for a time, persuaded Egerton to dismiss him, and brought the matter before the Court of High Commission. Though the commissioners judged the marriage valid, there now began for Donne a long and difficult period of unemployment, during which, however, his love for his wife never faltered.

Marriage to Anne More

Churchman and preacher. Donne first found shelter in the house of Lady Egerton's son, Sir Francis Wolley, at Pyrford, Surrey. Here he left his family for about a year (1605–06) while he travelled in France and Italy. From 1606 until 1611 he had his own house at Mitcham, south of London, and an apartment in London itself. For some time he seems to have assisted Thomas Morton (later bishop of Durham) in controversial tracts against the Catholics, more particularly in defense of King James I's exaction of the oath of allegiance as head of the Church of England. Donne's own contribution was an effective piece of propaganda in support of the King called *Pseudo-martyr* (1610), in recognition of which the Oxford University conferred on him the degree of master of arts. An offshoot of the controversy was his satiric fantasy against the Jesuits, *Ignatius his Conclave* (published in Latin and in English in 1611).

Donne often felt despondent at the lack of direction in his life. He had thoughts even of suicide, but these were exorcised in a casuistical work, *Bianthanatos* (written *c.* 1608; published 1646), in which he argued the grounds on which one could rightly take one's own life. More

especially, he experienced a deep personal religious struggle. Morton had more than once urged him to enter the ministry of the Church of England, most strongly in 1607, when he offered Donne a benefice he was himself vacating. Donne refused as being unworthy and as having no convincing vocation for holy orders. He tried unceasingly to secure secular employment but without success. By the time he had written *Pseudo-martyr*, moreover, King James made it clear that he wished Donne to enter the church and that he did not favour secular employment for him.

Theological views

Donne had, of course, been for some time a convinced Anglican. In 1596–97 he had served a Protestant queen against her Catholic enemies, and he must have at least formally renounced Catholicism before he could enter Egerton's service. He eventually maintained that any church asserting the main tenets of the Christian religion was a true church; in "indifferent" matters (government, ritual, and so on) the practice of one's own nation should be followed. Though he never doubted the truth of Christianity, his lack of complete conviction concerning his own salvation persisted. His struggle for assurance is expressed in religious poems written between 1607 and 1613, which include some of the finest in English. Donne was able to support his growing family by means of the belated payment of his wife's dowry and by the patronage (after 1607) of Lucy Russell, countess of Bedford. To her and to other noble ladies he wrote characteristically original verse letters, and he composed elegies on the deaths of some of Lady Bedford's kinsfolk.

In 1611 Donne found another patron, Sir Robert Drury, with whom he toured in France and the Low Countries in 1611–12, who provided the Donne family with accommodations (until 1621) attached to the Drury residence in London. The death of Drury's young daughter, Elizabeth, in 1610 had called forth a funeral elegy from Donne; and in 1611 and 1612 he wrote *Anniversaries*, also in commemoration of her death, though here the theme is used mainly as a pretext for a profound satirical "anatomy" of a decaying world, a contemplation of the soul's flight to share the joys of heaven. Still seeking state employment, Donne wrote verses on the death of Prince Henry (1612) and on the marriage of Princess Elizabeth to Frederick V, elector of the Palatinate (1613). He sat in the "Addled" Parliament of 1614 and served on four select committees. He wrote an epithalamion (a nuptial poem) in celebration of the wedding of the notorious Countess of Essex to the court favourite, Robert Carr, earl of Somerset, whom Donne had been cultivating for help in securing an official position. Late in 1614 a court post seemed at last within his grasp, but the King again refused to appoint Donne to any position outside the church. Convinced finally that he was called to the ministry at the hands of God's representative, the King, Donne completed studies in Greek and Hebrew, finished the *Essayes in Divinity*, and began a new career in the church.

Ordination into the Church of England

He was ordained deacon and priest by Bishop John King on January 23, 1615. Preferment soon followed. He was made a royal chaplain and received, at the King's command, the degree of doctor of divinity from the University of Cambridge. He occupied the rectory of Keyston, in Huntingdonshire (the King's gift) from 1616 until 1621; the rectory of Sevenoaks in Kent (an unsolicited gift from his old master, Egerton) from 1616; that of Blunham, in Bedfordshire, from 1622; and the vicarship of St. Dunstan-in-the-West, London, from 1624 until his death. He was scrupulous, however, in confining his tenure of those benefices that carried "cure of souls" to the two allowed by law to royal chaplains. From October 1616 he occupied the important post of reader (preacher and spiritual director) at Lincoln's Inn. His preaching soon developed great power and eloquence.

In August 1617, Donne's wife died in giving still-birth to their 12th child. Her loss caused him great sorrow but also deepened the devotion and dedication of his religious life. For a long time after his bereavement Donne's health gave his friends concern, and in 1619 he was freed from his duties at Lincoln's Inn to go as chaplain with the embassy of James Hay, Viscount Doncaster to the princes of Germany and Bohemia on the eve of the Seven Years' War, the great conflict involving all the major European powers except Turkey. On November 22, 1621, Donne was installed as dean of St. Paul's Cathedral. Many duties fell to him, and Donne was most conscientious in performing them. The register of his administration at St. Paul's proves his efficiency and integrity.

Dean of St. Paul's Cathedral

The death of King James in 1625 made little difference to Donne's status. The new king, Charles I, admired him as preacher and as poet. The only poems that Donne wrote after his ordination, however, were three fine "Hymns," two or three Holy Sonnets, and an elegy on the death of Lord James Hamilton. His creative talents were occupied almost wholly in preaching and in composing the prose *Devotions upon Emergent Occasions*, written during his recovery from a serious bout of relapsing fever (1624). He published only a few more important sermons; but in 1625, when he was sheltering in the house of Magdalen Herbert (Lady Danvers) from the severe plague that scourged London, and again in 1630, when he was trying to recover from his last illness at the home of his son-in-law, Donne transcribed from his notes more than 100 other sermons, which were published after his death by his son. His last years were saddened by the death of friends and patrons, of children (only six of whom survived their father) and, in January 1631, of his aged mother. Though emaciated by illness, Donne preached at court his last sermon (published posthumously as *Death's duell*) on February 25, 1631; the next day he attended his last meeting of the governors of the Charterhouse. He conducted cathedral business for the last time on March 21. On March 31, 1631, Donne died, and on April 3 he was buried in his cathedral. According to Walton, Donne had caused a drawing to be made of himself in his shroud during his last illness; from this Nicholas Stone carved an effigy in white marble, which, alone of all the monuments, survived the destruction of old St. Paul's in the Great Fire of 1666 and which may still be seen in the south aisle of the present cathedral.

Posthumous reputation. When Donne published his *Anniversaries* he apologized to his friends for having "descended to print anything in verse," for, at the time in which he lived to print one's verses smacked of commercialism. Only one of his poems in English survives in his own handwriting; most were preserved in manuscript copies made by and passed among a relatively small but admiring coterie of poetry lovers. The contents of the first two (posthumous) editions of his *Poems* (1633, 1635), apart from the few already printed, were in fact derived from collections of such transcripts. The *Poems* were sufficiently popular to be published eight times within 90 years of Donne's death, but his work was not to the general taste of the next century, when he was regarded as a great but eccentric "wit." From the beginning of the 19th century, however, perceptive readers began to recognize Donne's poetic genius. In the 20th century his sermons as well as his poems have attracted great attention but have had little literary influence upon other writers. He has, however, been established as a poet of great distinction (though uneven in quality), and as the master of what has been called the Metaphysical school of poets, has again been a strong influence both in the writing of poetry and in determining the emphases of literary critics and historians.

MAJOR WORKS

POEMS: The two *Anniversaries* (*An Anatomie of the World* and *Of the Progres of the Soule;* 1611 and 1612) were published during Donne's lifetime. The first (incomplete) collection of his verse was published posthumously in 1633. To the period before Donne's marriage (1601) belong most of his satires and elegies and many of his songs and sonnets. To the period between his marriage and ordination (1615) belong the verse letters, the rest of the elegies and songs and sonnets, and the first religious verses. Among these works are many of Donne's finest poems, including (titles followed by first lines): "Song" ("Goe, and catche a falling starre"); "The good-morrow" ("I wonder by my troth, what thou, and I/Did, till we lov'd?"); "The Exstasie" ("Where, like a pillow on a bed"); "A Valediction: forbidding mourning" ("As virtuous men passe mildly away"); "The Baite" ("Come live

with mee, and bee my love"); "The Sunne Rising" ("Busie old foole, unruly Sunne"); "The Canonization" ("For Godsake hold your tongue, and let me love"); "The triple Foole" ("I am two fooles, I know, /For loving, and for saying so"); "Twicknam garden" ("Blasted with sighs, and surrounded with teares"); "A nocturnall upon S. Lucies day . . ." ("Tis the yeares midnight, and it is the dayes"); "The Will" ("Before I sigh my last gaspe, let me breath"); "The Funerall" ("Who ever comes to shroud me, do not harme/ Nor question much"); and "The Relique" ("When my grave is broke up again /Some second ghest to entertain"). To the period of Donne's life between 1607 and his death (1631) belong the Divine poems, including the Holy Sonnets (notably I, "Thou Hast made me"; VII, "At the round earth's imagin'd corners blow"; X, "Death be not proud"; XI, "Spit in my face yee Jewes"; XIV, "Batter my heart, three-person'd God"; XVIII, "Show me deare Christ, thy spouse"). Numbering per H.J.C. Grierson (see Bibliography below).

PROSE: Donne's prose published during his lifetime includes *Pseudo-martyr* (1610); *Ignatius his Conclave* (1611; the Latin version, *Conclave Ignati*, had appeared that same year); *Devotions upon Emergent Occasions* (1624); *Foure Sermons on Speciall Occasions* (1625; reprinted with corrections and one further sermon, 1626); *Death's duell* (1630). Posthumous editions of Donne's prose include *Juvenilia* (1633); *LXXX Sermons* (1640); *Biathanatos* (1646); *Fifty Sermons* (1649); *Essayes in Divinity* (1651); *Letters to severall persons of honour* (1651); *Paradoxes, problems, essays, characters . . .* (1652); *XXVI Sermons* (1661). The principal editor of these writings was John Donne the younger.

BIBLIOGRAPHY. GEOFFREY L. KEYNES, *A Bibliography of Dr. John Donne*, 4th ed. (1972), contains full details of manuscripts, editions, and scholarly works. ROBERT C. BALD, *John Donne: A Life*, ed. by W. MILGATE (1970), is the definitive biography, and mentions all the available sources and documents. The fullest collection of Donne's letters is that in EDMUND W. GOSSE, *The Life and Letters of John Donne, Dean of St. Paul's*, 2 vol. (1899, reprinted 1959). The standard edition of the *Poems* is that of H.J.C. GRIERSON, 2 vol. (1912), but it is being replaced by four volumes: *The Divine Poems* (1952) and *The Elegies and the Songs and Sonnets* (1965), ed. by HELEN GARDNER; and *The Satires, Epigrams and Verse Letters* (1967) and the *Epithalamions, 'Anniversaries' and Epicedes* (forthcoming), ed. by W. MILGATE. The best introduction to the poems is JAMES B. LEISHMAN, *The Monarch of Wit: An Analytical and Comparative Study of the Poetry of John Donne*, 7th ed. (1965). Standard editions of the prose works are: *The Sermons of John Donne*, ed. by GEORGE R. POTTER and EVELYN M. SIMPSON, 10 vol. (1953–62); *Devotions upon Emergent Occasions*, ed. by JOHN SPARROW (1923); *Ignatius His Conclave*, ed. by T.S. HEALY (1969); *Essays in Divinity*, ed. by EVELYN M. SIMPSON (1952); and *Paradoxes and Problems*, ed. by GEOFFREY L. KEYNES (1923). There is a facsimile of *Biathanatos*, with a bibliographical note by J. WILLIAM HEBEL (1930); *Pseudo-Martyr* (1610) has never been reprinted. The best guide to the prose is in EVELYN M. SIMPSON, *A Study of the Prose Works of John Donne*, 2nd ed. (1948).

(W.Mi.)

Don River

One of the great rivers of the European portion of the Soviet Union, the Don has been a vital artery in that nation's history ever since the days of Tsar Peter the Great, who initiated a 14-map hydrographic survey of its course. Throughout the world, the river is associated with images of the turbulent and colourful Don Cossacks, a famous series of novels by Mikhail Sholokhov, a Nobel Prize winner, and a series of large-scale engineering projects that have enhanced the waterway's economic importance. The Don rises in the small reservoir of Shat, located in the central Russian uplands near the city of Novomoskovsk. It flows first in a generally southerly direction for 1,162 miles (1,870 kilometres), draining a basin of 163,000 square miles (422,000 square kilometres), before it enters the Gulf of Taganrog (Taganrogsky Zaliv) in the Sea of Azov. It is one of the major rivers of the Russian Soviet Federated Socialist Republic, lying between the equally important basins of the Volga (to the east) and the Dnepr. In its middle and lower courses, from the confluence with the Chërnaya Kalitva to its mouth, the Don describes an enormous eastward-bulging arc as far as its junction with the Ilovlya. Near the top of the arc, the vast Tsimlyansk Reservoir (Tsimlyanskoye Vodokhanilishche) begins, joined at the city of Kalach-na-Donu by the important Volga–Don Canal. Between the Ilovlya and the canal, the westward-swinging Volga approaches to within a mere 50 miles or so of the Don. From its source in the Tula Oblast (region), the Don crosses the Lipeck, Voronezh, Volgograd, and Rostov Oblasti, through the forest-steppe and renowned steppe zones of the Soviet Union. It gathers in the waters of numerous tributaries, the most famous of which are the Krasivaya Mecha, Sosna, Chërnaya Kalitva, Chir, and Donets (right bank), and the Voronezh, Khopër, Medveditsa, Ilovlya, Sal, and Manych (left bank). The river winds throughout its course, and the fall along its length is about 620 feet (190 metres). (For related information, see RUSSIAN SOVIET FEDERATED SOCIALIST REPUBLIC; SOVIET UNION; RUSSIAN STEPPE.)

Important tributaries

The river course. In the upper portion of the Don—that is, as far downstream as the southeastward bend—the river flows along the eastern edge of the central Russian heights through a generally narrow valley. The right bank is well developed, reaching heights of 160 feet at the cities of Dankov and Lebedyan, and its limestone and chalk rocks are cut into by ravines and gullies. The left bank borders a flatter floodplain, and the river itself widens intermittently into small lakes; its low water depth ranges from less than a few feet on the shoals to 33 feet, with a breadth widening to 650–1,300 feet.

In the middle course, to the beginning of the Tsimlyansk Reservoir, the valley widens to about four miles, and its path is marked by floodplains, more small lakes, and dried-up old courses; the banks, especially the right bank, become steeper, with chalk, limestone, and sandstone predominating. The river width narrows to 330–1,300 feet, and the low-level depth is about 50 feet.

The lower course is dominated by the 160 miles of the Tsimlyansk Reservoir, named after a giant hydroelectric plant built between 1952 and 1955. Up to 24 miles in width, the reservoir has an average depth of 28.9 feet. Finally, the lower section of the Don has a valley width of 12–19 miles, with a huge floodplain and a braided river channel as much as 66 feet deep.

The river basin. The landscape of the upper and middle Don Basin is characterized on the right bank by undulating plains cut into by jagged gorges and on the left bank by the smoother, pond-dotted topography of the Oka-Don Plain (Oksko-Donskaya Nizmennost). Farther downriver the vast open landscapes of the steppes predominate. Rich, black soils fill almost the entire basin, though there are patches of gray forest soil in the north, where forests cover up to 12 percent of the area.

The climate of the basin is moderately continental, with average January temperatures ranging from 18° F to 12° F (−8° C to −11° C), while July readings reach 66° F to 72° F (19° C to 22° C). Annual precipitation diminishes from 23 inches in the north to 14 to 15 inches in the south.

The Don runs mostly through treeless countryside, but plantations have been made in parts of the upper reaches in the hope that tree roots will help bind together and strengthen the riverbanks. In the gentler lower regions the banks themselves have been tilled. The river is rich in fish, especially in the lower course.

Water flow. The long-term fluctuations in the water level of the Don reach about 39 feet in the upper course, 26 feet in the middle course, and 20 feet in the lower. The highest levels are in the springtime, the lowest in autumn and winter. At the mouth of the Don strong winds from the sea cause increases in the water level. Interesting variations occur in the amount of water flowing past given points along the river course: in the upper course, at the city of Georgy-Dezh, an average flow of 8,857 cubic feet per second is achieved by flows ranging from 1,480 to approximately 395,000 cubic feet per second, and there are corresponding variations as the total volume recorded increases at stations downstream. At the city of Kalach-na-Donu it has been calculated that about 65 percent of the annual flow occurs during April and May, with about 7 percent in March, before the influx of melted snow. Below Tsimlyansk Reservoir, the flow has been brought under partial control: at Nikolay-

Fluctuating water levels

evskaya, for example, 34 percent of the annual volume occurs in spring, 33 percent in summer, 22 percent in autumn, and 11 percent in winter.

The Don begins to freeze, in its northern portion, about the middle of November and is clear of ice by about April 10. In the lower course the river is frozen between the end of November and the end of March at Kalach-na-Donu, and from mid-December to the beginning of April at Rostov-na-Donu.

The imprint of man. In the upper river basin an extensive network of ponds aids irrigation. The ponds are also used for building up stocks of fish. In the lower course of the Don an extensive irrigation network has been developed, channelled through the main Don canal, which emerges from the reservoir at the Tsimlyansk hydroelectric plant, branching out into, among others, the Lower Don, Upper Sal, and Proletariat canals. The plant at Tsimlyansk incorporates a hydroelectric station, a fish elevator, two navigable sluices, an irrigation canal, a 1,581-foot concrete overflow dam, and an eight-mile earthen dam. The full capacity of the reservoir is less than five cubic miles (22 cubic kilometres), but the average capacity is about half this amount.

The significance of the Don as a navigable waterway greatly increased with the construction of the Volga–Don Canal (Volgo-Donsky Kanal). The river itself is navigable from the mouth to the city of Georgy-Dezh (a distance of 842 miles [1,355 kilometres]) and in the spring as far as the village of Khlevnoye, 990 miles (1,590 kilometres) from the sea. Navigation in the lower course has been facilitated greatly by the Tsimlyansk project. At the mouth of the Don, navigational difficulties occasionally occur as a result of a fall of water level following a wind-induced water movement away to the south, while many strengthening and dredging operations are necessary to maintain and improve navigation in the upper reaches. The largest ports are Kalach-na-Donu, Tsimlyansk, and Rostov-na-Donu. (A.M.Ga.)

Dorgon

Although he never bore the title of emperor during his lifetime, the Manchu prince Dorgon played a leading part in establishing the Ch'ing dynasty in China and was the first of his house to wield effective power there.

Dorgon was born in Yenden, Manchuria, on November 17, 1612, the 14th of the 16 sons of Nurhachi, founder of the Manchu state, who in 1616 proclaimed himself emperor of China but died in 1626 before making good his claim to imperial title. Under his successor, Abahai, Dorgon received the title of an imperial prince, *hosoi beile*. He distinguished himself in the wars against the Chahar Mongols that began in 1628 and was elevated to prince of the first degree (*jui ch'in-wang*). Dorgon commanded one of the two army groups that breached the Great Wall and sacked 40 cities in the Chinese provinces of Hopeh and Shantung during Abahai's campaigns to subjugate China in 1638–39. He also participated in the capture of the cities of Sung-shan and Chin-chou that resulted in a significant expansion of Manchu authority.

Rise to power

On Abahai's death in 1643, Dorgon was nominated his successor but declined, reportedly because of loyalty to the dead emperor. Instead, he and the older Prince Jirgalang became regents for Abahai's five-year-old son, Fu-lin. The fact that Dorgon executed two princes when he discovered their plot to put him on the Imperial throne is characteristic of the high moral standards for which he is praised by historians.

When in April 1644 the troops of the Chinese rebel Li Tzu-ch'eng conquered Peking (the capital of China then ruled by the Ming dynasty), Dorgon, on the advice of a Chinese counsellor, led an expeditionary force into China. His former principal enemy, the Chinese general Wu San-kuei, joined forces with him rather than allow Li Tzu-ch'eng to establish his own dynasty; and the combined armies inflicted a heavy defeat on Li Tzu-ch'eng's troops. Dorgon entered Peking in June 1644, but the last emperor had already hanged himself in April. After pursuing the fleeing troops of Li Tzu-ch'eng, Dorgon turned his attention to the stabilization of his administration, prudently enlisting the cooperation of several outstanding Chinese experts. He established Peking as the capital and, adopting many Chinese customs, laid the basis for Manchu rule in China.

Entry into Peking

The youthful Fu-lin entered Peking on October 19, 1644, and 11 days later was proclaimed emperor under the name Shun-chih. In 1644 Dorgon subdued the provinces of Shensi, Honan, and Shantung; Kiangnan, Kiangsi, Hopeh, and part of Chekiang followed in 1645; and the provinces of Szechwan and Fukien were added in 1646. Rebellious Ming troops were pushed back into the southwestern provinces of the country, and Dorgon suppressed revolts of the Mongolian tribes in Central Asia.

He took over the highly developed administrative system of his Chinese predecessors, re-engaging Chinese experts and recruiting new civil servants through the proven method of selection and examination. Adam Schall von Bell, a German Jesuit missionary, served him as mathematician, director of the Imperial Board of Astronomy, and adviser on the manufacture of artillery. All these measures contributed to the generally favourable acceptance of the new dynasty, notwithstanding the forcible expropriation of land and the introduction of Manchurian customs, such as the pigtail.

Relegating Prince Jirgalang to the functions of assistant prince regent, Dorgon in 1644 began to gather more and more power in his hands, even venturing to impose humiliations on his nephew Haoge and other Imperial princes who opposed him. Receiving the title of Imperial father regent in 1648, he personally led the campaign against a rebellious Chinese general in Shansi. He also designed the plans for the construction of his own palaces in Jehol; here he intended to spend his remaining years as feudal overlord, but he died on December 31, 1650, during a hunt at Kharahotun, near the Great Wall. He was posthumously proclaimed emperor and given the temple name of Ch'eng Tsung.

Dorgon's sudden death created confusion and disorder in the empire. Since he had left no male heirs, disturbances broke out, especially among the corps of the white banner unit that had been under his command. Internal shifts on the political scene brought his former enemies to power; they had succeeded in obtaining the promulgation of an Imperial decree of March 1651 declaring that Dorgon had been a usurper. He was posthumously deprived of his princely rank, along with other honours; his relationship to the Imperial house was disavowed; and a petition of two officials attempting to redeem his reputation was rejected. Only after the Ch'ien-lung emperor, in 1773, honoured Dorgon's services in establishing the new dynasty and restored his neglected grave was Dorgon finally fully rehabilitated.

Posthumous reputation

BIBLIOGRAPHY. ERICH HAUER, "Prinz Dorgon," *Ostasiatische Zeitschrift*, n.s., 13:9–56 (1926), a translation with introduction of the extensive biography from the official biography collection of 1765; HELLMUT WILHELM, "Ein Briefwechsel zwischen Durgan und Schï Ko-fa," *Sinica*, 8:239–245 (1933), translation of the famous letter to the Ming general; FRANZ MICHAEL, *The Origin of Manchu Rule in China* (1942, reprinted 1965), a concise, basic survey of the historical development of the beginning Manchu empire; FANG CHAO-YING, "Dorgon," in ARTHUR W. HUMMEL (ed.), *Eminent Chinese of the Ch'ing Period, 1644–1912*, vol. 1, pp. 215–219 (1943), a résumé of a biography of Dorgon.

(M.Gi.)

Dormancy

There are few environments in which organisms are not subject to some kind of environmental stress. Some animals, such as birds, migrate vast distances to avoid unfavourable situations; others reduce environmental stresses by modifying their behaviour and the habitats (immediate surroundings) that they occupy. Arctic lemmings, for example, are able to avoid severe winter weather by confining their life in winter to activities beneath the snow cover.

Still another mechanism used by some organisms to avoid stressful environmental conditions is that of dormancy, an inactive state accompanied by a lower than

normal rate of metabolism—the chemical processes responsible for the activity, nourishment, and growth of an organism—during which an organism conserves the amount of energy available to it and makes few demands on its environment. Most major groups of plants and animals have some representatives that can become dormant; more species hibernate than become dormant, however. Periods of dormancy vary in length and in degree of metabolic reduction, ranging from only slightly lower metabolism during the periodic, short-duration dormancy of deep sleep to more extreme reductions for extended periods of time. The longest recorded period of dormancy is that for the seeds of *Spergula arvensis*, which sprouted after being dormant for 1,700 years.

GENERAL FEATURES

Value of dormancy. In terms of evolution, dormancy seems to have evolved independently among a wide variety of living things, and the mechanisms for dormancy vary with the morphological and physiological makeup of each organism. For many plants and animals, dormancy has become an essential part of the life cycle, allowing an organism to pass through critical environmental stages in its life cycle with a minimal impact on the organism itself. When lakes, ponds, or rivers dry up, for example, aquatic organisms that can enter a period of dormancy survive, while others perish. Moreover, animals that can become dormant during the extreme cold of winter can extend their ranges into regions where animals incapable of dormancy cannot live. Dormancy also ensures that these animals will be free from competition during their periods of activity. Thus, dormancy is an adaptive mechanism that allows an organism to meet environmental stresses and to take advantage of environmental niches that otherwise would be untenable at certain times.

Causes of dormancy. The dormant state that is induced in an organism during periods of environmental stress may be caused by a number of variables. Those of major importance in contributing to the onset of dormancy include changes in temperature and photoperiod (length of exposure to light) and the availability of food, water, oxygen, and carbon dioxide. In general, because organisms normally exist within a relatively narrow temperature range, temperatures above or below the limits of this range can induce dormancy in certain organisms. Temperature changes also affect such other environmental parameters as the availability of food, water, and oxygen, thus providing further stimuli for dormancy. In Arctic regions, for example, certain animals become dormant during the winter months because food is less abundant. In desert biomes, on the other hand, the summer months, which may be periods of reduced food availability, intense heat, or extreme aridity, stimulate some desert organisms to become dormant. The lack of water during summer periods of drought or winter periods of freezing, as well as annual changes in the duration and intensity of light, particularly at high latitudes, are other environmental factors that can induce dormant states.

Cyclical environmental changes

Under natural conditions, most of the environmental variables that influence dormancy are interrelated in a cyclical pattern that is either circadian (daily) or annual. Fluctuations in the major daily variables—light and temperature—can induce rhythmical changes in the metabolic activity of an organism; annual fluctuations in temperature and photoperiod can influence the availability of food and water. Concentrations of oxygen and carbon dioxide normally do not vary on a cyclical basis but as a result of habitat selection, such as burrowing in the mud, seeking a den, or other similar activities, in which the metabolic responses of the organism can alter the oxygen and carbon dioxide concentrations in its environment.

In an attempt to determine the relative influence of environmental factors upon dormancy, they have been varied experimentally. Investigations indicate that an organism, after it has adapted to a sequence of cyclical rhythms, tends to maintain its adaptive behaviour even though the environmental stimulus that originally elicited such behaviour is no longer present. For example, the Arctic ground squirrel (whose winter period of dormancy is referred to as hibernation), when taken into the laboratory, supplied with adequate amounts of food and water, and exposed to constant temperature and light, exhibits periodic torpor (extreme sluggishness)—an innate behavioral pattern that operates independently of environmental cues. Other animals frequently will continue to respond as if they were exposed to the cyclical changes of their home environments after they are removed from their natural habitats.

DORMANCY IN BACTERIA AND FUNGI

Once in a dormant state, such organisms as bacteria and certain parasites are resistant to a wide variety of environmental conditions that otherwise would be fatal. While dormant, they can withstand boiling temperatures or the harmful effects of acids and other hostile chemicals. Moreover, the protective mechanisms of dormancy have in many cases extended far beyond those needed for adapting the stressful environmental situation that originally induced it; for example, dormancy makes certain organisms particularly immune to the effects of environmental pollution.

Spores

Bacteria and fungi become dormant by forming small, generally rounded spores (bodies that can develop into new organisms without fertilization) that are highly resistant to both physical and chemical agents. Yeast cells, for example, are usually killed in five to ten minutes by moist heat at temperatures of 50°–60° C (120°–140° F). To kill their spores, however, temperatures of 70°–80° C (160°–180° F) for a similar time-span are required. Cells of many other fungi are killed in 30 minutes by moist heat at a temperature of 62° C (144° F), but their spores can withstand 80° C heat for 30 minutes before being killed. Similarly, bacterial cells that can be destroyed in five to ten minutes by moist heat at temperatures of 55°–65° C (130°–150° F) have spores that can withstand temperatures of 120° C (250° F) for approximately 15 minutes or longer before being killed. In order to destroy the spores of *Clostridium botulinum*, the bacterium responsible for botulism, they must be heated to temperatures of 180° C (360° F) for ten minutes. The resistance of spores to high temperatures explains the care necessary in the preparation of canned foods, which otherwise might harbour spores that survive and germinate to produce toxins (poisons).

Although heat may destroy micro-organisms, low temperatures, no matter how extreme, seldom kill all members of a sporeforming microbial population. Micro-organisms maintained at subfreezing temperatures display no detectable metabolic activity, but they become active again when the temperature is increased. Similarly, dried spores of micro-organisms remain viable indefinitely; some can be subjected to extreme dehydration in the frozen state and sealed under a vacuum, yet the desiccated organisms remain viable for many years.

DORMANCY IN VASCULAR PLANTS

Vascular plants, such as ferns, conifers, and flowering plants, contain special tissues or vessels for the circulation of fluids. Found everywhere on land except at very high altitudes and latitudes and in permanently arid deserts, such plants respond in various ways to cold.

Responses to cold. Among the annuals, the entire plant dies in the fall; only its dormant seed survives the winter (overwinters) to produce a new generation. In biennials, such as carrots, the plant overwinters by using food reserves stored in the roots; the leaves die off during the first winter. The second-year plant, which arises from the root and a small piece of stem, produces seeds that survive during the second winter; the rest of the plant dies. Only the perennial plant can overwinter in a dormant state and survive for centuries. As winter approaches, the amount of colloid (substances divided into fine particles that are dispersed in a medium) in the cells of perennial plants increases, with a subsequent reduction in the amount of free water in them. Thus, there is less damage to the cell if freezing occurs. In herbaceous perennials, such as the dandelion, seeds are produced an-

nually. The aboveground parts of the plant persist only through the relatively short growing season and then die off in the fall, leaving the roots and underground stems; these contain sufficient stored food to last through the winter and to produce new shoots in the spring.

Among trees and shrubs, evergreens retain their leaves, and the entire plant overwinters at a reduced metabolic rate. Deciduous plants shed their leaves in the fall, and the rest of the plant winters at a lower metabolic level, using the stored food reserves accumulated during the growing season.

Buds

The buds and embryonic leaves that develop during the growing season remain dormant until they sprout in the spring to produce new branches and foliage. Buds of bulbs, tubers, and other plant organs may also display dormancy. Some plant buds, known specifically as dormant buds, may never develop except under unusual conditions, such as when a small branch is broken off. Then the dormant bud begins to grow.

The role of seeds. That the annual and biennial plants overwinter as seeds indicates that the seed is a fairly resistant part of a plant's life cycle. Plant seeds hold the record for the longest known periods of dormancy: well-documented cases have established that seeds of the Manchurian lotus were still fertile after 1,000 years; and, as mentioned above, seeds of *Spergula arvensis* found in Denmark germinated after 1,700 years. Seeds buried to a depth of more than eight centimetres (three inches) normally do not germinate, apparently because the carbon dioxide level is too high and the oxygen level too low to allow germination. The ideal environment for seed survival over long periods of time seems to be in soils that are slightly to moderately moist and deficient in oxygen.

Dry seeds can withstand temperature extremes better than any other plant part without losing their viability. If, for example, dry sugar beet seeds are exposed to a temperature of −180° C (−290° F) for 30 minutes, 96 percent of them will still germinate. Furthermore, these seeds will also germinate if subjected to a temperature of 103° C (217° F) for 16 hours.

Although the majority of seeds germinate readily as soon as environmental conditions are appropriate, many seem to require a distinct dormant period. Even with proper environmental conditions, such seeds will not germinate until they have first gone through a dormant stage. The length of seed dormancy may be one winter or may extend for several years. There are two major types of seed dormancy, seed-coat and embryo.

Seed-coat dormancy. Seed-coat dormancy occurs when the seed coat either is impermeable to water or oxygen or is resistant to the expansion of the embryo. That carbon dioxide cannot escape from a seed must also be considered as a factor contributing to its dormancy. Seeds that are impermeable to water include those of many legumes, such as the locust, clover, and alfalfa. The spiny fruit of the common cocklebur (*Xanthium*) produces two seeds the coats of which are impermeable to oxygen. In nature, germination of the cocklebur seeds is aided by frost, heat, or other environmental conditions that gradually erode the seed coat and allow oxygen to enter. Because the upper seed of the cocklebur fruit requires more oxygen for germination than the lower one, the lower seed usually germinates the first year; the upper seed germinates a year later. Although the coats of seeds such as pigweed (*Amaranthus*) allow the passage of both oxygen and water, the embryos are unable to break through, and the seeds do not germinate. In such cases the seed coat must be cracked or weakened so that the embryo is not injured, after which germination readily takes place. Various conditions serve to crack seeds in nature: *e.g.*, fire, being stepped on, freezing, being ingested and passing through a digestive tract, and chemical or physical abrasion.

Embryo dormancy. Embryo dormancy may result from either an undeveloped (rudimentary) embryo or the failure of the embryo to begin growing. In the case of a rudimentary embryo, several months may elapse before the embryo has developed to the point of emerging from the seed; the buttercup (*Ranunculus*) is an example.

After-ripening

There are also certain seeds in which the embryo does not respond even if the seed coat is removed and the seed placed in optimal germinating conditions. In these seeds germination occurs only after a series of changes have taken place in the embryo; this process is referred to as afterripening. Afterripening sometimes occurs during the winter, allowing the seed to germinate in the spring. Often, however, afterripening continues over a period of years; some seeds from the plant germinate each year. This seems to be a mechanism for enabling a species to survive unfavourable growth conditions that may extend over several growing seasons. Afterripening is initiated by a variety of conditions, depending on the seed. In the hawthorn, an increase in acidity is required; in peaches and cherries, the requirement seems to be moist soil and exposure to temperatures ranging from 0° to 10° C. In such cases, afterripening may take from one to six months. Chemical treatment may also encourage it.

DORMANCY IN PROTOZOANS AND INVERTEBRATES

Cysts and cystlike structures. *Protozoans.* Many parasitic and free-living protozoans (one-celled animals) exhibit a dormant stage by secreting a protective cyst. The stimulus for cyst formation in free-living protozoans may be temperature changes, pollution, or lack of food or water. *Euglena*, a protozoan that encysts to avoid environmental extremes, apparently has two kinds of cysts. Apparently one is formed only to avoid stressful conditions; the other is formed for the same reason but also involves asexual reproduction, resulting in a cyst that may contain up to 32 daughter organisms, which emerge under proper environmental conditions.

The role of cysts in parasitic protozoans

Free-living protozoans form cysts around themselves to avoid environmental extremes, but cysts are a part of the life cycle of parasitic protozoans. The causative agent of amebic dysentery, *Entamoeba histolytica*, is found in the intestine of infected individuals, in whom it forms cysts that pass to the outside in feces. When food or water containing cysts enters the digestive tract of another person, the amoebas are released from the cysts and infect the new host. Without encystment, which allows the organism to live in a dormant state in an unfavourable environment (*e.g.*, water), amebic dysentery could be much more easily controlled. Protected by the cyst wall, however, the dormant contents of the cyst can survive for weeks. Although they are not particularly resistant to drying, the cysts of *E. histolytica* can withstand temperatures of up to 68° C (154° F) for five minutes. They are also resistant to certain chemicals.

Invertebrates. Dormant cysts are formed during the life cycles of invertebrate parasites such as the oriental liver fluke (*Clonorchis sinensis*). The cyst stage of this organism develops in fish muscle; if the fish is eaten raw or undercooked, the encysted fluke is transferred to a new host. The encysted stage of the trichina worm (*Trichinella spiralis*), which causes trichinosis, is found in the muscle cells of hogs; it is also an invertebrate parasite in which the dormant stage is an essential part of the life cycle. When undercooked pork is eaten, the cyst wall is dissolved by digestive juices, and the worm is able to make its way into the tissues of a new host.

The cystlike forms found in many other invertebrate groups are all dormant stages that preserve the species during times of environmental stress. All freshwater sponges and some marine species survive cold or drought by forming gemmules within the body of the adult sponge. These structures, which are surrounded by a resistant covering, are released when the sponge dies and disintegrates. When conditions are appropriate, the cell mass escapes from the covering and forms a new sponge.

Rotifers are microscopic aquatic animals that produce winter eggs with thick and resistant coverings similar to protozoan cysts; the eggs may remain dormant for long periods. They can survive drought or freezing and may be dispersed by wind or carried by animals. Thus, the cyst serves not only for survival of the egg under adverse conditions but also for dispersal. Some freshwater bryozoans develop disklike buds, or statoblasts, that are surrounded by a hard, chitinous (horny) shell. These statoblasts are

the dormant structures that survive when the bryozoan dies in the fall or during a drought; they form a new bryozoan colony when favourable environmental conditions again prevail.

Among mollusks, land snails remain largely dormant throughout the day, with the soft head and foot withdrawn into the shell. During periods of drought or cold, they retreat into their shells and secrete a membrane (the epiphragm) of mucus and lime that covers the opening of the shell and resists desiccation. Slugs, on the other hand, bore into the ground and secrete a mucus mantle around themselves for protection during periods of unfavourable environmental conditions. Among the arthropods, many freshwater forms develop dormant cystlike stages that resist desiccation and allow the species to survive unfavourable periods.

Diapause in insects. Many insects undergo periods of reduced metabolic activity called diapause. Diapause, which may occur during any stage of the life cycle—egg, nymph, larva, pupa, or adult—is usually characterized by a cessation of growth in the immature stages and a cessation of sexual activity in adults. In some insects, it is a reaction to unfavourable environmental conditions; in others, such as certain moths and butterflies, diapause is a necessary stage of the life cycle. The 17-year larval and pupal periods of the cicada are examples of diapause. This form of dormancy is particularly common among insects that live in arid desert areas, where during the dry and hot summers, the insects usually hide themselves in the soil at suitable depths or under any available protective objects.

Insect survival in winter

Insects may overwinter as egg, larva, nymph, pupa, or adult; because they can stand very low temperatures, few of these forms die if the winter temperatures are within their normal range. Even rather fragile forms, such as mosquitoes and butterflies, survive in sheltered, relatively dry places out of doors. Some butterflies even survive the winter in low shrubbery, where they may be completely covered by snow and ice for three or four months. Other insects prepare for winter by constructing nests or cocoons; still others seek suitable hiding places.

Among some insect species, diapause lasts only until favourable environmental conditions return, after which the insect immediately resumes its normal activities. In other species, favourable environmental conditions alone do not break the diapause; some other stimulus, such as cold or food, is necessary. The eggs of the mosquito *Aedes vexans*, for example, remain in diapause until the damp soil on which the eggs are laid is flooded to form a pool suitable for the larvae. Although the eggs of another mosquito, *Aedes canadensis*, are laid in the same soil as those of *Aedes vexans*, they will not hatch—even after flooding—until they have been subjected to cold. Thus, when both species lay their eggs together in early summer, those of *Aedes vexans* hatch in pools formed by late summer rains, but those of *Aedes canadensis* overwinter and hatch in the spring rain pools. Not only are certain conditions required to break diapause but in some species, such as certain cutworms, a specific length of time must elapse before the stimuli are effective.

The onset of diapause depends upon a combination of environmental factors operating on the regulatory mechanisms—*i.e.*, nervous and endocrine systems—of the insect. Photoperiod and temperature influence the endocrine function of the brain, which synthesizes and secretes a substance (hormone) that controls other endocrine organs, specifically the prothoracic glands. Under the stimulation of the brain hormone, the prothoracic glands secrete a hormone called ecdysone. When stimulation by the brain hormone ceases, ecdysone is no longer secreted, and, in its absence, all insect growth and metamorphosis are halted. Thus, provision is made for the overwintering of immature insects in a state of developmental standstill. With the arrival of more favourable conditions, ecdysone is again secreted, and development resumes. Because many insect species have more than one generation of progeny per year, the prothoracic glands do not cease functioning except at some stage in the life cycle of the brood that must overwinter.

DORMANCY IN COLD-BLOODED VERTEBRATES

Two kinds of dormancy can be distinguished in vertebrates on the basis of body temperature. Most vertebrates are poikilothermous, or cold-blooded, because the body temperature follows that of the environment and is not kept constant by internal (homoiostatic) mechanisms. The second group, the homoiotherms, maintain a constant body temperature regardless of the environmental (ambient) temperature; called warm-blooded animals, they include birds and mammals.

Fishes and amphibians. The metabolism of poikilothermous animals is most influenced by the environmental variables of temperature, nutrition, and photoperiod. Photoperiod, the daily length of light exposure, has a marked metabolic effect in both fishes and amphibians; fishes, however, remain active throughout the year, although the activity may be limited by temperature, as in those fish that rest on the bottom or in mud during cold periods. Brief superficial freezing and supercooling (without freezing) to temperatures below the freezing point of body fluids are experienced by resistant species, but it has not been established that fishes that have been frozen solid can become active when thawed. In the Arctic, no fishes are found in lakes that freeze solid in the winter. Because most fishes do maintain some kind of activity year round, they cannot be said to become dormant in the sense in which the word is used in this article.

In addition to light and temperature, another environmental stress imposed upon fish is drought. Lungfishes, as represented by the African lungfish (*Protopterus*), burrow deeply into the mud when their water supply is diminished. They surround themselves with a cocoon of slime and remain inactive, using a lunglike air bladder for respiratory purposes because their gills are nonfunctional during this period of dormancy. They rely on fat reserves as an energy source, and in order to conserve water, they excrete urea rather than ammonia. This is because ammonia as an excretory product is highly toxic; animals that excrete ammonia require large quantities of water to dilute it below toxic levels. Urea is a semi-solid substance of low solubility, and requires little or no water for its excretion. (Desert animals and many insects excrete urea.)

Protective niches

During periods of drought or cold, amphibians seek protective niches (*i.e.*, habitats that supply the factors necessary for existence) in which to remain dormant until the return of favourable environmental conditions. Overwintering of frogs and salamanders frequently involves their aggregation in large numbers in a moist terrestrial niche, such as a rotting log, the mud on banks or bottoms of marshes and ponds, or in springs. The more terrestrially oriented amphibians, such as toads, may pass the winter in solitary burrows on land. During dry seasons, frogs may be dormant in a mud cocoon.

Reptiles. *Effects of temperature.* Because reptiles depend on external sources of heat to keep warm, they survive during periods of low temperature by seeking a place where the temperature will not fall below freezing, except temporarily. The commonest niche for reptilian dormancy is almost always found underground at a depth dependent on the thermal conductivity of the soil relative to the minimum temperature reached. This factor alone can control the distribution of reptiles. None can survive in the Arctic or Antarctic in places in which the subsoil is permanently frozen; and relatively few can exist in areas near these regions, even if suitable sites for dormancy were available, because the short summers would prevent the completion of life cycles. Although the distribution of snakes at high latitudes or altitudes is limited, the adder has been found at 10,000 feet (3,300 metres) in the Swiss Alps and about as far north as the Arctic Circle. The Himalayan pit viper has been found at an altitude of 16,000 feet (5,000 metres).

Dormancy in reptiles may display a circadian rhythm, a seasonal one, or both; it is a state of torpor directly induced by low temperature. When the adder, for example, experiences temperatures of about 8°–10° C (46°–50° F), it begins to search out suitable niches in which to rest. Its dormancy ends on the first sunny days after the maxi-

mum temperature has reached 7.5° C (45.5° F). Because these conditions vary, the adder's period of dormancy extends from 275 days in northern Europe to 105 days in southern Europe and is about two weeks in the United Kingdom, where the Gulf Stream provides warmth.

Reptiles also normally become dormant during the hottest parts of summer, but the physiology of summer dormancy is quite different from that of winter. As already mentioned, winter dormancy is a state of torpor, induced by a low temperature, that becomes more pronounced as the temperature falls. There is, however, a wide range between the animal's normal, active (coenothermic) temperature and the lowest temperature at which it can exist. At high temperatures, on the other hand, there is a much narrower range between the coenothermic temperature and temperatures that cause death. In other words, reptiles can tolerate colder temperatures much better than they can tolerate higher ones. For this reason, during hot weather they must seek refuge underground or in cool, shady places, where they remain physiologically active but must forego all normal activity because of the restricted nature of the cooler niche. Desert reptiles, in particular, exhibit such temperature responses daily.

During its dormancy, the amount of water needed by a reptile is less than at other times and is normally supplied by water produced from the metabolism of the animal's own stored food reserves, particularly fat. In areas in which alternating wet and dry seasons occur, reptiles maintain a longer period of dormancy during the dry season. This behaviour may be related more to the lack of available water than to temperature, because in such areas the onset of the seasonal monsoons elicits a period of increased reptile activity.

Winter dens of reptiles

Because there is only a limited number of suitable sites available for dormancy, several snakes, usually of the same species, may be found in each niche. As many as 100 or more snakes have been taken from one winter den. Occasionally, lizards and toads may also be found in the same den, but stories of snakes that share denning sites with small birds and mammals have been difficult to substantiate. It is much more usual to find that the entry of snakes into the burrow of a prairie dog or some other warm-blooded animal is followed by the evacuation of the original occupant.

Effects of latitude. Changes in latitude not only alter the lengths of the dormant and active periods of reptiles but also affect their circadian rhythms because of the changes in the proportions of night to day. Many species of snakes, including the adder, are normally active in the early evening. In the northerly latitudes (*e.g.*, northern Europe, such as Scandinavia and Finland), where the length of the active season is reduced by as much as two-thirds, these snakes become active throughout the day to take advantage of every warm hour in order to complete the necessary portions of their life cycle. Even this increased activity during the shorter summer season, however, does not compensate for the latitude. Growth and development slow to such a point that sexual maturity is delayed, and the reproductive period requires two years rather than one; young are produced only every other year instead of every year, as at lower latitudes.

DORMANCY IN WARM-BLOODED VERTEBRATES

Hibernation and estivation

The term hibernation is often loosely used to denote any state of torpor, inactivity, or dormancy that an organism might exhibit. Properly speaking, however, use of the term should be confined solely to warm-blooded homoiotherms; *i.e.*, birds and mammals whose feathers or fur serve as insulation to reduce heat radiating from the body and aid in the maintenance of constant body temperatures, which normally are independent of those of the environment. Because warm-bloodedness gives animals an internal physiological stability, they are less dependent on many environmental restrictions, particularly those limitations imposed on organisms by ambient temperatures. For example, only two species of reptiles are found north of the Arctic Circle, but great numbers of birds live and breed there. Warm-bloodedness also signifies a high metabolic rate, a factor that undoubtedly influences normal learning, which depends heavily on the frequency and recency of experiences. Because periods of lowered metabolism interrupt continuous learning experiences, they may explain in part why birds and mammals are so much easier to train than any other animal. The benefits of warm-bloodedness require the expenditure of large amounts of energy through the year and make a heavy demand on available food supplies.

The term hibernation is also used to delineate the dormant state only during winter. In arid regions a reverse phenomenon is seen in which the animal becomes torpid during the hot, dry, barren summer; such hibernation is called estivation. As a means of avoiding environmental stresses, hibernation and estivation are not common devices among warm-blooded animals and they are far less common among birds than among mammals.

Some warm-blooded organisms exhibit thermic instability, a heterothermous condition that allows their metabolic rate to be reduced, with a commensurate reduction in body temperature. Heterothermy is a transitional state between cold-bloodedness and warm-bloodedness; the animal is awake and moving during its temperature fluctuations. The body temperature, although not as constant as in humans, is not so low as to force the organism into deep hibernation. Among mammals, two monotremes, the spiny anteater and the duckbill platypus, are thermally unstable; many of the marsupials, including the opossum, the pouched mouse, and the native cat (a weasel-like spotted marsupial of the family Dasyuridae), are also unable to maintain a fixed body temperature.

The true hibernator not only possesses adaptations that enable it to respond as a homoiothermous animal during certain periods of the year but can also adapt to stressful environmental situations and become essentially a poikilothermous animal during other periods. An animal exposed to food shortages, low temperatures, or lack of water, for example, may "turn off its thermostat" and hibernate until the environment becomes more favourable. Unlike poikilotherms, however, hibernators still retain a measure of temperature control and can change their metabolic levels as required. They can arouse themselves to full activity, whatever the environmental temperature, whereas the arousal of a poikilotherm is dependent upon increased environmental temperatures.

Changes in preparation for hibernation

During the period prior to hibernation an animal must make a considerable number of gradual physiological and metabolic adjustments that appear to be correlated with temperature, light, and the availability of food. No one set of conditions applies equally to all hibernators: some store food, others do not; some become excessively fat, others gain a more moderate amount of weight. Generally, as the season advances and as the hibernator becomes progressively more prepared for hibernation, there is an increase of fat deposition and a general readjustment of body temperature, metabolism, and heart rate to lowered levels of activity.

Although no single factor or condition can be said to determine when an animal will go into hibernation, specific changes include an increase in the quantity of magnesium in the blood and a reduction in the activity of endocrine glands, such as the pituitary, thyroid, and adrenals. A reduction in gonadal activity has also been observed; hibernation does not occur when the gonads are in an actively functional state. Perpetuation of the species requires that the animal be warm and active during the mating and pregnancy periods.

There appears to be a relationship between sleep and hibernation; available evidence suggests that hibernation is entered into from a state of sleep. If hibernation is to be considered a form of sleep, then it must be considered a remarkably complex one. Hibernation and sleep are somewhat similar in that essential body processes continue during both periods at a lowered level. In sleep, the heart beats less rapidly, and breathing is slower; the body produces less heat, necessitating that a sleeping person be protected from the cold.

Hibernation in birds. *Temperature variations.* Birds normally have higher and more variable temperatures than do mammals. Whereas mammalian temperatures

normally range between 36° and 39° C (97° and 102° F), avian temperatures range between 37.7° and 43.5° C (99.9° and 110.3° F), with the majority between 40° and 42° C (104° and 108° F). Although the nesting temperature of most passerine species (perching songbirds) is about 40.5° C (104.9° F), primitive bird species—like primitive mammals—have lower temperatures than do the more advanced species. The kiwi, for example, has an average coenothermic body temperature of 37.8° C (100° F). In general, the temperatures of small birds fluctuate more than do those of large birds. The temperature of a house wren (*Troglodytes*) may fluctuate 8° C (14° F) in 24 hours, that of a robin (*Turdus*) fluctuates about 6° C (11° F), and that of the domestic duck only about 1° C (2° F).

The circadian period of activity and rest in birds is accompanied by a temperature cycle. Birds active in the daytime have their highest temperatures late in the afternoon and their lowest in the early morning. Nocturnal species, however, such as owls and the kiwi, have their maximum body temperatures at night, when they are most active. Seasonal temperature variations are also found in birds, and, like mammals, certain birds exhibit thermic instability. Although some are capable of maintaining a highly stable body temperature, others have a fluctuating body temperature. A torpid poorwill (*Phalaenoptilus nuttallii*) is an example of a bird that demonstrates both thermic instability and true hibernation. Its coenothermic body temperature is relatively constant; it can, however, through the influence of a thermoregulatory centre (the hypothalamus) in the floor of the brain, become essentially poikilothermous. Under such influence, its body temperatures approximate those of the environment.

Energy conservation. Considering that hibernation and estivation are devices to avoid such factors as stressful extremes of temperature, lack of water, unavailability of food, or lessened photoperiod, they also must be energy-conservation devices for the animals concerned. Even short periods of torpor can conserve energy. The efficiency of this energy-conservation system can be demonstrated by comparing the smallest bird, the hummingbird, which exhibits circadian torpor, with the shrew, the smallest mammal, which remains active throughout a 24-hour period. Oxygen consumption is an indicator of metabolic rate, and at an environmental temperature of 24° C (75° F) during the day, an awake but resting hummingbird consumes about 14 millilitres of oxygen per gram per hour. At dusk, the rate drops first to a sleeping level and then plunges to a torpid level of about 0.8 millilitre of oxygen per gram per hour. Just before daybreak, the bird awakens for another activity period. The hummingbird has the highest metabolic rate and the greatest metabolic range of any vertebrate. The shrew, in contrast, consumes about the same amount of oxygen as the hummingbird does during the day and even increases the amount slightly at night.

Oxygen consumption

The hummingbird uses about 10.3 calories (units of heat energy) during each 24-hour period if it sleeps at night without becoming torpid but only 7.6 calories if it becomes torpid. As it wakes from the torpid state, its temperature increases about 1° C (2° F) per minute to a maximum; the entire process takes less than 30 minutes and sometimes as little as ten minutes. The energy required to warm the tissues of the hummingbird is relatively small; a hummingbird that weighs four grams expends only 0.114 calorie to warm its body from 10° to 40° C (50° to 104° F). This is only 1/85 of the total 24-hour expenditure of energy of a hummingbird in nature.

The behaviour of the hummingbird can be contrasted to that of a larger bird, such as the poorwill, which is a nocturnal, insect-catching bird. During an average 24-hour day, the poorwill has brief periods of activity at dusk and just before dawn, the total of which is scarcely more than an hour. The temperature of the poorwill during these periods of activity, which are correlated with the bird's feeding habits, is between 40.5° and 43.1° C (104.9° and 109.6° F). Between periods of activity, the bird rests quietly, and its body temperature drops 1° to 3° C (2°–5° F). During periods when a supply of flying insects is not available, the bird hibernates in depressions in rocks or other suitably protected places, to which it returns each year. When hibernating, the bird's temperature is frequently within 1° C (2° F) of that of the environment; as a result, the energy saved is great. A poorwill whose body temperature is 5° C (41° F) has a metabolic rate only 3 percent of that at normal body temperature. Because the poorwill is a larger bird than a hummingbird, it may take more than an hour for it to emerge from hibernation.

Hibernation in mammals. It takes longer for larger animals than for smaller ones to go into hibernation because heat must radiate from the body before the temperature can be lowered. Thus, it would require a considerable amount of time for large birds or mammals to go into and emerge from hibernation each day, as do bats and hummingbirds. A 200-kilogram (440-pound) bear, for example, would need 5,100 calories to warm from 10° to 37° C (50° to 99° F). Unlike the hummingbird, which uses only 1/85 of its total daily energy expenditure to emerge from hibernation, the amount expended by a bear would be equivalent to its full 24-hour energy budget. Even if there were enough time in 24 hours for a large animal to enter into and emerge from dormancy, therefore, it would be metabolically extravagant, thus defeating the purpose of hibernation.

Actually, the most common misapplication of the term hibernation is in relation to the bear, which is not a true hibernator. Its body temperature, which normally averages 38° C (100° F), drops during its winter lethargy to about 34° C (93° F), seldom getting below 31.2° C (88.2° F). Hence, a bear's temperature during the winter does not approximate that of the environment. This is indicative of winter rest rather than true hibernation. During this inactive period, the bear sleeps but is, nonetheless, warm and capable of activity when stimulated, unlike a true hibernator. Moreover, it is also during this period when females give birth to cubs that suckle and are maintained by maternal warmth until they emerge from the den in the spring. This behaviour is in contrast with that of the Arctic ground squirrel, whose normal temperature is the same as that of the bear but whose temperature during hibernation drops to near freezing and, in some cases, to a degree or two below 0° C (32° F).

Although certain mammals are said to hibernate, they do not necessarily enter a state of deep hibernation during winter. Instead, for weeks at a time they may be inactive and lethargic in behaviour, with a slightly depressed body temperature. The chipmunk (*Eutamias*) is an example of what has been termed a shallow hibernator, as are bears and raccoons. Superficial hibernation, apparently a compromise between the minimum energy requirements of a deep hibernator and the high energy expended by an animal that remains active during the winter, saves energy without the stress of hibernation. The animal can thus conserve food, while still being able to escape from predators and such dangers as flooding of its burrow. The main energy source during the winter in this shallow hibernation state is food stored in the winter nest. There are instances, however, of shallow hibernators, such as the chipmunk, that enter a state of deep hibernation, particularly if without food.

Superficial hibernation

True mammalian hibernation. Omitting the thermally unstable mammals, the true mammalian hibernators are those whose lowered body temperatures approximate that of the environment and those who require extensive and complex physiological changes to turn from a warm-blooded animal into an essentially cold-blooded one for an appreciable length of time. Only three orders of placental mammals display such behaviour: the Insectivora, as exemplified by the hedgehog; the Chiroptera, the bats; and the Rodentia, including the marmot, hamster, dormouse, hazel mouse, and ground squirrel.

A typical mammalian hibernator, such as the Arctic ground squirrel, finds a protected environmental niche—in this case, a burrow beneath the surface—and builds a nest of grass, hair, and other materials to provide still

further insulation. The usual hibernating position is one of being curled up in a ball with the extremities tucked tightly against the body so there is a minimal surface-to-volume ratio. After the temperature of the animal has dropped near that of the ambient temperature, it appears to be dead: its respiration is imperceptible, about three irregular breaths per minute; it does not react to outside stimuli in an observable fashion; nor does it react to being handled and uncurled, although such handling will trigger wakening mechanisms.

The internal organs, such as the digestive tract and the endocrine glands, are almost totally inactive. Because the process of hibernation necessitates the mobilization of all body resources, it places great demands on the tissues, all of which are directed toward the problem of maintaining the animal's metabolism at the minimal level necessary for life during the hibernating period. This means that all activity not immediately germane to the process of living at the lowest possible metabolic level ceases. Even bones and teeth deteriorate during hibernation. The hibernator apparently is balanced on a very narrow line between the maintenance of life at a level that makes recovery from hibernation possible and a reduction of metabolism to a level that will lead to death. Evidence obtained from tissues indicates that the process of hibernation is a precarious method of survival at best and one from which many animals do not awaken. As a mechanism of species survival, hibernation seems effective; for the survival of the individual, however, it is an uncertain and dangerous process.

The hibernator does not remain in a continuous state of hibernation from the time it enters in the fall until it emerges in the spring. Hibernating Arctic ground squirrels, for example, awaken at intervals of every three weeks or less. During this time the animals may move about and sometimes emerge from the burrow. These periods of arousal are more frequent at the beginning and end of a hibernation period than in mid-hibernation; and the lower the temperature at which an animal hibernates, the fewer the awakenings.

During the period of hibernation about 40 percent of the total body weight is lost, an average of about 0.2–0.3 percent per day. One period of arousal and wakefulness consumes more heat and energy than many days in hibernation. About 90 percent of the total heat production and weight loss during hibernation takes place during the arousal periods; only 10 percent is required to maintain the animal in hibernation. Thus, in the case of an unusually long or hard winter, the animal may be called upon to use all of its available energy sources in periodic arousals; it then enters one final hibernation period from which it does not awaken. Animals that store food in the nest have a chance to renew their energy requirements by eating when they awaken periodically.

Types of mammalian hibernators

Entrance into hibernation. Hibernating mammals can be divided into four major groups according to the way they enter hibernation. One group is exemplified by the golden hamster; it waits a variable time of from one to three months in the cold and then enters hibernation in one major temperature reduction. This is accomplished when the biochemical and physiological preparations have been sufficient to lower the animal to a level at which it is receptive to the hibernating stimulus, which causes the abandonment of the temperature differential between ambient and body temperatures.

A second group, of which the pocket mouse (*Perognathus*) is an example, prepares for hibernation relatively rapidly, waiting only a few days before becoming torpid in one major temperature decline. The third group, which constitutes most of the mammalian hibernators, includes ground squirrels and marmots. These animals wait only a few days before entering hibernation and then go through a series of steps of torpor and arousal, each one at successively lower body temperatures, until the level dictated by the stage of preparation for hibernation is reached.

The fourth group, which includes most of the bats, becomes inactive in the poikilothermous manner; that is, the body temperature follows the ambient temperature. Even though the bat seems ready to hibernate at any season, survival during hibernation depends upon more adequate preparation than is necessary for the transitory periods of torpor. Bats not only exhibit true hibernation during the winter but also have natural periods of hypothermia (subnormal temperature), which are unrelated to hibernation, during the remainder of the year.

The woodchuck, the dormouse, and the California ground squirrel enter hibernation in successive stages, with a complete or nearly complete awakening between each one. In the woodchuck, an initial decline in temperature is followed by an arousal. During the second decline there is a lower and more pronounced fall in body temperature, followed by a less pronounced rise. This process continues until the body temperature is essentially the same as that of the environment.

Heart rate and circulation. The body temperature of a hibernating mammal is affected by changes in respiration, heart rate, and oxygen consumption; all are apparently mediated by a part of the nervous system. The heart rate decreases prior to a decline in body temperature. In the woodchuck, the rate may drop from 153 to 68 heartbeats per minute within 30 minutes. In the California ground squirrel, the heart may beat as slowly as once a minute at 5° C (41° F). In contrast, the hearts of non-hibernators generally will not beat at all at temperatures below 10°–20° C (50°–70° F).

Preparation of the Arctic ground squirrel for hibernation

As an Arctic ground squirrel prepares for hibernation, its heart rate and its blood pressure decrease. Both may be detected before a decrease in body temperature can be noted. When the animal enters hibernation, temperatures of both the heart and abdominal regions are identical, indicating an even blood flow between the anterior (front) and posterior (rear) parts of the body. As the body temperature drops, the resistance to blood flow in the peripheral parts of the circulatory system increases because of the increased viscosity (resistance to flow) of the chilled blood and the constriction of the distal arterioles (small arteries) of the body. This peripheral resistance maintains blood pressure at relatively high levels in the deeply hibernating squirrel, even when the heart beats only three or four times a minute.

Neural changes. The nervous system of hibernators also is acclimated; certain specific structures and pathways are seemingly maintained to regulate and coordinate metabolism as temperatures drop. This adaptation of the nervous system enables changes in the environment to be perceived, even when the animal is torpid. In the Arctic ground squirrel, measurements of the general electrical activity of the brain indicate a 90 percent reduction when the animal is in hibernation, at which time brain temperatures approximate 6° C (43° F). During hibernation, both the peripheral nervous system (all the nerves outside the brain and spinal cord, which constitute the central nervous system) and the spinal cord have an increased sensitivity to certain stimuli; in addition, the areas of the brain that regulate temperature as well as cardiac (heart) and respiratory function remain active at ambient temperatures, below which the mammalian nervous system normally ceases to function.

Changes in the circulatory system involving constriction (narrowing) of posterior vessels and the favouring of anterior circulation allow the brain temperature of hibernators to remain a few degrees warmer than the environmental level. This enables the temperature of the brain to remain constant despite fluctuations in the temperature of the skin.

Endocrine activity. The male sex hormone testosterone stimulates reproductive activity. The golden hamster will not hibernate if injected with more than five milligrams of a hormonal preparation. Hibernation is also prevented if the animal is fed or injected with thyroid hormones or thyroid-stimulating extracts. The latter would seem to implicate the thyroid as another endocrine gland that plays an important role in hibernation. There is, in fact, a seasonal progression and regression of thyroid activity in hibernators; maximal activity occurs in the spring and minimal activity in the fall. And because hibernation does not take place in the absence of the

adrenal glands, it appears that a minimal adrenal activity is also necessary for hiberation and survival.

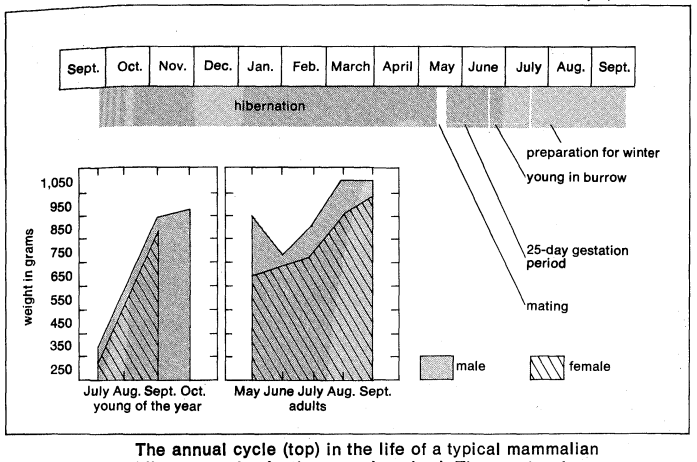
Importance of timing

The importance of timing in the annual rhythm of activity and dormancy can be demonstrated: when hibernators are exposed to cold temperatures in spring and summer, they react as do all homoiotherms by increasing their thyroid activity and metabolic rate to maintain normal body temperature. But if they are exposed to cold temperatures in the fall, the thyroid activity and metabolic rate of hibernators are lowered. In some species, a combination of decreased food and lower ambient temperature is required to reduce activity of the thyroid gland and to produce hibernation, although cold alone is sufficient in ground squirrels and the dormouse.

Although hibernation does not take place during periods of gonadal activity or stimulated thyroid activity, it can occur during increased activity of the pituitary gland. This would suggest that there is a dissociation of cellular growth and hormone synthesis that is normally controlled by hormone secretion of the pituitary and its target organs. Thus, the triggering mechanism for the resumption of normal endocrine activity apparently resides elsewhere than in the pituitary. The function of the hypothalamic region of the brain in regulating appetite, fat deposition, water intake, and diuresis (increased excretion of urine), as well as in the control of temperature and sleep, would appear to make it a key area in directing life processes of the hibernator. Furthermore, the fact that the hypothalamus regulates the pituitary and other endocrine glands not only supports this thesis but also indicates that this area of the brain is the prime, or master, regulator of the entire hibernation process.

Reproductive cycles. The Arctic ground squirrel may spend more than half its life in hibernation (see Figure).

From W.V. Mayer, *Hibernation*



The annual cycle (top) in the life of a typical mammalian hibernator, the Arctic ground squirrel. The graphs show typical weight patterns of young and adult males and females during the aboveground activity period (see text).

It thus must be able to breed, rear young, maintain its home burrow, and prepare for the period of hibernation during an activity period of less than six months. This requires considerable adaptation of both metabolic and behavioral patterns. Prior to entering hibernation in late September or early October, there is a renewal of sexual activity in the testes of males, and, throughout the period of hibernation, they continue to grow. On the Arctic slope in early May, the male ground squirrel emerges from its burrow. As it utilizes the remaining fat and eats the stores of seeds and other food still in the nest, the male reaches a period of reproductive readiness. Mating takes place in the middle of May, and the young are born in the middle of June, after a gestation period of about 25 days. By the middle of July the young are above ground and eating the green Arctic vegetation, which they continue to eat until the onset of hibernation. By October, both the young of the year and the adults from the previous year weigh nearly 1,000 grams.

In the bat, the reproductive cycle is interrupted by hibernation. Gonadal activity in the male reaches its maximum in the fall, when copulation with the female occurs. The animals then hibernate, and the production of sperm in the male ceases. The sperm deposited in the female are stored in her reproductive tract throughout the period of hibernation; fertilization occurs the next spring, when the eggs are ovulated (released from the ovaries) within a few days after awakening from hibernation.

Delayed fertilization

The only exception to the general hibernation–reproduction pattern of bats is the vespertilionid bat (*Miniopterus*), in which there is no delayed ovulation and fertilization. In this species the eggs are ovulated soon after copulation, in the fall, and fertilized immediately. During the ensuing period of hibernation embryonic development is initiated and slowed, but it does not actually cease. The young are born in the early summer, soon after hibernation ends. The introduction of hibernation during pregnancy makes the gestation period several months longer than in non-hibernating tropical members of the same genus.

Cyclical reproductive activity has thus become adapted to the shortened activity season available to the hibernator. But although the annual sequence of reproductive events is known, the external stimuli that regulate the reproductive cycles of bats and other hibernators are not known. More knowledge is needed concerning the endocrine and nervous mechanisms that presumably regulate reproductive processes internally. It has been suggested that the pituitary–gonadal relationship influences the hibernating cycles as well as the reproductive cycle, hence both the latter and homoiothermism are controlled by a common mechanism. Such a suggestion is attractive in that the mechanism solves the regulation problems, but more needs to be known of the way in which hibernation directly or indirectly modifies the action of endocrine and neural mechanisms that direct the reproductive cycle.

Protection from disease and radiation. Hibernating organisms have a certain degree of resistance to infectious diseases that appears to be attributable to at least three factors, all of which are related to temperature. One is the fact that the lowered temperature of the host and the commensurate slowing of its metabolic processes prevent the multiplication of parasites to a greater extent than they affect the host's defensive mechanisms. Second, lower temperatures are more harmful to the development of a disease organism than to the host, as has been shown with the parasite *Trichinella spiralis*. In bats hibernating at 5° C (41° F), only larvae have been recovered from the intestines; but mature adult worms have been recovered from the intestines of bats kept at 35° C (95° F). The third factor is that the influence of low temperature on the chemical composition of the host tissues may also affect infectious organisms.

Hibernation also seems to protect animals from radiation. When ground squirrels are irradiated with radioactive cobalt while hibernating, they are found to be more resistant to the effects of the radiation than are squirrels irradiated while warm and active. This resistance, which is apparent over a wide range of doses, suggests that protective mechanisms function in the hibernating animal. In both hibernating and non-hibernating animals, repair processes within cells occur the first day after irradiation; however, when the metabolic requirements of cells are small, as in hibernation, the injured cells seem to be more capable of repair, and survival is greater. The large metabolic requirements imposed on injured cells of warm and active animals appear to render them incapable of an adequate repair response.

Awakening from hibernation. The process of awakening in the Arctic ground squirrel takes about three hours. There is a rapid rise in heartbeat and a decrease in peripheral circulatory resistance; the area around the head and heart warms more rapidly than the posterior part of the animal. This differential vasodilatation (widening of the blood vessels) in the anterior part of the body is a unique and vital part of the awakening process. The concentration of active circulation in this region results in a high blood pressure and an efficient and rapid warming. If a drug is administered during awakening

that causes vasodilatation throughout the body, there is a marked drop in blood pressure even though the heart may almost double its rate; thus, the heart cannot maintain a high blood pressure at this time if all blood vessels are dilated. Later during the arousal process, after the anterior part of the body has been warmed, the posterior part of the animal warms rapidly.

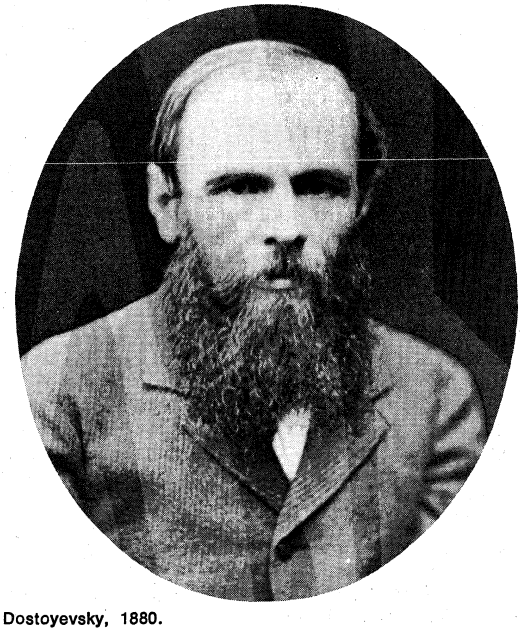
Despite the deterioration of glands and tissues and the drastic reduction of all metabolic activity during hibernation, within 24 hours after arousal, all the squirrel's physiological processes are essentially normal. This rapid repair and recovery mechanism is one that requires further study.

BIBLIOGRAPHY. H.W. WOOLHOUSE (ed.), *Dormancy and Survival* (1969), a symposium incorporating a wide variety of review articles covering dormancy in a broad spectrum of organisms from bacteria through mammals; K.C. FISHER (ed.), *Mammalian Hibernation: Proceedings of the Third International Symposium on Natural Mammalian Hibernation* (1967), 25 papers dealing directly or indirectly with the topic of mammalian hibernation accompanied by over 1,300 bibliographical references relative to this topic; X.J. MUSACCHIA and J.F. SAUNDERS (eds.), *Depressed Metabolism* (1969), a variety of studies relative to hibernation, hypothermia, and thermic instability; W.V. MAYER, "Hibernation" (1964), a popular pamphlet concerned with hibernation in both birds and mammals; J.P. HANNON and E. VIERECK, *Comparative Physiology of Temperature Regulation* (1962), deals with temperature regulation in both cold- and warm-blooded animals, including hypothermia and hibernation; C. KAYSER, *The Physiology of Natural Hibernation* (1961), an intensive look at hibernation in birds and mammals, including hypothermia and estivation, with emphasis on functional changes; C.P. LYMAN and A.R. DAWE (eds.), *Mammalian Hibernation* (1960), hibernation and hypothermia in birds and mammals, including articles on such thermally unstable forms as bears.

(W.V.M.)

Dostoyevsky, Fyodor

Dostoyevsky's *Crime and Punishment*, *The Idiot*, *The Possessed*, and *The Brothers Karamazov*, in all of which art is used as a medium for conveying the wisdom of life and the emotions of the soul, have won him a reputation as one of the world's greatest novelists.



Dostoyevsky, 1880.
By courtesy of the Literary Museum of the Institute of Russian Literature, Leningrad

Early life and literary beginnings. Fyodor Mikhaylovich Dostoyevsky was born in Moscow on November 11 (October 30, old style), 1821. His father, a former army surgeon, was a stern man and rigid in domestic matters. Years later, the son, mindful of his insecure, middle class family, called himself an "intellectual proletarian." The background that influenced his literary interests was entirely different from that of his rivals, Ivan Turgenev and Leo Tolstoy, who were members of the cultured gentry class. Finishing his early education in Moscow, Dostoyevsky entered the military engineering school of St. Petersburg (now Leningrad) at 16. What time he could steal away from drill and the science of fortifications, he spent reading Russian and other European literature, especially melodramatic fiction that encouraged his taste for plots of violence and crime. Little is known about these formative years, but he appears to have enjoyed nights out with his fellow cadets, good food and drink, interesting conversation, music, the theatre, and the company of girls. He was also a passionate dreamer about fame, self-sacrificing deeds, and idealistic friendships.

Shortly after graduating, Dostoyevsky took the bold step of resigning his commission to devote all his time to writing. He had hardly any financial means—his mother had died, and his father, who had been murdered by serfs, left little money. He had, however, already finished the manuscript of a short novel, *Poor Folk* (1846), and with misgivings he allowed a young friend to submit it to a renowned literary critic, Vissarion Belinsky. Many years later Dostoyevsky recalled the ecstasy he experienced when the critic summoned him and praised the artistic instinct with which he had revealed the hidden nature of his hero. "The truth has been revealed and announced to you as an artist, it has been brought as a gift; value the gift and remain faithful to it, and you will be a great writer!"

Poor Folk is hardly a major effort and is marred by a beginner's technical faults, but Belinsky's praise was prophetic. He had recognized in the work the first Russian social novel, for it tells of an impoverished, elderly clerk's hopeless struggle for respectability while concealing his love for an orphaned girl in a sentimentally expressed, paternal affection. An uncommon insight into the tragic futility of poor people in love is revealed, people victimized by cruel circumstances of contemporary society. Dostoyevsky's handling of the subject won an enthusiastic response from readers, for he added a new dimension—an intense psychological interest in which the hero's conflicts are studied from within. In a letter to his brother Mikhail, he explained his approach: "I proceed by analysis and not by synthesis, that is, I plunge into the depths, and, while analyzing every atom, I search out the whole." In fact, he had begun his own school of Russian realistic fiction.

First novel's psychological foreshadowings

In contrast, the youthful author did not create an impressive personal image in literary circles and society salons, where he was lionized after his initial success. This short, fair-haired man with small, gray eyes, sickly complexion, and nervously twitching lips was awkward in his movements and ill at ease in such social situations. He took refuge in his writing and quickly produced another short novel, *The Double* (1846). This literary exploitation of the "split personality" (Golyadkin, a minor civil servant, suffers growing persecution mania, leading him to encounter another man looking exactly like him who is the leader of a conspiracy against him) bored readers and cost him Belinsky's critical support. In its less clinical aspects, however, the "double," the man or woman of divided self, played a significant role among the characters of his famous novels.

A series of sketches, short stories, and another short novel, published between 1846 and 1849, received little notice. He might have won back his audience with an ambitious full-length novel that he planned, *Netochka Nezvanova* (1849), the story of a young girl's love for her wayward stepfather, a preview of ideas, images, and devices that would repeatedly appear in later works; but only three lengthy episodes had been published when his arrest for alleged subversion brought an end to its writing, as well as to this first literary period. These early tales reflect both literary influences and his own observations of St. Petersburg life. Psychological and spiritual self-examination, however, also entered into the intense analysis of the thought and feelings of his characters. If only in a fugitive manner, this first period suggests the main direction of his future creative development.

Exile to Siberia. Caught up in the movement for political and social reform during the repressive rule of Tsar Nicholas I, Dostoyevsky had participated in weekly discussions at the home of an idealist, Mikhail Petrashevsky, in which the ideas of French utopian Socialists were debated. There is also evidence that he took part in secret meetings of a smaller inner group that planned to print illegal radical pamphlets. The government, fearful of revolutionary contagion from the West, ordered the arrest of members of the Petrashevsky Circle in April 1849. After a long investigation, 21 of them, including Dostoyevsky, were condemned to be shot. Memory of the experience during the grisly preparations for the execution, before the Tsar's commutation was announced, haunted the pages of his later fiction. His sentence was changed to four years of hard labour in a prison at Omsk, in Siberia, and thereafter four more years as a soldier in the ranks.

Radical leanings

Dostoyevsky accepted his punishment as a necessary atonement for what he believed a serious crime. He came to regard as "extraordinary people" many of the simple convicts among whom he lived in chains, stench, and hard labour; but spiritual agony overwhelmed him at times, and he dates his first attack of epilepsy, an affliction that persisted for years, to this period. The only book allowed in prison was the New Testament, which he read frequently. It assuaged the bitterness of prison and taught him a new faith in Christ, who alone could raise up the sinner and promise the humble of heart a new existence. His prison experience was important for his future development as a writer and thinker. Youthful radicalism gave way to respect for established order and belief in the messianic mission of the common people; Christ's teaching, in the sense of salvation by suffering, and the spirituality of the Russian Orthodox Church took on deeper meaning; and prison provided him with rich material for further study of the insulted and injured.

After his release in 1854, he found serving as a soldier in the Siberian town of Semipalatinsk more boring, in a sense, than prison. He worked hard at his duties, eventually became a junior officer, made a few friends, and repeatedly begged his brother to send him books and periodicals to repair the huge gaps in his reading. The only event of consequence was his marriage in 1857 to a consumptive widow with one son, which turned out to be an unhappy union. Added financial responsibilities intensified his desire to resume writing and regain his literary position after years of enforced silence. He began with a comic short story that had no connection with various designs for fiction he had thought out in prison. "Uncle's Dream" (1859) is an amusing tale, a satire written in the manner of the earlier Russian novelist Nikolay Gogol on the snivelling society of a provincial town for which Semipalatinsk must have been the model. This story was quickly followed by a more ambitious short novel, *The Friend of the Family* (1859), the main character of which, Opiskin, another dual nature, saves the work from mediocrity. Critics paid attention to neither story. Not long after their publication, Dostoyevsky was allowed to return to his beloved St. Petersburg, a free man, just ten years after he had left it in chains.

Literary renewal. Radicals in the capital were eager to glorify him as a former political prisoner, but he spurned them and their ideas, especially their ridiculing of religion. His sympathies were on the side of the social reforms advocated by the new emperor, Alexander II. The first collected edition of his writings appeared in 1860, and the next year, with the cooperation of his brother, he began the magazine *Vremya* ("Time"). Its announced position was the ideological reconciliation of the leading intellectual groups—Westernizers and Slavophiles. Both factions were urged to unite with the masses on behalf of Russia's national salvation. His elaboration of this popular approach in journalism and fiction contributed to the magazine's success. Old and new like-minded friends, such as the poet Apollon Maykov and the critics Apollon Grigoryev and Nikolay Strakhov, rallied around him and *Vremya*, and they had considerable influence on his future political, social, and artistic views.

Founding of *Vremya*

The House of the Dead (1861–62) revived Dostoyevsky's earlier literary fame and helped the initial popularity of *Vremya*, in which it appeared. Turgenev acclaimed the work, and Tolstoy valued it as Dostoyevsky's best production. Though represented as the memoirs of a man condemned to penal servitude for murdering his wife, it is really a vivid account of Dostoyevsky's prison experiences. Striving for objectivity, he describes prison existence and offers penetrating psychological studies of unusual convicts, symbolically suggesting in moving incidents the pain of lost liberty among these human outcasts. At about the same time, he published serially in *Vremya* his novel *The Insulted and Injured*, the story of a woman's right to offer her love to a man in defiance of family and convention. It annoyed critics but delighted the reading public. At least several portrayals, such as the heroine Natasha, his first fully drawn female representative of emotional ambivalence, little Nelly, a character reflecting his deep understanding of child psychology, and the self-willed villain Valkovsky, are harbingers of more impressive characters in the major novels.

First trip abroad

By the summer of 1862, *Vremya*'s earnings enabled Dostoyevsky to realize an old dream—his first trip abroad. The visit provoked his notable magazine article "Winter Notes on Summer Impressions" (1863), in which he declared that the evils of European civilization that he had observed fortified his faith in Russia's high destiny if it could escape the poison of the West. In that year, however, the government banned *Vremya* for what it considered an unpatriotic article by Strakhov. In this crisis Dostoyevsky went abroad again on borrowed money, ostensibly for treatment of his epilepsy but really to try his luck at the gaming tables of Wiesbaden in Germany and to keep a rendezvous with Polina Suslova, a young contributor to his magazine with whom he had already become intimate. His luck with both was execrable, but continued experience with this strange woman's mixed feeling of love and hate entered into his conception of the "infernal women" of his novels.

After his return to Russia, a small legacy enabled Dostoyevsky and his brother to start a new magazine, *Epokha* ("Epoch"), in the first issue of which appeared the beginning of his extraordinary *Notes from the Underground* (1864). Though in part a satire on radical Socialists who believe that man can be governed by rational self-interest, the nameless hero is also a profound analyst of himself, a supremely alienated individual for whom no truth is absolute and every good is relative. His dualism is caused by a fundamental conflict between will and reason. The work marks an altered approach to characterization in the hero's emphasis on self-examination in which the focus is on the spiritual life of dislocated man in a real and acceptable world. In essence, *Notes from the Underground* is a philosophical introduction to the forthcoming cycle of great novels, for nearly all their motifs involving moral, religious, political, and social ideas appear in it.

Period of the great novels. Between 1864 and 1865 misfortunes overwhelmed Dostoyevsky—his wife and brother died, and the debt-ridden magazine collapsed. Threatened by debtor's prison, he fled abroad with an advance on a novel from a dubious publisher. Again, his hope was in gambling, to which he had become a passionate devotee. Perhaps with marriage in mind—he had had several unsuccessful love affairs in St. Petersburg—he had once more arranged a meeting with Polina Suslova. Deserted by her at Wiesbaden, he also lost all his money at roulette and was reduced to pawning his clothes. He begged friends for loans to pay hotel bills and return to Russia. One letter was to the editor of a periodical for an advance on still another novel, which he described as *Crime and Punishment*. Money eventually arrived, and he returned to Russia in October 1865.

The conception of *Crime and Punishment* (1866) probably dates back to Dostoyevsky's prison days. Infinite concern for its artistic details is reflected in many notebook pages on the novel's planning (such notebooks, which also exist for later novels, shed much light on his

creative methods). In part, the work is a social novel in which money is a basic problem. Related to it is the materialistic thinking of radical youth, one of whom is the work's impoverished hero, Raskolnikov. A nihilist in intellectual rebellion against society, he struggles between good and evil, reason taking the place of the living process of life; and his destructive theory that humanitarian ends justify evil means leads him to murder. In prison his intellectual pride, which had driven him to violate the moral law, gives way to a realization that happiness cannot be achieved by a reasoned plan of existence but must be earned by suffering. Such secondary characters as Marmeladov, his wife, the prostitute Sonya, and Svidrigaylov are also brilliantly portrayed. The work provides a new dimension to the suspense of the familiar murder mystery by infusing into it compelling philosophical, religious, and social elements. The novel achieved immediate success. Critics and readers were captivated by its innovations, narrative intensity, and the spiritual glow that illumines the darkest recesses of the criminal and the morally debased.

Before completing *Crime and Punishment*, Dostoyevsky recalled that he risked severe financial penalties for failure to fulfill within a month his earlier contract with the unethical publisher. He hired a young stenographer, Anna Snitkina, and dictated a short novel in time. *The Gambler* (1866) is a minor effort with several powerful scenes that were inspired by his passion for gambling and his love–hate relations with Polina Suslova. The next year he married the stenographer, and, to escape creditors and begging in-laws, they went abroad, remaining away four years. Often they lived in abject poverty, moving from country to country. His young wife endured all this, his epileptic seizures, incessant gambling, and the tragic death of their firstborn, her devotion to him and his genius never faltering. This second marriage was one of real love and the most fortunate event of his life.

Out of these harsh circumstances emerged Dostoyevsky's second masterpiece, *The Idiot* (1868–69). Its starting point was the account of a criminal trial in the Russian press. Such events he called "fantastic realism," but, in his fictional use of them, he shifted the emphasis from the external world to that of the mind and heart of his characters. Though concerned with problems of average Russians, he elevates them to universal significance in his search to find "man in man." He wrote in his notebook: "They call me a psychologist. It is not true. I'm only a realist in the higher sense; that is, I portray all the depths of the human soul." The chief idea of *The Idiot*, he wrote to his niece, "is to portray a positively beautiful man [*i.e.*, in a moral sense]. . . . There is only one positively beautiful man is the world—Christ." But human weaknesses mar the pure moral nature of his hero, Myshkin, and they condition his involvements with the Yepanchin and Ivolgin families, with Rogozhin, and with those two unusual rivals for Myshkin's love, Aglaya and Nastasya. In superb scenes these and other characters, who are given to sensuality, acquisitiveness, and crime, test Myshkin's moral feelings. Though his faith and radiant personality draw them to him, his message of service, compassion, and brotherly love fails. His experiences are symbolic of Christ's among the Pharisees. In the end the sinning people he moved by his goodness are rendered unhappy, and he himself lapses into idiocy.

Having already spent the advance on another novel, of which he had written nothing, Dostoyevsky accepted ready cash from a new publisher for a short story that turned into a short novel. *The Eternal Husband* (1870), a subtle psychological study of a betrayed husband who seeks revenge on his wife's seducer, represents no particular advance in the author's creative art. Meanwhile, his thoughts concentrated on a vast plan, a series of possibly five connected novels under the title *The Life of a Great Sinner*, the notes for which are extant. Its protagonist would be guilty of heinous crimes against God and man, but, at the end of his spiritual pilgrimage, he would redeem himself and achieve salvation. This work was never written, but from its outline Dostoyevsky borrowed ideas, scenes, and characters for his last three novels, the first of which was *The Possessed*, begun in 1869 and completed in 1872.

Sources for *The Possessed*

The main plot of *The Possessed* was suggested by sensational press reports of a Moscow student's murder by fellow revolutionists who suspected he intended to betray them. Worked into this plot are certain features and figures of the plan of *The Life of a Great Sinner*, especially aspects of the novel's central character, Stavrogin. In a work packed with action and drama, the revolutionary plotters are satirized as dolts and rascals. Their intended victim, the reformed Shatov, reflects Dostoyevsky's ideological opposition to revolution—an expression of nationalist faith in tsarist Russia's future that can have meaning only as a part of the country's faith in the Christ of the Orthodox Church. The enigmatic Stavrogin dominates the novel. His magnetic personality influences the amusing old liberal Stepan Verkhovensky and his revolutionary son Pyotr, as well as the renegade radicals Shatov and Kirilov, and he has a fatal attraction for the principal female characters Lizaveta, Darya, and Marya. With his loss of faith in God, however, the innate goodness of his nature has atrophied. His violation of the little girl symbolizes his complete submission to evil. Despite Dostoyevsky's didactic intention, the work is saved from becoming an inflated purpose novel by the sheer force of his art. Rarely had he handled with such artistic skill his favourite combination of sensational and ideological elements.

Long before he finished *The Possessed*, Dostoyevsky fell ill, desperately needed money, and insisted he could not complete the novel abroad. His worried publisher sent funds to enable him to return to St. Petersburg. The novel's success was enhanced by impressive public readings from his works, and Dostoyevsky was again in demand at social gatherings. In 1873 prominent friends helped to secure for him the editorship of the conservative weekly *Grazhdanin* ("The Citizen"). After a year he resigned because he found the work too confining and the publisher excessively reactionary. By this time his capable wife had undertaken the publication of his writings, from which a substantial profit was realized.

Editor of *Grazhdanin*

In 1876 Dostoyevsky revived as a separate monthly publication a column, "The Diary of a Writer," which he had contributed to *Grazhdanin*. He continued it for over a year, with a few additional issues in 1880 and 1881. It is devoted mostly to his views on significant current events, literary reminiscences and criticism, and occasional sketches and short stories, of which two are among his best: "A Gentle Spirit" (1876) and "The Dream of a Ridiculous Man" (1877). He also used it, however, as a medium for his striking ideas on broad social, political, and religious problems. Journalism and literature were closely allied in his mind, for he believed that the interrelation of art and reality must be based on observation of daily existence. *The Diary* attracted numerous readers and is of importance for the study of his life, philosophy, and fiction, especially of his last two novels.

In *The Diary* he describes the theme of *A Raw Youth* (1875)—the confession of an illegitimate son, Arkady Dolgoruky, about his adventures in St. Petersburg, where he seeks to win the affection of his father, Versilov, another dualist character and mouthpiece of Dostoyevsky's favourite convictions—for instance, the conviction that the Russian is unique, the perfect cosmopolitan, whereas the European intellectual stands on the brink of destruction because of his revolutionary materialism and denial of Christ. In stressing Versilov's ambivalence, Dostoyevsky for the first time clarifies his special interest in the subject. The psychological determinants of dualism as reflected in the thoughts, feelings, and actions of his imaginary men and women were observed by him in self-examination of the dualism of his own nature as well as that of others. Dostoyevsky recognized the main plot's strangulation by several subplots, telling his wife in some despair: "There are four novels in *A Raw Youth*." The novel is usually ranked below his others.

National renown

By the time Dostoyevsky began *The Brothers Karamazov* (1879–80), he was nationally recognized as a celebrated author. Important people sought him out; he

was asked to give the public eulogy at the funeral of a famous editor and writer, N.A. Nekrasov; the Academy of Sciences elected him as an associate member in literature; and his speech at the 1880 commemoration of the poet Aleksandr Pushkin electrified a distinguished audience with its ringing prophecy of Russia's world mission. He preferred, however, to live quietly at Staraya Russa, a small resort town some distance from St. Petersburg, with his wife and two children, Fyodor and Lyubov. There he followed a strict regimen of walks and writing, with his devoted wife as his amanuensis. She took down in shorthand his dictation of *The Brothers Karamazov*, the literary effort for which he had been preparing for most of his creative life. It is a story of patricide into the sordid unfolding of which Dostoyevsky introduces a love–hate struggle with profound psychological and spiritual implications. Throughout the whole novel persists a search for faith, for God—the central idea of the work. Alyosha, the youngest brother, identified with the Christian ideal, loves life more than its meaning; Dmitry loves life but its meaning evades him; Ivan, who is more concerned with life's meaning than with life itself, is the most absorbing character and the mental image of his creator. Ivan's ambivalence is centred in the cosmic struggle of man with God. He begins with an act of rebellion and ends in metaphysical insurrection against God's world. Ivan is concerned with those "accursed questions" that motivated Dostoyevsky's own search for faith—the problems of sin and suffering and their relation to the existence of God. The dramatization of Ivan's repudiation of God's world is concentrated in the famous "Legend of the Grand Inquisitor." The answer follows in the novel's next section in the preaching of the monk Zosima that the secret of universal harmony is not achieved by the mind but by the heart, by feeling, and by faith.

Dostoyevsky intended to exemplify Zosima's precepts in action in a sequel to the novel in which Alyosha would become the hero. A few months after the completion of *The Brothers Karamazov*, however, he died in St. Petersburg, on Feburary 9 (January 28, old style), 1881.

Assessment

Today, Dostoyevsky is among the most widely read 19th-century novelists, perhaps because he effectively dramatized in his fiction moral, religious, and political problems that disturbed generations between World Wars I and II and afterward. Friedrich Nietzsche, the German philosopher and poet, admitted indebtedness to him, and one pre-Nazi German critic declared that after Martin Luther there had been no greater spiritual influence in Germany. The 20th-century novelist André Malraux asserted that in France Dostoyevsky had deeply affected the intellectual history of his generation. The 20th-century French philosopher Jean-Paul Sartre paid tribute to Dostoyevsky's condemnation of the tyranny of reason as having helped inspire his own Existentialist beliefs. Though Lenin is reported to have said of his fiction, "I have no time for such trash," Dostoyevsky is widely read in the U.S.S.R., and notable Soviet novelists have been influenced by his works. If the test of a great author is his capacity to impose his vision on his readers and to transform their experiences, then Dostoyevsky's anti-heroes have had an effect on the creation of many characters in 20th-century U.S. fiction, portrayals that seem overwhelmed by the infirmity of doubt.

MAJOR WORKS

Bednye lyudi (1846; *Poor Folk*); *Dvoynik* (1846; *The Double*); *Netochka Nezvanova* (1849; *Netochka Nezvanova*); "Dyadyushkin Son" (1859; "Uncle's Dream"); *Selo Stepanchikovo i ego obitateli* (1859; *The Village of Stepanchikovo and Its Inhabitants*—known in its English translation as *The Friend of the Family*); *Zapiski iz myortvogo doma* (1861–62; Notes from the House of the Dead—familiar in English as *The House of the Dead*); *Unizhennye i oskorblyonnye* (1861; *The Insulted and Injured*); "Zimniye zametki o letnikh vpechatleniyakh" (1863; "Winter Notes on Summer Impressions"); *Zapiski iz podpolya* (1864; *Notes from the Underground*); *Prestupleniye i nakazaniye* (1866; *Crime and Punishment*); *Igrok* (1866; *The Gambler*); *Idiot* (1868–69; *The Idiot*); *Vechny muzh* (1870; *The Eternal Husband*); *Besy* (1871–72; *The Devils*—well known in English as *The Possessed*); *Dnevnik pisatelya* (1873–74, 1876, 1877–81; *The Diary of a Writer*); *Podrostok* (1875; *A Raw Youth*); "Krotkaya" (1876; "A Gentle Spirit"); "Son smeshnogo cheloveka" (1877; "The Dream of a Ridiculous Man"); *Bratya Karamazovy* (1879–80; *The Brothers Karamazov*).

BIBLIOGRAPHY. Substantial bibliographical listings may be found in various English and Russian works cited below.

Editions, journalistic writings, letters, and notebooks: *Полное собрание сочинений Ф.М. Достоевского,* 30 vol. (first four volumes dated 1972), is the most complete edition, with all texts corrected on the basis of original manuscripts and extensive commentary provided. This edition contains a great deal of hitherto unpublished material. The only edition in English that contains most of the fiction is *The Novels of Fyodor Dostoevsky,* trans. by CONSTANCE GARNETT, 12 vol. (1921–23). Many separate novels and collections of tales have repeatedly appeared in English. Some journalistic writings have also been published in English: *The Diary of a Writer,* trans. by BORIS BRASOL, 2 vol. (1949); *Winter Notes on Summer Impressions,* trans. by R.L. RENFIELD (1955); and *Dostoevsky's Occasional Writings,* trans. by DAVID MAGARSHACK (1963). The fullest edition of letters, pending completion of the Soviet definitive edition, is *Ф. М. Достоевский, Письма,* ed. by A.C. Долинин, 4 vol. (1928–59). In English selections have appeared: *Letters of Fyodor Mikhailovich Dostoevsky to His Family and Friends,* trans. by ETHEL C. MAYNE (1914); *Dostoevsky: Letters and Reminiscences,* trans. by S.S. KOTELIANSKY and J. MIDDLETON MURRY (1923, reprinted 1971); *The Letters of Dostoevsky to His Wife,* trans. by ELIZABETH HILL and DORIS MUDIE (1930); JESSIE COULSON, *Dostoevsky: A Self-Portrait* (1962). Valuable Russian editions of notebooks on the five major novels have been translated into English: *The Notebooks for "Crime and Punishment,"* ed. and trans. by EDWARD WASIOLEK (1967); *The Notebooks for "The Idiot,"* ed. by EDWARD WASIOLEK, trans. by KATHARINE STRELSKY (1967); *The Notebooks for "The Possessed,"* ed. by EDWARD WASIOLEK, trans. by VICTOR TERRAS (1968); *The Notebooks for "A Raw Youth,"* ed. by EDWARD WASIOLEK, trans. by VICTOR TERRAS (1969); *The Notebooks for "The Brothers Karamazov,"* ed. and trans. by EDWARD WASIOLEK (1971).

Biography and reminiscences: The latest and best informed biography is in Russian: Л.П. Гроссман, *Достоевский,* 2nd ed. (1965). Other works of importance include N.N. STRAKHOV, *Fyodor Dostoevsky: A Study* (1921; orig. pub. in Russian, 1883); А.П. Суслова, *Годы близости с Достоевским* (1923); ANNA DOSTOEVSKY, *Dostoevsky Portrayed by His Wife: The Diary and Reminiscences of Mme. Dostoevsky,* trans. by S.S. KOTELIANSKY (1926); Андрей Достоевский, *Воспоминания* (1930); HENRI TROYAT, *Dostoevsky* (1939; Eng. trans., *Firebrand: The Life of Dostoevsky,* 1946); E.H. CARR, *Dostoevsky* (1949); AVRAHM YARMOLINSKY, *Dostoevsky: His Life and Art,* 2nd rev. ed. (1957), a standard, up-dated biographical treatment in English; ROBERT PAYNE, *Dostoevsky: A Human Portrait* (1961); *Ф.М. Достоевский в воспоминаниях современников,* 2 vol. (1960); А.Г. Достоевская, *Воспоминания* (1971).

Criticism and interpretation: *Творчество Достоевского. Сборник статей и материалов,* ed. by Л.П. Гроссман (1921); *Ф.М. Достоевский, Статьи и материалы,* 2 vol., ed. by А.С. Долинин (1922–24); JOHN MIDDLETON MURRY, *Fyodor Dostoevsky: A Critical Study* (1923), a view of early English reaction; Л.П. Гроссман, *Поэтика Достоевского* (1925); ANDRE GIDE, *Dostoevsky* (1923; Eng. trans., 1925); М.М. Бахтин, *Проблемы творчества Достоевского* (1929); JANKO LAVRIN, *Dostoevsky: A Study* (1943); NIKOLAI BERDYAEV, *Dostoevsky* (1957; orig. pub. in Russian, 1923), a philosophic review of his works; VYACHESLAV IVANOV, *Freedom and the Tragic Life: A Study in Dostoevsky,* 3rd ed. (1960; orig. pub. in Russian, 1932), a symbolic and mythic study; R.E. MATLAW, *The Brothers Karamazov: Novelistic Technique* (1957); VLADIMIR SEDURO, *Dostoyevski in Russian Literary Criticism, 1846–1956* (1957); GEORGE STEINER, *Tolstoy or Dostoevsky: An Essay in the Old Criticism* (1959), stimulating but impressionistic; В.У. Кирпотин, *Ф.М. Достоевский, творческий путь* (1960); ERNEST J. SIMMONS, *Dostoevsky: The Making of a Novelist* (1950, reprinted 1962), a critical study of the development of Dostoevsky's creative art that makes extensive use of his notebooks; EDWARD WASIOLEK, *Dostoevsky: The Major Fiction* (1964); DONALD FANGER, *Dostoevsky and Romantic Realism* (1965); ROBERT L. JACKSON, *Dostoevsky's Quest for Form* (1966); R.L. BELKNAP, *The Structure of The Brothers Karamazov* (1967); KONSTANTIN MOCHULSKY, *Dostoevsky: His Life and Work* (1967; orig. pub. in Russian, 1947), one of the best and most comprehensive works on Dostoyevsky's life and writings; RICHARD PEACE, *Dostoevsky: An Examination of the Major Novels* (1971).

(E.J.Si.)

Draft Animals

Before the Romans developed the waterwheel, man's only supplements to or replacements for his own muscle power were animals he trained to work for him. Even now, after the harnessing of steam, oil, and electricity, over 80 percent of the world's cultivable land is still tilled by men with the help of animals. This article surveys the animal species used and the ways man has used them, for traction, as pack animals, and for other agricultural operations not involving traction. For the use of horses for riding see the articles entitled RIDING AND HORSEMANSHIP; and HORSE RACING. For information on the biology of the various species see HORSE; ARTIODACTYLA; etc. Care and feeding are covered in LIVESTOCK AND POULTRY FARMING; and ANIMAL FEED.

Draft animals were in use long before man began to keep written records, as indicated from the remains of animals, harness, and chariots in ceremonial graves and from tomb paintings, temple decorations, statuettes, coins, and seals (*e.g.*, signet rings). Though animals were first domesticated some 9,000 to 11,000 years ago, there is no evidence of their exploitation for work until many years later. Archaeological records suggest that reindeer sledges were used in northern Europe before 5000 BC, and land sledges drawn by oxen were used in ancient Mesopotamia before 3500 BC. Pack asses are represented in ancient Egyptian art of about the same period. With the invention of the plow and the wheel (before 3000 BC) the use of animals for draft became common in Mesopotamia and spread outward from there in all directions. Wheeled vehicles drawn by animals reached the Indus Valley by 2500 BC, Europe by 2000 BC, Egypt by 1600 BC, China by 1300 BC, and Britain about 500 BC.

BOVINES

Cattle were the first animals used for work and have always been the most important farm animal the world over. In the West, during the last few centuries, they were replaced by horses and these in turn by tractors.

Common cattle. Though the earliest archaeological records of domestic cattle date back to the 7th millennium BC, the first representations of plows do not appear until some 3,000 years later. It is not known whether this is simply a gap in the records or whether cattle were first domesticated for some other purpose, such as religious ritual, and only later adopted for farm work. Until 1000 BC cattle were the only animals used for heavy transport.

The original draft team, for sledge, cart, or plow, consisted of a pair of oxen. This pattern continued through Greek and Roman times. In many Mediterranean countries the situation is little changed even today. For the heavier soils of northern Europe, a larger team was needed; there the normal complement was four or eight.

Competition between horse and ox

Only after the Norman Conquest (1066) was the horse used for farm work in Britain. At that time began the long competition between horse and ox, which was finally settled in favour of the swifter horse, though in a few places the ox team survived into the 20th century, some as sentimental anachronisms.

Both draft oxen and draft horses were brought to the New World. In the early days of America, in fact, cattle were valued more for draft than for meat. Covered wagons on the western trails were largely drawn by oxen. The use of cattle for heavy work continued at least until the end of the 19th century. Competitive ox-pulls, using a sledge loaded with stones, were still a feature of local fairs in New England as late as the 1930s.

While the horse revolution came (and went) in western Europe, oxen remained the standard draft team throughout Asia and North Africa. South of the Sahara, particularly in tsetse fly areas where livestock do not thrive, native agriculture still relies primarily on human muscle power (hoe cultivation). Though the ox is used as a pack and riding animal by nomadic cattle owners in countries such as Chad, Cameroon, and Senegal, the idea of using oxen as draft animals was introduced comparatively recently (1920s) by Europeans in limited areas of East and West Africa.

The greatest users of ox power in Africa south of the Sahara were the Dutch colonists. Soon after they settled at the Cape of Good Hope in 1652, they obtained cattle from the Hottentots and quickly developed them into an excellent breed of draft oxen—the Africander, which played a crucial role in opening up the country.

Selection of work oxen is based chiefly on their conformation. For rapid work in light soils, the rangy long-legged type is chosen. For slow, heavy work a correspondingly heavy, thick-set animal is preferred. Some trials with scientific performance tests have been made. In India a dynamometer has been used to measure the draft power of oxen. At livestock markets in Taiwan, cattle show their strength by pulling a cart with locked wheels. Temperament is also of vital importance.

Though cows, bulls, and castrates are all used for work, the castrate (the ox) is much the most favoured. In central Europe many breeds are still used for milk and meat as well as draft. Experiments, both in central Europe and India, have shown that even cows in milk can work a normal day with little effect on milk yield.

Types of buffalo

Water buffalo. In southern Asia and in Egypt, the domesticated Asiatic buffalo (*Bubalus bubalis*) is an important work animal. There are two main types: the dark gray "swamp" buffalo (which still closely resembles the wild animal in appearance) and the black "river" buffalo. In the rice-growing countries of southeast Asia, the swamp buffalo is the principal draft animal. It is rarely milked, but is usually eaten at the end of its working life. In India, Pakistan, southwest Asia, and Egypt, where various dairy breeds of river buffalo have been developed, their use for meat and draft is secondary to milk production. The buffalo is also raised in some countries of southeast Europe, where it is now being transformed from a draft into a meat animal.

The water buffalo is very suitable for work in humid climates and on water-logged land. Although (unlike the zebu) not adapted to dry conditions, it is used for draft work on roads because it can pull heavy loads.

Yak. On the cold high plateaus of Tibet, and the neighbouring provinces of China (especially Tsinghai and Szechwan), as well as on the southern slopes of the Himalayas (northern Nepal, Bhutan, and the North East Frontier Agency of India), the yak (*Bos grunniens*) is the principal beast of burden. It is also used for riding, draft, and milk. Its hybrid with cattle, the dzo (also spelled zho and zo), is used similarly. In Tibet the dzo is often preferred to the yak for plowing.

EQUINES

Wild horses are native to northern Asia and Europe, onagers ("half-asses" or "half-horses") to central and southwest Asia, and wild asses to North Africa. Each was domesticated in its native area and was first used for load carrying or traction.

Donkey or ass. The ass is the oldest pack animal. It was used in Egypt possibly as early as 3500 BC, and certainly by 2500 BC. From the Nile Valley its use gradually spread westward around the Mediterranean, southward in Africa as far as Kenya, and eastward across Asia to China. Although the donkey has been primarily a load carrier, it has been ridden and used for draft (*e.g.*, plowing) as well. Donkeys were recently introduced into southern Africa for draft purposes.

Though essentially a beast of arid regions, the ass is not as specialized as the camel. The very modesty of its food requirements makes it an ideal poor man's animal.

Mule. The mule is the infertile hybrid between mare and jackass. It combines the size and strength of the horse with the resistance of the donkey. The surefooted mule is still popular, both for pack and draft, in rugged terrains and mountainous areas, particularly in the Mediterranean and South America.

Onager. In the 3rd millennium BC, onagers served as draft animals in ancient Mesopotamia. There are representations in Sumerian art of a yoke of four onagers

abreast pulling a heavy two-wheeled chariot or a four-wheeled cart. At the beginning of the 2nd millennium, the more amenable horse was introduced from the north and replaced the onager, whose domestication then ended.

Horse. The horse was domesticated in the Ukraine or Central Asia, probably toward the beginning of the 3rd millennium BC. It was probably ridden before being harnessed to a vehicle. Introduced to the Middle East, it was taken by invading Hyksos to Egypt, where it first appears in art about 1580 BC. Subsequently the horse and chariot spread over the whole of the ancient world from the Mediterranean to China. In contrast to the situation in their northern homeland, however, horses were not used for riding or agricultural work until much later.

Horse plowing

In England, the horse was first used for harrowing, which it could do faster than an ox. Horse plowing did not start until the end of the 12th century. At first horses joined the ox team; mixed yokes were common.

Disagreement over the relative merits of horse and ox as draft animals has lasted for 500 years and in some countries (*e.g.*, Japan) has not yet ended. The ox has a long strong pull for the old heavy plow, and does not require shoeing. As a ruminant it can be fed on coarse fodder (roughage). It can even be eaten at the end of its working life. With the wheeled plow, however, the horse is a much quicker worker, and in medieval northern Europe, where an oat crop could easily be grown, the horse largely supplanted his slow-gaited rival.

In the 17th century, the heavy horses that had been developed in northwest Europe for carrying feudal knights became available for agriculture, where they formed the basis of heavy draft breeds such as Belgian, Percheron, and Clydesdale.

The reign of the plow horse in western Europe and North America was ended in the 20th century by the large internal-combustion engine. Changeover was particularly rapid in the large farms of the U.S. By contrast, in the small peasant properties of central and eastern Europe, the horse is still an important farm animal.

In central and eastern Europe, the Netherlands, Sweden, and Finland special performance tests with a dynamometer measure pulling and carrying ability of farm horses. It is interesting to note that, in proportion to its body weight, a small horse can pull a heavier load than a large horse. In the Western world, the horse has returned to its original role as a sport animal, although the sports are now riding and racing rather than war and hunting.

Zebra. The zebra, the equine native to eastern and southern Africa, has never been systematically domesticated, but occasionally individuals have been tamed and teams trained to pull carts or carriages for amusement.

OTHER DRAFT ANIMALS

Dog. The use of dogs to pull sledges in the Arctic is well known—whether it be the Husky of North America or the Samoyed and Laika of Siberia. In America the original arrangement was a fan hitch of 12–15 dogs, each with its separate attachment to the sledge. More usual now is a pair hitch involving about eight dogs in double file which may pull loads of 675 kilograms (1,500 pounds) or more. In the U.S.S.R. smaller teams (4–6 dogs) and lighter loads (up to 158 kilograms or 350 pounds) are more usual.

Dogcarts

It is less well known that dogs were used until recently in western Europe for pulling small carts. In Britain this practice was forbidden by law in 1885 (since then "dogcart" has meant a light pony-drawn carriage with a box under the seat for carrying dogs). In Belgium, the Netherlands, and parts of France, Germany, and Switzerland, small carts drawn by one or two dogs were used by farmers to bring their produce (especially fruit, vegetables, and dairy products) to the town market until just after World War II. The dogs used were of Mastiff type and pulled loads of 135–180 kg (300–400 lb).

Reindeer. Reindeer drew sledges in northern Europe long before horses were harnessed for the task. In Lapland, a single reindeer harnessed to a sledge can pull a load of 75 kg (165 lb), or even 100–150 kg (220–330 lb) at a slower pace under good conditions. The Samoyeds harness two to four or even more animals per sledge; these pull heavier loads. The Lapps use their animals for packing in summer. In Siberia, reindeer are also used as pack animals and for riding; a good male carries 65 kg (150 lb) for 80 kilometres (50 miles) a day.

Camel. The camel is the desert animal par excellence. As a beast of burden it is irreplaceable. The one-humped camel or dromedary supplies all the needs of the pastoralists in the semidesert areas of North Africa and southwest Asia as far east as Rajasthan in India. In neighbouring agricultural areas, though chiefly used for carrying bulky loads (cotton, groundnuts, straw, green forage), it also turns the *sakia* (Persian wheel) to lift loads and drive mills. It is sometimes yoked to a plow, occasionally teamed with ass or ox.

In Central Asia, northern China, and Mongolia, the two-humped or Bactrian camel is used as a pack animal. In China an average load is 115–135 kg (250–300 lb) carried at 3 km per hour (2 mi per hour) for 40 km (25 mi) a day; the strongest animals can carry 270 kg (600 lb) and cover 1,050 km (650 mi) in 30 days. In India, an adult male dromedary can carry a load of 225–295 kg (500–650 lb) at 3–4.5 km per hour (2–3 mi per hour) and cover 32 km (20 mi) per day; the best animals can manage up to 405 kg (900 lb) over short distances.

Elephant. Domesticated Indian elephants are represented on seals from the Indus Valley civilization dating from 2500 BC; they have presumably been used for riding and as beasts of burden ever since. In India, Sri Lanka (formerly Ceylon), and Burma, their chief use now is in lumbering—they drag logs through the jungle and push floating logs around bends and off sandbanks ("aunging"). They also serve as baggage animals. Mature elephants can carry a load (excluding gear) of 180–360 kg (400–800 lb) depending on their size.

The African elephant was used in war both in Ptolemaic Egypt and in Carthage, playing a notable role in Hannibal's invasion of Italy in 218 BC. But excessive exploitation and increasing desiccation of the environment led to extinction of the North African variety. South of the Sahara, elephants have not been domesticated.

Llama. The llama is the only large animal domesticated in pre-Columbian America. Its habitat is restricted to the Andean highlands. Since the Incas did not have wheeled vehicles, they used the llama exclusively as a pack animal. In its present area of distribution in south Peru and Bolivia, its chief use is still as a pack animal. Normal load is only 25–30 kg (55–66 lb), exceptionally up to 35 kg, or 77 lb. It cannot travel more than 25–30 km (15–20 mi) per day.

Sheep and goat. Some large breeds of sheep and goat are used as beasts of burden at altitudes of 3,500–5,500 metres (12,000–18,000 feet) in Tibet and over the Himalayan passes between Tibet and India. In northern Kashmir these pack animals are of the Chanthan breed of sheep and Kel breed of goat. In the Punjab they are known as Kangra Valley animals and belong to the Biangi and Gaddi breeds of sheep and the Chigu breed of goat. They can carry a load (*e.g.*, of salt) of 4.5–18 kg (10–40 lb) for 15 km (10 mi) a day.

It is not known whether fact or fancy inspired the scene portrayed on a signet ring from Crete (15th century BC) that shows a pair of wild goats drawing a chariot, but it is true that domestic goats have been used as draft animals in western Europe.

USE OF DRAFT ANIMALS

Use of animal feet

In agriculture. *Use without implements.* For some agricultural operations, the herd animals' feet alone are sufficient. In ancient Egyptian tomb paintings sheep are shown being driven over the fields to tread in the seed corn after it has been broadcast. Later in the year the same procedure was used for threshing—animals were driven around the threshing floor treading out the grain. In the East this method of threshing is still employed using either cattle or buffalo.

In Southeast Asia, water buffalo may be used for culti-

vating the rice fields by puddling the flooded fields with their large hoofs. In Sri Lanka the same activity is used in puddling clay for bricks.

Draft power for agricultural implements. The first cultivating tool was a hoe pushed or pulled by a man or (more commonly) a woman. Later, two oxen were attached and the hoe became a plow. Since then the plow has been the principal agricultural implement; even as late as 17th-century England, plowing and carrying in the harvest were the only farm tasks for which animals were used. There was no spring cultivation and seeds were broadcast by hand. Parallel with the introduction of horses, the number of operations and implements increased. Horses were first used for harrowing and later for sowing (pulling the seed drill), rolling (to break up clods), hoeing, haymaking (pulling the mowing machine, the swathe turner, and the rake), harvesting (pulling the reaper or reaper-binder), in addition to pulling carts and plowing.

In less sophisticated agricultures, where cattle still dominate, the range of operations is smaller. Around the Mediterranean and in Asia, cattle or buffalo are used for plowing and for subsequent cultivation with the harrow. They are also used for threshing by pulling a sledge through the grain on the threshing floor. This method is still standard in Egypt and, until recently, was common (often with mules) in other Mediterranean countries.

Stationary machinery. Animals have been used to pull rotary devices at least since Roman times. In India and the Sudan, camels still turn the sugarcane crusher and oil-extraction mill. There and elsewhere animal power still lifts irrigation water out of wells, rivers, or canals. The blindfolded animal, whether horse or mule as in Spain, ox or buffalo as in Egypt, still endlessly turns the waterwheel, or *sakia*. In the West mechanical power has long since replaced the ass that turned a treadmill to raise water from the well, as well as the hard-worked dog that used to run in a wooden wheel to turn the spit, the butter churn, or the blacksmith's bellows.

In transport. *Pack animals.* Before the invention of the sledge and the wheel, loads were carried first by man and then on his animals. Humans still fulfill this function all over Africa and the Orient. Donkeys were the first pack animals and are probably still the most important, though mules can carry heavier loads. Horses are rarely employed. Camels and elephants are invaluable for bulky and heavy loads. Yaks, llamas, sheep, and goats carry light loads in high mountains. Cattle are used in West Africa.

Pulling wheelless vehicles. Animals were used for drawing sledges before invention of the wheel made the more efficient cart possible. The snow sledge or sleigh drawn by dogs, reindeer, or horses remains in use where wheels slip, get stuck, or break the ice. Over rock or soil the sledge is now only encountered in rare situations where the going is so rough or muddy that wheels are useless. In Thailand, the Philippines, and Borneo, for instance, buffalo draw a light wooden sledge through forest or over cultivated land; the ox-drawn sledge is seen in Botswana. Elephants, oxen, or buffalo may draw logs on a sledge or may pull the log itself as a skid. Even in England, the agricultural sledge was used until recently.

Towing barges

Barges on the canals of Europe were commonly towed by horses; the towpath beside the canal was designed for these animals. The load that one horse could pull on a canal was of the order of 50 metric tons (55 short tons) (30 metric tons [33 short tons] on a river), compared with only 126 kg (280 lb) on a pack horse. For wagons the load varies from one to eight metric tons (one to nine short tons) according to the nature of the road. Oxen still haul fishing boats ashore along the Portuguese coast.

Wheeled vehicles. In primitive agriculture, carts and wagons are the chief forms of transportation on the farm and from farm to market. With the development of macadamized roads in the 18th century, the horse-drawn carriage became an important means of transport for mails and passengers. During the 19th century, heavy horses pulled carts and drays on farms, coach horses drew stagecoaches and post chaises between cities, ponies were hitched to wagons in coal mines and traps on country roads, and light horses pulled innumerable cabs, carriages, carts (later, even fire engines and omnibuses) within towns (see WAGONS AND CARRIAGES).

In the West only an occasional horse-drawn milk delivery cart or brewer's dray is seen today; the last pit ponies will shortly vanish from English coal mines. Farther east in Europe, horses still pull the farm wagon. In Egypt there is a division of labour: the farm work is done by cattle or buffalo; on the roads the wagons are pulled by horses. In Asia the ox carts penetrate the cities, and the horses are restricted to passenger transport.

DRAFT HARNESS AND ACCESSORIES

Harness. The original ox team was harnessed by a simple wooden yoke to the pole of the cart or plow. In this antique but still used harness, the yoke rests on the back of the animal and the pull is on the projecting withers. Vertical bars on each side of the neck prevent sideslip and a rope round the neck holds the yoke in place and prevents a cart being tipped up. The hump of the zebu gives a larger projection for the yoke to rest on. With a single ox, the yoke can be attached to the shafts of a cart or to the plow by ropes. Alternatively the yoke can be attached to the head or horns; this is common in southern Europe. Another type of attachment occasionally used is the breastband.

With equines a yoke cannot be used because the withers are not sufficiently prominent. So with the introduction of onagers and horses for draft the breastband was employed. Unfortunately, as the horse pulls, the band presses on its windpipe and chokes it. This inefficient harness limited the use of horses for heavy draft work until the introduction of the rigid padded collar, invented in China in the 1st century BC and introduced in Europe about the 10th century AD. With a horse collar pressure falls on the front of the shoulders (*i.e.*, on the skeleton); the only limit to load or speed is the strength and willingness of the horse.

The horse collar

The modern draft harness is based on the collar, usually with the addition of a saddle to support the shafts of a cart and hence take some of the weight. Wooden hames (projections) fit into the collar and the shafts are attached to the harness by straps or chains (the traces). The crupper (loop passing under a horse's tail and buckled to the saddle) prevents the saddle riding forward and the breeching (part of the harness that passes over the animal's buttocks) prevents the vehicle forcing the saddle forward when backing or going downhill.

In the absence of shafts, the traces are joined directly to the vehicle or to swingletrees (pivoted swinging bars to which the traces of a harness are fastened) to give an even pull. Two horses are harnessed to a plow by means of two sets of swingletrees.

Though collars have been tried on oxen to prevent sores caused by galling of the yoke or the choking effect of the breastband, the old system is still the most popular.

Packsaddles. A proper packsaddle consists of a pad to protect the back, a wooden framework to support the pack (with straps around neck and tail), and a pack frame that is loaded and placed on the animal's back.

For camels the pack is very extensive and consists of two bags filled with grain or straw, one on each side of the animal. The framework, reduced to one or two wooden bars fastened to each bag, is held in place by ropes around neck, belly, and tail.

Control. Cattle and camels are guided by a single nose rope, horses by a bit and reins. The nose rope, or guiding rein, is threaded through the nasal septum; a wooden peg or plug prevents it from pulling through. In China (including Hong Kong and Taiwan) the buffalo's nose is ringed and the cord attached to the ring. Sometimes control is effected by a rope looped around the horns.

Horses are controlled by a metal bit through the mouth to which the reins are attached. The bridle holds the bit in position. This system dates from 1500 BC. Blinkers may be added so that the horse can only see ahead.

Shoes. On the paved Roman roads, horses needed foot protection to prevent wear and slipping. The iron shoe was used in Rome from the 1st or 2nd century BC and in northern Europe from the 8th century AD. Where oxen are used on roads they too are shod. In Italy and Portugal, for instance, oxen wear special double shoes to fit their divided hoofs. In the Far East, shoes tend to be more primitive; in Java *trompahs* for buffalo are made from old automobile tires.

CONCLUSION

While animals are actually working their efficiency is as high as a tractor's (about 25 percent), but over their lifetime their energy output in useful work represents only a small percentage of the potential energy of their fuel. It is especially low if they are used for only a short period each year. On the other hand, animals supply milk and manure as well as power, require only modest protection from the weather, and, compared with tractors, require little skilled maintenance. Animals can use local products as fuel and are self-reproducing. A tractor requires expensive (often imported) fuel and its capital cost is high. For these reasons, it is only when the standard of living is high and agriculture is on a large scale that farmers can afford to replace draft animals by tractors. Though they have virtually disappeared from the Western world, elsewhere draft animals' numbers are declining only slowly and will long retain the place they held in agriculture for the last 5,000 years.

BIBLIOGRAPHY. The history of domestic animals is fully documented in F.E. ZEUNER, *A History of Domesticated Animals* (1963); and in R. TROW-SMITH, *A History of British Livestock Husbandry to 1700* (1957). The history of their utilization for work is surveyed in C.J. SINGER *et al.* (eds.), *A History of Technology*, 5 vol. (1954–58). Information on the management of specific draft animals may be found in the following: W.C. MILLER and E.D.S. ROBERTSON, *Practical Animal Husbandry*, 7th ed. (1959); B.S. VESEY-FITZGERALD (ed.), *The Book of the Horse* (1946); G. WILLIAMSON and W.J.A. PAYNE, *An Introduction to Animal Husbandry in the Tropics*, 2nd ed. (1965); W.R. COCKRILL, "Observations on Control, Restraint and Slaughter of the Water Buffalo in the Far East," *Vet. Rec.*, 80:225–229 (1967); C.L.B. HUBBARD, *Working Dogs of the World* (1947); and A.J. FERRIER, *The Care and Management of Elephants in Burma* (1947).

(I.L.M.)

Drafting

Drafting is the act of delineating a drawing of an object, its structure, or its parts in the graphic language universally employed to communicate the engineering intent of a design, or problem. Engineering drawings consist of lines representing surfaces, edges, and contours of objects, supplemented by symbols, dimensional sizes, and notes. In a broad sense, drafting is the graphical development of an engineering problem or record.

The term engineering graphics has come into use in place of engineering drawing because a wide range of related graphic detail must often be included, such as wood trim, ornate brickwork patterns, corbelling instructions, and even load-design formulas for certain cantilevered beams.

HISTORY

Among the earliest records of engineering graphics is a stone engraving of a fortress plan made by the Babylonian engineer Gudea about 2000 BC. The Roman architect Vitruvius, in his treatise on architecture (*c.* 27 BC), referred to technical drawing, emphasizing that the architect should have a working knowledge of drawing so that he could better show the work to be constructed. He also discussed the procedures and practices to be followed in making drawings, thus providing the first textbook of engineering drawing.

Villard de Honnecourt's *Sketchbook* (*c.* 1225–50), well known to students of medieval art and technology, has been the subject of extensive research and commentary. Although small in comparison with the works of Vitruvius and Leonardo da Vinci, it contains several technical drawings of church structures, church furniture, and geometrical devices. Leonardo da Vinci (1452–1519) illustrated his work profusely with technical drawings. His treatise on painting is regarded as the first book printed on the theory of projection drawing. Gaspard Monge (1746–1818), however, is regarded as the first to make an organized record of the principles of engineering drawing that are used today. His *La Géométrie descriptive* (1801) was the first formal treatise on modern engineering drawing. In it he developed the theory of projecting views of an object onto three mutually perpendicular coordinate planes (such as are formed by the front, side, and top of a cube), and then revolving the horizontal (or top) and profile (side) planes into the same planes as the vertical (front) plane.

The first organized thesis

Most original ideas for mechanical devices are first expressed in a freehand drawing made simply with a pencil. Freehand drawings and sketches can clarify and record various technical ideas and verbal expositions. With them the designer can organize his thoughts, record his ideas, and quickly lay out various solutions to a problem.

The fundamental theory for all orthographic (mutually perpendicular) projection drawing is derived from descriptive geometry as organized by Monge.

In a broad sense, the subject of engineering drawing includes many different kinds of projection, descriptive geometry, and graphics; but its most common form currently is orthographic projection drawing. In orthographic projection, any object, regardless of its complexity, can be viewed from any direction.

CLASSIFICATION OF DRAWINGS

Mechanical drawing. Most engineering drawings are made with instruments, hence the name mechanical drawing. Instrument-made drawings provide the greatest accuracy, neatness, and legibility.

Various instruments are available to simplify the making of drawings, such as pencils with varying degrees of hardness; T-squares and triangles for drawing parallel, perpendicular, and diagonal straight lines; scales for measuring; dividers for transferring distances; compasses for constructing circles and arcs; a drawing board to provide a flat surface for the paper; drafting tape to hold the paper on the board; and erasers. Other instruments include templates (patterns) for drawing irregular curves and various symbols, protractors to measure angles, erasing shields, and pencil sharpeners.

In orthographic projection drawing, an object is viewed with a point of sight set at infinity, making the visual rays, or lines of sight, from the eye to the object parallel. Since the visual rays are parallel, distances on the object in a direction perpendicular to the line of sight will show true lengths; surfaces of the object in a plane perpendicular to the line of sight will show true size and shape; distances on the object in a direction parallel to the line of sight will show as zero; and any surface of the object parallel to the line of sight will show as a line. Distances and plane surfaces inclined to the line of sight will appear foreshortened.

Point of sight

To show the distances in three coordinate directions (height, width, and depth), two views of an object will be required. With lines of sight perpendicular to each other, each view will show two of the three coordinate distances. Adjacent views have lines of sight perpendicular to each other to show these distances in three mutually perpendicular directions, again citing the front, side, and top of a cube as an example.

Any two adjacent views with lines of sight perpendicular to each other have their common dimension in alignment with each other in the two views, as is shown in Figure 1. Thus, at any distance the top or bottom view is in vertical alignment with the front view. The side view, either right or left, is in horizontal alignment with the front view; and an auxiliary view is in alignment with the view from which it is projected.

The views are placed on the sheet in sequence of adjacent sides of the object. Main views are the so-called principal views: the top (or bottom), front (or rear),

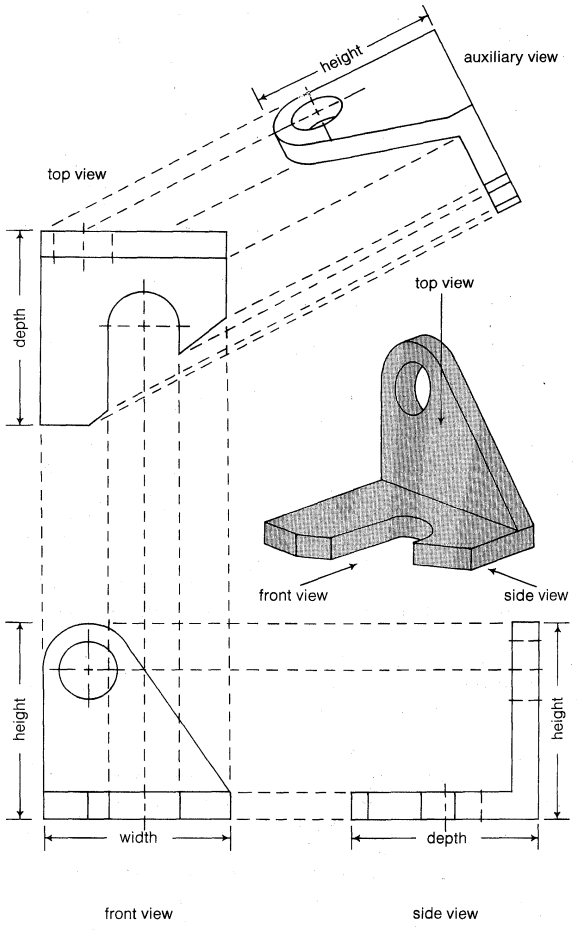


Figure 1: Views in alignment.

and right side (or left side). Other views, which can be taken from any direction, are known as auxiliary views.

Arrangement of views. The arrangement of views may be in either of two systems: third-angle projection or first-angle projection. In both systems, the principal views, the front, top, and side, are obtained by looking directly at the front, top, and side, respectively.

In third-angle projection, the top view is drawn in a location vertically above the front view, and the side view is placed horizontally in direct line with the front view. In both the top and the side views, the front of the object is positioned toward the front view; thus the right side view is that seen to the right of the front view, and the left side view is that seen to the left of the front view. Third-angle projection is an almost universally accepted method in the United States.

In first-angle projection the top view is drawn in a location vertically below the front view; the right side view is placed horizontally in line with the front view but located on the left of the front view; and the left side view is located at the right of the front view. In depicting both the top and side views, the front of the object is facing away from the front view. First-angle projection is the commonly accepted method in many European countries.

An auxiliary view is taken from any direction except that of one of the three principal views. A first-auxiliary view is one with a line of sight perpendicular to the line of sight of one of the principal views. A second-auxiliary view is a view with a line of sight that, in turn, is perpendicular to the line of sight of a first auxiliary view.

Many objects have curved or angular surfaces that are not perpendicular to the line of sight of any of the principal views; therefore, they do not appear in their true size and shape in the principal views. In order to show the true size and shape of such surfaces, it is necessary to obtain a view with a direction of sight perpendicular to the surface.

Auxiliary views are constructed by projecting from a given view to obtain measurements at right angles to the projecting lines that are common to the two adjacent views. In the auxiliary view, measurements parallel to the projecting lines are the same as the corresponding measurements in the other views adjacent to the one from which the projection is made.

Working drawings. A working drawing describes an object so completely that it can be produced without further information. It provides specifications as to shape, size, material, and finish. It may suggest the shop processes required for production of the object, but it does not specify that only certain processes may be used.

In general, the main (or front) view should show the object in its operating or usual position, or in the position it occupies in the assembly. In many cases, however, its position is not important, and the view that shows the characteristic shape of the object is selected for the front view. One-view drawings are used whenever views in more than one direction are unnecessary, *e.g.*, for flat parts made of thin material. Many parts, such as cylinders, require only two views. More views would merely duplicate the views already shown. More than two views are necessary for objects that are made up of combined geometric shapes.

One-view drawings

Lines that make up a drawing have been thoroughly standardized, and their use on drawings is universally recognized. Visible object lines are continuous thick lines. Hidden lines are broken, or dashed, medium thick lines. Dimension lines, extension lines, leaders, and section crosshatch lines are thin continuous lines. Centre lines are alternate long–short dashed thin lines. Phantom lines are long dashed thin lines. Cutting plane lines are medium dashed heavy lines.

To save drafting time, many standard features, such as threads that frequently occur on machine parts, are represented by conventional symbols. In the conventional representation, the helical curves of the thread for both the crest and the root may be represented by straight lines drawn at right angles to the length of the thread; or, for further simplification, threads may be represented by hidden lines parallel to the axis at the root diameter of the thread, as shown in Figure 2.

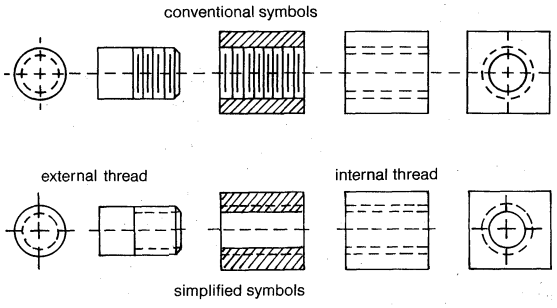


Figure 2: Thread representation.

Standards have been developed for a considerable number of graphic symbols to be used on drawings and diagrams, for gear teeth, pipe fittings, hydraulic fixtures, electrical devices, and other units.

Sections. By means of properly selected views, the external features of the most complicated objects may be fully described. The showing of complicated interiors of many parts, however, results in a confusion of hidden lines that may be extremely difficult to interpret. Such objects can be adequately described by the use of sectional views. The object is imagined as cut through the interior by a cutting plane, and the resulting object is shown as a view. The portion of the object cut by the cutting plane is crosshatched by section lining.

Various conventional methods of passing cutting planes result in certain sectional views, such as half section, full section, broken-out section, revolved section, removed section, auxiliary section, offset section, aligned section, and partial section, as shown in Figure 3.

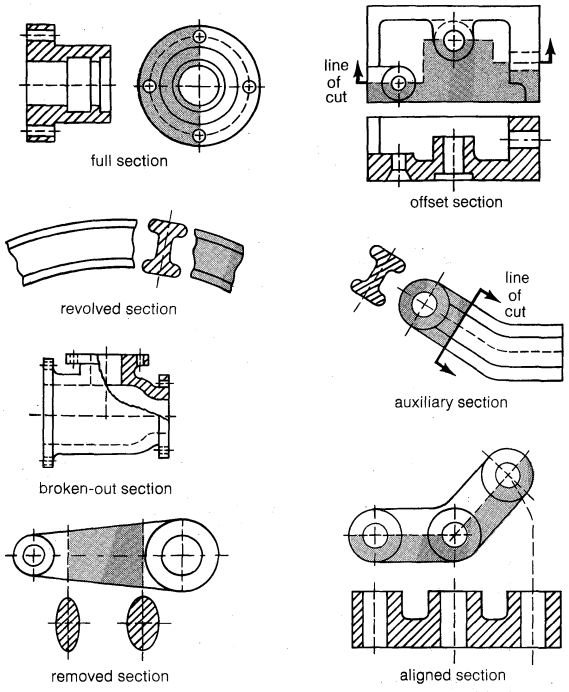


Figure 3: Conventional sections.

Drawings should be completely dimensioned so that it will not be necessary for the reader to scale the drawing or assume any dimensions. Drawings should be made accurately to scale, however, to maintain correct proportions between the various features of the object. The scale is usually selected to be some convenient multiple or fraction of full size, such as full size, half size, quarter size, one-tenth size, double size, or ten times full size, although any intermediate scale may be used.

Dimensions. Dimensions, the size specifications added to the shape description as shown by the orthographic drawings, consist of the numerical values of the measurements, directed to the proper location on the part by means of dimension lines to indicate the extent of the measurement and by extension lines to indicate the relevant surfaces or locations on the object. The units of measurement must also be specified (see Figure 4).

Use of decimal dimensions

In the past, common fractions were considered adequate for dimensions; but as products became more complicated and precise sizes became necessary in order to have interchangeable parts, more accurate specifications were required. It then became necessary to turn to the decimal system. Today, decimal dimensions are widely and increasingly being used in industry.

There are many types of manufacturing in which units and common fractions are still used, however, because extreme accuracy is not required or is relatively unimportant. Architectural and structural work is an example. Such work is usually one of a kind, and much fitting is done in the final assembly. These drawings are commonly dimensioned in units and common fractions.

Tolerances. It is impossible to produce products to an exact size. In actual practice, exact sizes are not essential. Only varying degrees of accuracy are necessary according to the functional requirements of the product. Tolerances are therefore the means of specifying dimensions to the degree of accuracy required.

A tolerance is the specification of the permissible variation in the size of a part or in the location of a feature with reference to other features of a part. Tolerances are expressed in the same units and to the same degree of accuracy—that is, the same number of decimal places—as the dimensions to which they apply. Tolerances may be expressed either as plus-and-minus values to be added to the basic dimension or as maximum-and-minimum values of the dimension (Figure 4).

It usually is necessary to supplement the direct dimensions with notes, which should be brief and carefully worded to permit no misinterpretation. Local notes are directed to the proper place on the drawing by means of leaders. General notes, applying to the entire drawing rather than to a specific portion of it, are placed either in one corner of the drawing or centred beneath the views to which they apply.

Lettering. Dimensions and notes are lettered on the shape description shown on the drawing in order to complete the size description. Lettering should be suitable for easy and rapid execution. It must also be legible after reproduction, particularly for photographic reduction and subsequent enlargement where small spaces between the letter elements may tend to fill in.

Standard lettering

The form of lettering generally recognized as standard for use on drawings is single-stroke Gothic. Single stroke refers to the width of a single stroke of the pen or pencil used in making the line elements of the letters.

All lettering that is a part of the drawing should be in capital letters. Lowercase letters may be used for notes and memorandums.

Either vertical or inclined letters may be used. Vertical letters are slightly more legible than inclined letters but more difficult to execute.

A detail drawing is the complete representation of a single part, including all necessary views, dimensions, and notes. Frequently, each individual part is shown on a separate sheet.

An assembly drawing shows the assembled machine or structure with all detail parts in their functional positions. Assembly drawings vary in character according to use: design assemblies or layouts, general assemblies, working drawing assemblies, outline or installation assemblies, and check assemblies.

Diagrams. Diagrams are simplified drawings of electrical or hydraulic circuits describing the connections and functions of such circuits. Standards also have been developed for a number of graphic symbols used on these diagrams. A circuit diagram shows, by means of single lines and graphic symbols, the course of an electric or hydraulic circuit or system of circuits and the connections of an installation or its component devices or parts. A connection diagram usually shows the general physical arrangement of the component devices, but the internal connections of the unit assemblies or equipment may be omitted. A schematic diagram shows, by means of lines and graphic symbols, the connections and functions of a specific circuit arrangement. It facilitates tracing through the circuit and its functions.

Pictorial drawing. Pictorial representation designates the method of projection that shows the object approxi-

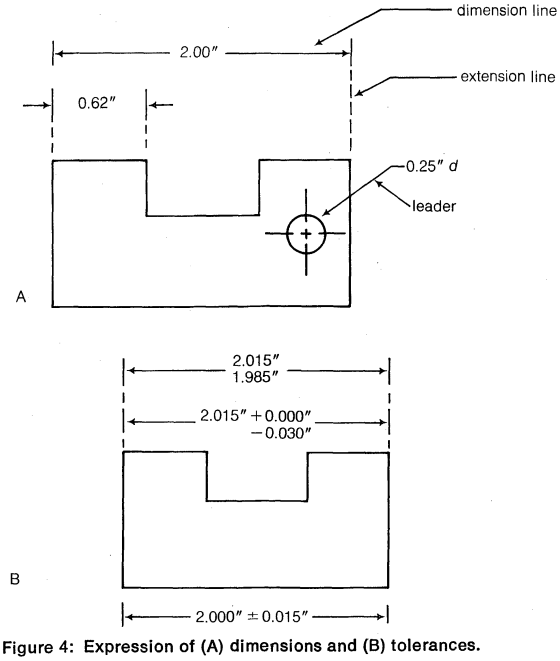


Figure 4: Expression of (A) dimensions and (B) tolerances.

mately as it would be seen by the eye. There are three main divisions: axonometric, oblique, and perspective.

Axonometric. Axonometric is a type of pictorial orthographic projection drawing where the line of sight is inclined to the principal faces of the object and to the principal axes of the object. Thus, all dimensions in the direction of the principal axes are foreshortened. The faces of the object not perpendicular to the line of sight are distorted.

Often, axonometric projections are most easily constructed by taking a second-auxiliary view of the object (Figure 1). If the second-auxiliary view is taken in such a direction that all three principal axes are foreshortened by different amounts, the resulting view is a trimetric projection. If the second-auxiliary view is taken in such a direction that two of the three principal axes are foreshortened the same amount and the third by a different amount, the resulting view is a dimetric projection. If the second-auxiliary view is taken in such a direction that all three principal axes are foreshortened the same amount, the resulting view is an isometric projection.

With the geometric method of positioning ordinary orthographic views of an object on a sheet, axonometric projections can be produced directly by projection from the views, without a first-auxiliary view.

By setting up scales to be used in the directions of the three principal axes, axonometric drawings can be produced directly without projection from the orthographic views. Since distances are proportional only in the direction of the principal axes, all measurements must be parallel to these axes. Such drawings are not true to scale, but this is seldom important because the main purpose of axonometric drawings is to show relationships. The most common drawing is isometric; the three axes are equally foreshortened and may be laid off to the same scale.

Oblique. Oblique projection is another form of pictorial representation, much like axonometric, employing a point of sight at infinity with parallel visual rays. Unlike axonometric, however, oblique projection requires a plane of projection, or picture plane, on which the view is projected.

Picture plane

The picture plane is set at an angle other than 90 degrees with the visual rays. It usually is parallel to one of the principal faces of the object, the face showing true size and shape in the oblique projection. The view is projected to the picture plane along the visual rays (Figure 5).

Oblique drawings also may be produced directly by repeating the width and height measurements as shown on an ordinary orthographic drawing and converting depth dimensions to any scale desired along a receding axis drawn at any angle except horizontal or vertical.

Perspective. Perspective drawings are the most realistic of any mechanically drawn pictorial representations, since they show the apparent convergence of parallel lines as they recede. Perspective drawing, like oblique, utilizes a picture plane on which the view is projected. The point of sight (or vanishing point) is at a finite location, however, and thus the visual rays are not parallel but converge at the point of sight.

A perspective drawing can be projected by the ordinary methods of multiview orthographic projection from the top and side views by constructing the front view as a view only of the picture plane (Figure 6). A more realistic view results if the object is turned so that two of the principal faces are inclined to the picture plane.

Architectural and structural drawings. Architectural and structural drawings embrace some standard practices and conventions unique to this field of drafting and unlike those in machine drawing. The fundamental principles, however, are the same as in ordinary machine drawing. In general, architectural and structural drawings are third-angle projection, but there are occasions when first-angle projection is used. Because of the size of the structure and the practical drawing sheet size, the architect or engineer cannot relate his views on one sheet as in a machine drawing in third-angle projection.

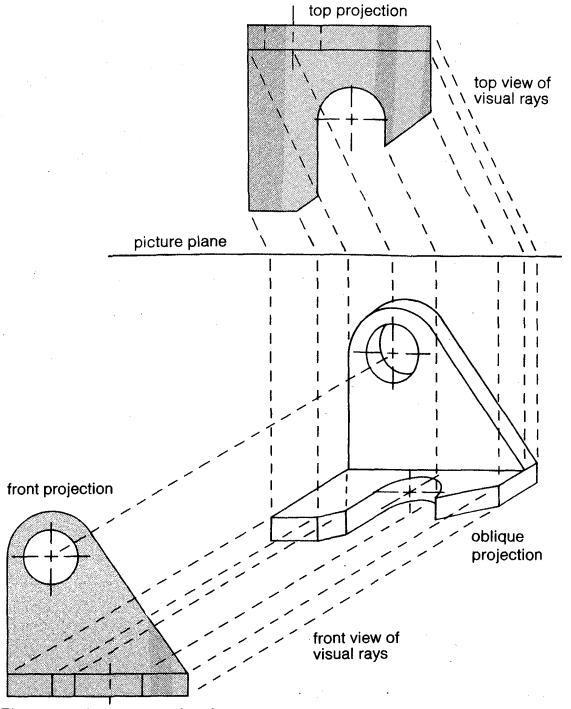


Figure 5: Oblique projection.

Instead, he uses one sheet at a time, constructing the plan view or building elevation he wishes to show. Rather than projecting directly from view to view as in machine drawing, the draftsman must resort to measurements made according to the rules of projection.

In machine drawing it is customary to show individual parts separately, but in architectural and structural drawing the common practice is to show the individual parts assembled. In other words, all the parts of a built-up structural member are detailed as far as possible in the places they occupy in the structure.

In addition to the usual two or three views, it frequently is desirable to show a bottom view of structural members. In architectural and structural drafting, such a view is made as a horizontal section looking down, instead of the regular bottom view looking up as is common in ma-

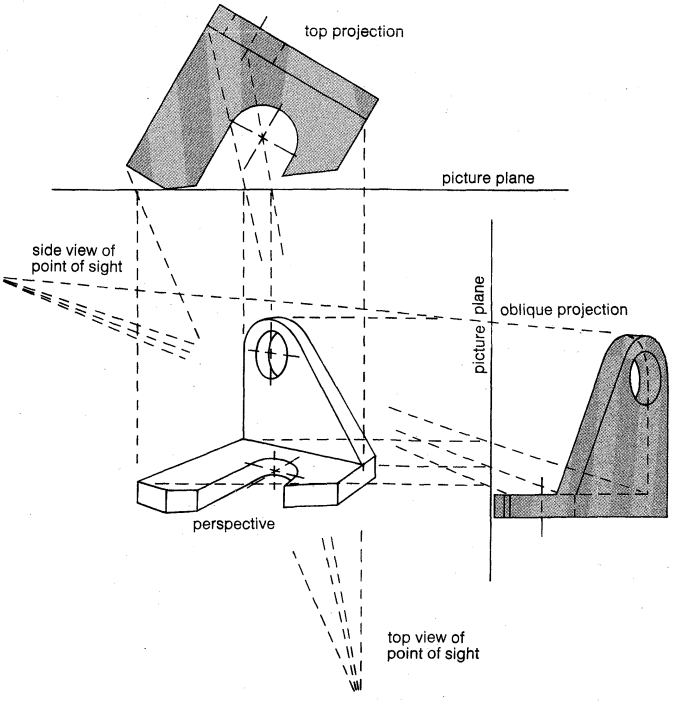


Figure 6: Perspective by projection.

chine drawing. The horizontal cutting plane is passed to show as little other detail as possible.

Pictorial graphs

Charts and graphs. A chart or graph is a simple pictorial representation of a group of related facts, statistics, or records, drawn so that the relationship may be understood at a glance. A pictorial or graphic description is easier to understand than a numerical tabulation or a verbal description.

Line graph. In a line graph, values of two related variables are plotted on coordinate-ruled paper and the points joined together consecutively to form a continuous jagged line or smooth curve. Line graphs can be drawn quickly and accurately; they permit a large number of plotted values in a compact space; they permit the comparison of the relative movements, or trends, of several curves on the same graph; intermediate values can be interpolated; and external values beyond the range of the plotted points can be extrapolated.

If one of the variables of a line graph is a unit of time, the graph is known as a time-series chart.

Polar charts. In a polar chart, two variables, one of linear magnitude and the other of angular magnitude, are plotted on a polar coordinate grid with respect to a pole, or origin, to form a continuous line, or curve. The angular variable is frequently a unit of time.

Bar charts. A bar chart is a graphic representation of numerical values by lengths of bars beginning at a base line, indicating the relationship between two or more related factors.

Circle charts. A circle chart, or sector chart, sometimes called a pie chart, is a circle divided into wedge-shaped portions to show the relative percentage of the component parts in relation to their total. Circle charts are a ready method for presenting data on a percentage basis and for emphasizing amounts rather than trends.

Flow charts. Flow charts are predominantly schematic representations of the flow of a manufacturing or chemical process.

Nomographs. A nomograph is a combination of functional scales arranged to represent a mathematical law or equation for computational purposes.

Solving three-variable equations

Basically, a nomograph is designed to solve a three-variable equation. A straight line joining known or given values on the scales of two of the variables intersects the scale of the third variable at a value that satisfies the equation represented. The scales may be either straight, curved, or circular. Two or more graphs frequently can be combined to solve an equation containing more than three variables.

Usually, nomographs have fixed scales and a movable alignment line. Movable-scale nomographs, however, can be constructed with a fixed-direction alignment line.

COMPUTER GRAPHICS

By reason of its speed and versatility, the use of the computer in the engineering design process and the direct preparation of engineering data in the form of engineering drawings or printout data is rapidly expanding. Input to the computer is by means of punched cards or tapes, magnetic tapes, character recognition devices, direct keyboard devices, or, for direct communication between the designer and the computer, cathode-ray tube devices.

In the cathode-ray tube devices, the designer makes his sketches with a special light pen on the face of a cathode-ray tube to provide the input data for the memory units of the computer. The designer can recall his sketches, revise them, change the scale, delete portions, or add material as he wishes, developing his design directly on the tube. The final drawing displayed on the tube can be converted photographically to a permanent record.

Using card or tape-controlled computers, the input data is in numerical rather than graphical form. Data for the lines, shapes, curves, and symbols are expressed mathematically, programmed in a computer language, and fed into the computer. The computer processes the data and converts it to a punched or magnetic tape, which in turn controls an electronic plotting machine. The plotting machine operates a pen or stylus according to the control tape, and supplies to the designer or draftsman a finished drawing, complete with all required lines, dimensions, and notes. The drawing may be the final product of the computer or it may be used to check the tape and the tape then used to guide the operations of numerically controlled programmed machine tools for the manufacture of the parts.

Engineers and draftsmen must have a thorough training in the computer process to provide the correct information to the programmer who prepares the input data for the computer. The computer will do only what it is programmed to do. Regardless of how automated the process becomes, the draftsman still has the responsibility for the communication of information that will guarantee the correctness of the finished product.

REPRODUCTION METHODS

In modern industrial practice, the engineering drawing represents a considerable economic investment, and control and protection of the original drawing are important. The original drawing is seldom handled directly by the workman, reproductions being used instead. Several methods of reproducing the original drawing are available, the most important currently being the blueprint and diazo processes.

Blueprint process. The blueprint process is the oldest, usually least expensive, and still a very commonly used method of reproduction. It is rapidly being replaced by the more convenient diazo process, however. The blueprint process was discovered in 1842 by Sir John Herschel, an English astronomer. The paper is sensitized with a coating of a potassium ferricyanide and ferric ammonium citrate solution, exposed to light under an original on translucent tracing paper, developed in water, and dried. The print consists of white lines on a dark-blue background. It is difficult to make highly readable changes, corrections, or additions to blueprints with an ordinary pen or pencil because of the background. Notes can be written on the print with an alkaline solution strong enough to bleach out the blue background.

Diazo process. The diazo process is similar to the blueprint process. The paper is coated with a light-sensitive diazo compound, exposed to light under a tracing, and developed by exposure to ammonia fumes. It provides prints with black, blue, or red lines on a white background; and changes, corrections, or additions can be added with an ordinary pen or pencil. Since the print is not immersed in a solution, the original drawing is reproduced the same size and undistorted.

Microfilm. Because a large number of drawings are required for a complex machine or structure, ordinary recording methods put a strain on filing, storage, and retrieval systems. The use of microfilm eases this burden.

In microfilming, the original drawings are photographically reproduced on film from $\frac{1}{8}$ to as little as $\frac{1}{40}$ of the original size. Film reproductions may be filed and stored in a roll, but more often frames are cut and mounted individually on aperture cards. Since the cards are essentially standard data cards used for electronic data processing, they may be handled on mechanical equipment capable of sorting, filing, and retrieving them. Cards may be viewed with a projection machine or on a microfilm reader or they may be reproduced in the original or any intermediate size on a printer.

BIBLIOGRAPHY. T.E. FRENCH and C.J. VIERCK, *Manual of Engineering Drawing*, 10th ed. (1966), a basic manual for students and draftsmen, covering all aspects of graphic communication; F.E. GIESECKE *et al.*, *Technical Drawing*, 5th ed. (1967), a comprehensive text and reference book on engineering drawing; W.J. LUZADDER, *Basic Graphics For Design, Analysis, Communications, and the Computer*, 2nd ed. (1968), basic fundamentals of graphics with applications to computer-aided design and drafting and numerically controlled machine tools; AMERICAN NATIONAL STANDARDS INSTITUTE, *Y14 Standards for Engineering Drawings and Related Documentation*, *Y10 Letter Symbols for Use on Drawings*, and *Y32 Graphical Symbols for Use on Drawings*, American national standard drafting practices for industrial drawings.

(F.L.S.)

Drake, Sir Francis

Sir Francis Drake, a 16th-century English admiral, was the greatest, as well as the most renowned, of the English seamen of the Elizabethan age.

Born *c.* 1540/43 in Devonshire, in southern England, on the Crowndale estate of Lord Francis Russell, second earl of Bedford, Francis Drake was the son of one of the latter's tenant farmers. His father was an ardent Protestant lay preacher, an influence that was to have an immense effect on Drake's character. His detestation of Catholicism had its origins not only in his father's teaching but in his own early experiences, when his family had to flee the West Country during the Catholic uprising of 1549. They made their way to Kent in southeast England and, in exchange for their former country cottage home, found lodging in one of the old naval hulks that were moored near Chatham on the south bank of the Thames Estuary. Had he stayed in Devon he might have become a yeoman farmer, but his family's poverty drove him to sea while still a boy. When Drake was about 13 years old, he was apprenticed to a small coastal vessel plying between North Sea ports. Thus, sailing one of the harshest stretches of water in the world, he learned early how to handle small vessels under arduous conditions. The knowledge of pilotage he acquired during these years was to serve him in good stead throughout his life. The old sea captain left Drake his ship when he died, so that Drake, thereafter, became his own master.

Early voyages. Drake might have spent all his life in the coastal trade but for the happy accident that he was related to the powerful Hawkins family of Plymouth, Devon, who were then embarking on trade with the New World—the New World that, as Drake never forgot, had been given by Pope Alexander VI to the kingdom of Spain. When he was about 23, dissatisfied with the limited horizons of the North Sea, he sold his boat and enlisted in the fleet belonging to the Hawkins family. Now he first saw the ocean swell of the Atlantic and the lands where he was to make his fame and fortune.

Voyages to the West Indies

On a voyage to the West Indies, as second-in-command, Drake had his first experience of the Spaniards and of the way in which foreigners were treated in their realms; their cargoes, for example, were liable to be impounded. At a later date he referred to some "wrongs" that he and his companions had suffered—wrongs that he was determined to right in the years to come. His second voyage to the West Indies, this time in company with John Hawkins, ended disastrously at San Juan de Ulúa off the coast of Mexico, when the English seamen were treacherously attacked by the Spanish and many of them killed. Drake returned to England in command of a small vessel, the "Judith," with an even greater determination to have his revenge upon Spain and the Spanish king (Philip II). Although the expedition was a financial failure, it served to make Drake's reputation, for he had proved himself an outstanding seaman. People of importance, including Queen Elizabeth I, who had herself invested in the venture, now heard his name. In the years that followed he made two expeditions in small boats to the West Indies, in order "to gain such intelligence as might further him to get some amend for his loss. . . ." In 1572—having obtained from the Queen a privateering commission, which amounted to a license to plunder in the King of Spain's lands—Drake set sail for America in command of two small ships, the "Pasha," of 70 tons, and the "Swan," of 25 tons. He was nothing if not ambitious, for his aim was to capture the important town of Nombre de Dios, Panama. Although himself wounded in the attack, he and his men managed to get away with a great deal of plunder—the foundation of his fortune. Not content with this, he went on to cross the Isthmus of Panama. Standing on a high ridge of land, he first saw the Pacific, that ocean hitherto barred to all but Spanish ships. It was then, as he put it, that he "besought Almighty God of His goodness to give him life and leave to sail once in an English ship in that sea." His name as well as fortune were established by this expedition, and he returned to England both rich and famous. Unfortunately, his return coincided with a moment when Queen Elizabeth and King Philip II of Spain had reached a temporary truce. Although delighted with Drake's success in the empire of her great enemy, Elizabeth could not officially acknowledge it. Drake, who was as politically discerning as he was navigationally brilliant, saw that the time was inauspicious and sailed with a small squadron to Ireland, where he served under the Earl of Essex, who was then engaged in suppressing a rebellion in that strife-torn land. This is an obscure period of Drake's life, and he does not emerge into the clear light of history until two years later.

By courtesy of the National Portrait Gallery, London



Drake, oil painting by an unknown artist. In the National Portrait Gallery, London.

Circumnavigation of the world. In 1577 he was chosen as the leader of an expedition intended to pass around South America through the Strait of Magellan and to explore the coast that lay beyond. The object was to conclude trading treaties with the people who lived south of the Spanish sphere of influence and, if possible, to explore an unknown continent that was rumoured to lie far in the South Pacific. The expedition was backed by the Queen herself. Nothing could have suited Drake better. He had official approval to benefit himself and the Queen, as well as to cause the maximum damage to the Spaniards. It was now that he met the Queen for the first time and heard from her own lips that she "would gladly be revenged on the King of Spain for divers injuries that I have received." He set sail in December with five small ships, manned by less than 200 men, and reached the Brazilian coast in the spring of 1578. His flagship, the "Pelican," which Drake later renamed "The Golden Hind," was only about 100 tons. It seemed little enough with which to undertake a venture into the domain of the most powerful monarch and empire in the world.

The voyage of "The Golden Hind"

Upon arrival in South America, it was discovered that there was a plot against Drake, and its leader, Thomas Doughty, was tried and executed. Drake was always a stern disciplinarian, and he clearly did not intend to continue the venture without making sure that all his small company were loyal to him. Two of his smaller vessels, having served their purpose as store ships, were then abandoned, after their provisions had been taken aboard the others, and, on August 21st, 1578, he entered the Strait. It took 16 days to sail through, after which Drake had his second view of the Pacific Ocean—this time from the deck of an English ship. Then, as he wrote, "God by a contrary wind and intolerable tempest seemed to set himself against us." During the gale, Drake's vessel and that of his second-in-command had been separated; the latter, having missed a rendezvous with Drake, ultimately returned to England, presuming that the "Hind" had sunk. It was, therefore, only Drake's flagship that made its way into the Pacific and up the coast of South America. He passed along the coast like a whirlwind, for the Spaniards were quite unguarded, having never known a hostile ship in their waters. He seized provisions at Valparaíso, attacked passing Spanish merchantmen, and captured two very rich prizes. "The Golden Hind" was below her watermark, loaded with bars of gold and silver, minted Spanish coinage, precious stones, and pearls, when he left South American waters to continue his voyage around the world. Before sailing westward, however,

he sailed to the north as far as 48° N, on a parallel with Vancouver, to seek the Northwest Passage back into the Atlantic. The bitterly cold weather defeated him, and he coasted southward to anchor just north of modern San Francisco. He named the surrounding country New Albion and took possession of it in the name of Queen Elizabeth. In his search for a passage around the north of America he was the first European to sight the west coast of what is now Canada.

In July 1579 he sailed west across the Pacific and after 68 days sighted a line of islands (probably the remote Palau group). From there he went on to the Philippines, where he watered ship before sailing to the Moluccas. There he was well received by the local sultan, and appears to have concluded a treaty with him giving the English the right to trade for spices. Drake's deep-sea navigation and pilotage were always excellent, but in those totally uncharted waters his ship struck a reef. He was able to get her off without any great damage and, after calling at Java, set his course across the Indian Ocean for the Cape of Good Hope. Two years after she had nosed her way into the Strait of Magellan, "The Golden Hind" came back into the Atlantic with only 56 of the original crew of 100 left aboard.

On September 26, 1580, Francis Drake brought his ship into Plymouth Harbour. She was laden with treasure and spices, and Drake's fortune was permanently made. He thus became the first captain ever to sail his own ship around the world—the Portuguese navigator Ferdinand Magellan having been killed before completing his circumnavigation—and the first Englishman to sail the Pacific, the Indian Ocean, and the South Atlantic. Despite Spanish protests about his piratical conduct while in their imperial waters, Queen Elizabeth herself came aboard "The Golden Hind," which was lying at Deptford in the Thames Estuary, and knighted the farmer's son.

Mayor of Plymouth

In the same year, 1581, Drake was made mayor of Plymouth, an office he fulfilled with the same thoroughness that he had shown in all other matters. He organized a water supply for Plymouth that served the city for 300 years. In 1585 he married again, his first wife, a Cornish girl named Mary Newman, whom he had married in 1569, having died in 1583. His second wife, Elizabeth Sydenham, was an heiress and the daughter of a local Devonshire magnate, Sir George Sydenham. In keeping with his new station, Drake bought himself a fine country house—Buckland Abbey (now a national museum)—a few miles from Plymouth. Drake's only grief was that neither of his wives bore him any children.

During these years of fame when Drake was a popular hero, he could always obtain volunteers for any of his expeditions. But he was very differently regarded by many of his great contemporaries. Such well-born men as the naval commander Sir Richard Grenville and the navigator and explorer Sir Martin Frobisher disliked him intensely. He was the parvenu, the rich but common upstart, with West Country manners and accent and with none of the courtier's graces. Drake had even bought Buckland Abbey from the Grenvilles by a ruse, using an intermediary, for he knew that the Grenvilles would never have sold it to him directly. It is doubtful, in any case, whether he cared about their opinions, so long as he retained the goodwill of the Queen. This was soon enough demonstrated, for in 1585 Elizabeth placed him in command of a fleet of 25 ships. Hostilities with Spain had broken out once more, and he was ordered to cause as much damage as possible to the Spaniard's overseas empire. Drake fulfilled his commission, capturing Santiago in the Cape Verde Islands and taking and plundering the cities of Cartagena in Colombia, St. Augustine in Florida, and San Domingo (Santo Domingo, Dominican Republic). The effect of his triumph in the West Indies was cataclysmic. Spanish credit, both moral and material, almost foundered under the losses. The Bank of Spain broke, the Bank of Venice (to which Philip II was principal debtor) nearly foundered, and the great German bank of Augsburg refused to extend the Spanish monarch any further credit. Even Lord Burghley, Elizabeth's principal minister, who had never approved of Drake or his methods, was forced to concede that "Sir Francis Drake is a fearful man to the King of Spain."

Defeat of the Spanish Armada. By 1586 it was known that Philip II was preparing a fleet for what was called "The Enterprise of England," and that he had the blessing of Pope Sixtus V to conquer the heretic island and return it to the fold of Rome. Drake was given carte blanche by the Queen to "impeach the provisions of Spain." In the following year, with a fleet of some 30 ships, he showed that her trust in him had not been misplaced. He stormed into the Spanish harbour of Cádiz and in 36 hours destroyed thousands of tons of shipping and supplies, all of which had been destined for the Armada. This action, which he laughingly referred to "as singeing the king of Spain's beard," was sufficient to delay the invasion fleet for a further year. But the resources of Spain were such that by July 1588 the Armada was in the English Channel. Lord Howard had been chosen as English admiral to oppose, with Drake as his vice admiral. It was, however, the latter's dash and fire that largely turned the scales, Drake himself managing to capture a rich prize during the long sea fight in the Channel. It was also Drake who prompted the use of fire ships to drive the Armada out of Calais, where it had taken refuge. Then, to delight his Protestant heart, "The Winds of God blew," so that the Spanish fleet was dispersed and largely wrecked. Drake was England's hero, achieving a popularity never to be equalled by any man until Horatio Nelson emerged more than 200 years later. Innumerable souvenirs were struck in his name, and he was immortalized in poems and broadsheets.

Last years

His later years were not happy, however. An expedition that he led to Portugal proved abortive, and his last voyage, in 1596, against the Spanish possessions in the West Indies was a failure, largely because the fleet was decimated by fever. Drake himself succumbed and was buried at sea off the town of Puerto Bello (modern Portobelo, Panama) on January 28th, 1596. Few men have been more famous in their lifetimes. As the Elizabethan historian John Stow wrote:

> He was more skilful in all points of navigation than any . . . He was also of perfect memory, great observation, eloquent by nature . . . In brief he was as famous in Europe and America, as Timur Lenk (Tamerlane) in Asia and Africa.

At a later date, Dr. Samuel Johnson wrote eulogistically of his character and bravery. But to the Spaniards he was, as their ambassador to England remarked, "the master-thief of the unknown world." He was "low of stature, of strong limb, round-headed, brown hair, full-bearded, his eyes round, large and clear, well-favoured face and of a cheerful countenance." A devout churchman and an able businessman, Sir Francis Drake was one of the world's greatest seamen. He embodied many of the virtues of expansionist Elizabethan England.

BIBLIOGRAPHY. There is no definitive biography of Drake. Of the comparatively short biographies available, the most useful is probably that by C.C. LLOYD (1957). J.A. WILLIAMSON, *Sir Francis Drake* (1951), is the work of an author expert in the naval and exploratory history of the period. E.D.S. BRADFORD, *Drake* (1965), is light and readable. Books placing Drake in his contemporary setting include J.A. WILLIAMSON, *The Age of Drake*, 5th ed. (1965); JULIAN CORBETT's old but still valuable classic, *Drake and the Tudor Navy*, 2 vol. (1898); and K.R. ANDREWS, *Drake's Voyages: A Reassessment of Their Place in Elizabethan Maritime Expansion* (1967). Collections of documents, such as letters and narratives referring to the voyages, include *New Light on Drake: A Collection of Documents Relating to His Voyage of Circumnavigation, 1577–1580*, ed. by ZELIA NUTTALL (1914); *The World Encompassed: Analogous Contemporary Documents Concerning Sir Francis Drake's Circumnavigation of the World*, ed. by N.M. PENZER (1926, reprinted 1969); *The Last Voyage of Drake and Hawkins*, ed. by K.R. ANDREWS (1972); and CALIFORNIA HISTORICAL SOCIETY, *Drake's Plate of Brass: Evidence of His Visit to California in 1597* (1937).

(E.Br.)

Drake Passage

The Drake Passage is a deep waterway, 600 miles (1,000 kilometres) wide, connecting the Atlantic and Pacific

oceans south of Tierra del Fuego, on which stands Cape Horn, the southern tip of South America. To the south, the waterway is bounded by the South Shetland Islands, situated about 100 miles north of the Antarctic Peninsula. Across this stretch of ocean the climate changes from the cool, humid, subpolar type found at Tierra del Fuego to the frozen conditions of Antarctica.

The Drake Passage played an important part in the trade of the 19th and early 20th centuries before the opening of the Panama Canal in 1914. The stormy seas and icy conditions made the rounding of Cape Horn through the Drake Passage a rigorous test for ships and crews alike, especially for the sailing vessels of the day. With the advent of giant oil tankers and with the growing congestion of the Panama Canal, the Drake Passage may again become a major sea route.

Exploration

Though bearing the name of the famous 16th-century English seaman and explorer, the Drake Passage was, in fact, first traversed in 1615 by a Flemish expedition led by Willem Schouten. Drake did not sail through the passage but passed instead through the Straits of Magellan to the north of Tierra del Fugeo, although he was blown back into the northern latitudes of the passage by a Pacific storm.

Submarine topography and bottom deposits. The passage has an average depth of about 11,000 feet (3,400 metres), with deeper regions of up to 15,600 feet near the northern and southern boundaries.

The sediments on the sea floor result from the interplay of glacial debris from Antarctica, material of biological origin, and eroded material from South America. Proceeding southward, the sediments are first mainly sandy and clayey silts, and then, later, silts with ice-rafted material (*i.e.*, material dropped from floating glacial ice as it melts). Near 58° S is a zone of numerous manganese nodules. Sediments accumulate at the rate of four inches every 1,000 years near the boundaries of the passage and at a little more than an inch every 1,000 years in the centre. The manganese nodule zone is swept clear of sediments by swift bottom currents.

Climate and ice conditions. The winds over the passage are predominantly from the west and are most intense in the northern half. The most commonly observed wind speeds are from 12 to 23 miles per hour, but winds of more than 45 miles per hour also occur.

The mean annual air temperature ranges from 41° F (5° C) in the north to 27° F (−3° C) in the south. The coldest temperatures, of −4° F (−20° C), occur in July. Cyclones (atmospheric low-pressure systems with winds that blow clockwise in the Southern Hemisphere) formed in the Pacific Ocean traverse the passage at the southern end. Cloud cover remains over the passage for about six-tenths of the year.

The sea ice cover extending northward from Antarctica varies from year to year as well as seasonally. In the late summer (February) the passage is ice free. In September, the maximum ice cover occurs; 25 percent to full cover extends to 60° S, with occasional ice floes reaching Cape Horn. Icebergs have been observed at different times across the whole extent of the passage and represent a danger to shipping.

Hydrography, temperature, and salinity. The water within the passage flows from the Pacific into the Atlantic, except for a small amount of water in the south that comes from the Scotia Sea. The general movement, known as the Antarctic Circumpolar Current, is the most voluminous in the world, moving at a rate of nearly 5,300,000,000 cubic feet a second. Surface water temperature varies from near 43° F (6° C) in the north to 30° F (−1° C) in the south, with the temperature altering sharply near 60° S. This transitional zone is called the Antarctic Convergence, or Polar Front; it separates the sub-Antarctic surface water from the colder and fresher Antarctic surface water. At depths of between about 1,600 and 10,000 feet (500 to 3,000 metres) there occurs the relatively warm and salty Circumpolar Deep Water Current.

Surface currents

The Antarctic Circumpolar Current, which is slightly accelerated by the passage, is strongest in the vicinity of the convergence. Near Cape Horn is a fairly swift coastal flow that is derived from a current that hugs the Chilean coast. After traversing the passage, the Circumpolar Current turns sharply to the north.

Throughout the passage as a whole, salinity and the amount of oxygen increases as one proceeds northward.

Fish and other wildlife. Photographs of the sea floor show the presence of numerous animals—either directly or indirectly by their tracks and fecal remains. The most common of these are such organisms as sea urchins and starfish; sponges are also found. There is also a relatively high abundance of plankton. Krill (small shrimplike crustaceans) are very abundant in the south, as are the blue and fin whales that feed on them. Squid also feed on krill and, in turn, form a basic food for the sperm whales. Krill are also important to the diet of the emperor penguin and the crabeater seal. Antarctic cods are the most common fish. Some of these fish and all of the cold water Chaenichthyidae (icefish) are "bloodless"—*i.e.*, they have no detectable hemoglobin in their blood.

Prospects for the future. The economic significance of the Drake Passage will most likely be confined to its importance as a shipping route. The mining of materials on the sea floor will probably remain uneconomic for a long time because of the difficulties of deep-sea mining in a polar environment. The importance of the passage to Antarctic oceanography and meteorology has made it the subject of frequent investigation.

BIBLIOGRAPHY. T. HATHERTON (ed.), *Antarctica* (1965), a general description of various scientific aspects of Antarctica, including oceanography; V.C. BUSHNELL and J.W. HEDGPETH (eds.), *Distribution of Selected Groups of Marine Invertebrates in Waters South of 35° S. Latitude*, Antarctic Map Folio Series no. 11 (1969), a detailed description of plankton found in Antarctic waters; A.L. GORDON and R.D. GOLDBERG, *Circumpolar Characteristics of Antarctic Waters*, ed. by V.C. BUSHNELL, Antarctic Map Folio Series no. 13 (1970), on the general distributions of temperature, salinity, and oxygen at various depths within Antarctic waters; ALAN VILLIERS, "The Defeat of Cape Horn," *The Oceans*, 2:6–16 (1969), a thrilling account of the history of man's navigation around the stormy tip of South America.

(A.L.G.)

Dramatic Literature

The term dramatic literature implies a contradiction in that "literature" originally meant something written and "drama" meant something performed. Most of the problems, and much of the interest, in the study of dramatic literature stem from this contradiction. Even though a play may be appreciated solely for its qualities as writing, greater rewards probably accrue to those who remain alert to the volatility of the play as a whole.

In order to appreciate this complexity in drama, however, each of its elements—acting, directing, staging, etc. —should be studied, so that its relationship to all the others can be fully understood. It is the purpose of this article to study drama with particular attention to what the playwright sets down. A similar approach is taken in the separate articles on the two main types of dramatic literature, TRAGEDY, and COMEDY. For the dramatic literature of particular genres, the reader also is referred to such articles as NO THEATRE; KABUKI THEATRE; COMMEDIA DELL'ARTE; and RELATED ENTRIES under DRAMATIC LITERATURE in the *Ready Reference and Index*. Thie history of dramatic literature is discussed in such articles as THEATRE, WESTERN; and THEATRE AND DANCE, EAST ASIAN; as well as in articles on the history of literature such as LITERATURE, WESTERN. Regional studies of both the theatre and literature will be found in articles such as SOUTH ASIAN PEOPLES, ARTS OF; and AFRICAN PEOPLES, ARTS OF.

GENERAL CHARACTERISTICS

Essential elements of a play

From the inception of a play in the mind of its author to the image of it that an audience takes away from the theatre, many hands and many physical elements help to bring it to life. Questions therefore arise as to what is and what is not essential to it. Is a play what its author thought he was writing, or the words he wrote? Is a play the way in which those words are intended to be em-

bodied, or their actual interpretation by a director and his actors on a particular stage? Is a play in part the expectation an audience brings to the theatre, or is it the real response to what is seen and heard? Since drama is such a complex process of communication, its study and evaluation is as uncertain as it is mercurial.

All plays depend upon a general agreement by all participants—author, actors, and audience—to accept the operation of theatre and the conventions associated with it, just as players and spectators accept the rules of a game. Drama is a decidedly unreal activity, which can be indulged only if everyone involved admits it. Here lies some of the fascination of its study. For one test of great drama is how far it can take the spectator beyond his own immediate reality and to what use this imaginative release can be put. But the student of drama must know the rules with which the players began the game before he can make this kind of judgment. These rules may be conventions of writing, acting, or audience expectation. Only when all conventions are working together smoothly in synthesis, and the make-believe of the experience is enjoyed passionately with mind and emotion, can great drama be seen for what it is: the combined work of a good playwright, good players, and a good audience who have come together in the best possible physical circumstances.

Drama in some form is found in almost every society, primitive and civilized, and has served a wide variety of functions in the community. There are, for example, records of a sacred drama in Egypt 2,000 years before Christ, and Thespis in the 6th century BC in ancient Greece is accorded the distinction of being the first known playwright. Elements of drama such as mime and dance, costume and decor long preceded the introduction of words and the literary sophistication now associated with a play. Moreover, such basic elements were not superseded by words, merely enhanced by them. Nevertheless, it is only when a playscript assumes a disciplinary control over the dramatic experience that the student of drama gains measurable evidence of what was intended to constitute the play. Only then can dramatic literature be discussed as such.

The late arrival of words in drama

The texts of plays indicate the different functions they served at different times. Some plays embraced nearly the whole community in a specifically religious celebration, as when all the male citizens of a Greek city-state came together to honour their gods; or when the annual Feast of Corpus Christi was celebrated with the great medieval Christian mystery cycles. On the other hand, the ceremonious temple ritual of the early Nō drama of Japan was performed at religious festivals only for the feudal aristocracy. But the drama may also serve a more directly didactic purpose, as did the morality plays of the later Middle Ages, some 19th-century melodramas, and the 20th-century discussion plays of George Bernard Shaw and Bertolt Brecht. Plays can satirize society, or they can gently illuminate human weakness; they can divine the greatness and the limitations of man in tragedy, or, in modern naturalistic playwriting, probe his mind. Drama is the most wide-ranging of all the arts: it not only represents life but also is a way of seeing it. And it repeatedly proves Dr. Samuel Johnson's contention that there can be no certain limit to the modes of composition open to the dramatist.

Common elements of drama. Despite the immense diversity of drama as a cultural activity, all plays have certain elements in common. For one thing, drama can never become a "private" statement—in the way a novel or a poem may be—without ceasing to be meaningful theatre. The characters may be superhuman and godlike in appearance, speech, and deed or grotesque and ridiculous, perhaps even puppets, but as long as they behave in even vaguely recognizable human ways the spectator can understand them. Only if they are too abstract do they cease to communicate as theatre. Thus, the figure of Death in medieval drama reasons like a human being, and a god in Greek tragedy or in Shakespeare talks like any mortal. A play, therefore, tells its tale by the imitation of human behaviour. The remoteness or nearness of that behaviour to the real life of the audience can importantly affect the response of that audience: it may be in awe of what it sees, or it may laugh with detached superiority at clownish antics, or it may feel sympathy. These differences of alienation or empathy are important, because it is by opening or closing this aesthetic gap between the stage and the audience that a dramatist is able to control the spectator's experience of the play and give it purpose.

The second essential is implicit in the first. Although static figures may be as meaningfully symbolic on a stage as in a painting, the deeper revelation of character, as well as the all-important control of the audience's responses, depends upon a dynamic presentation of the figures in action. A situation must be represented on the stage, one recognizable and believable to a degree, which will animate the figures as it would in life. Some argue that action is the primary factor in drama, and that character cannot emerge without it. Since no play exists without a situation, it appears impossible to detach the idea of a character from the situation in which he is placed, though it may seem possible after the experience of the whole play. Whether the playwright conceives character before situation, or vice versa, is arbitrary. More relevant are the scope and scale of the character-in-situation—whether, for example, it is man confronting God or man confronting his wife—for that comes closer to the kind of experience the play is offering its audience. Even here one must beware of passing hasty judgment, for it may be that the grandest design for heroic tragedy may be less affecting than the teasing vision of human madness portrayed in a good farce.

The role of style

A third factor is style. Every play prescribes its own style, though it will be influenced by the traditions of its theatre and the physical conditions of performance. Style is not something imposed by actors upon the text after it is written, nor is it superficial to the business of the play. Rather, it is self-evident that a play will not communicate without it. Indeed, many a successful play has style and little else. By "style," therefore, is implied the whole mood and spirit of the play, its degree of fantasy or realism, its quality of ritualism or illusion, and the way in which these qualities are signalled by the directions, explicit or implicit, in the text of the play. In its finer detail, a play's style controls the kind of gesture and movement of the actor, as well as his tone of speech, its pace and inflexion. In this way the attitude of the audience is prepared also: nothing is more disconcerting than to be misled into expecting either a comedy or a tragedy and to find the opposite, although some great plays deliberately introduce elements of both. By means of signals of style, the audience may be led to expect that the play will follow known paths, and the pattern of the play will regularly echo the rhythm of response in the auditorium. Drama is a conventional game, and spectators cannot participate if the rules are constantly broken.

By presenting animate characters in a situation with a certain style and according to a given pattern, a playwright will endeavour to communicate his thoughts and feelings and have his audience consider his ideas or reproduce the emotion that drove him to write as he did. In theatrical communication, however, audiences remain living and independent participants. In the process of performance, an actor has the duty of interpreting his author for the people watching him, and will expect to receive "feedback" in turn. The author must reckon with this in his writing. Ideas will not be accepted, perhaps, if they are offered forthrightly; and great dramatists who are intent on furthering social or political ideas, such as Henrik Ibsen, George Bernard Shaw, and Bertolt Brecht, quickly learned methods of having the spectator reason the ideas for himself as part of his response to the play. Nor will passions necessarily be aroused if overstatement of feeling ("sentimentality") is used without a due balance of thinking and even the detachment of laughter: Shakespeare and Chekhov are two outstanding examples in Western drama of writers who achieved an exquisite balance of pathos with comedy in order to ensure the affective function of their plays.

Dramatic expression. The language of drama can range between great extremes: on the one hand, an intensely theatrical and ritualistic manner; and on the other, an almost exact reproduction of real life of the kind commonly associated with motion picture and television drama. In the ritualistic drama of ancient Greece, the playwrights wrote in verse, and it may be assumed that their actors rendered this in an incantatory speech halfway between speech and song. Both the popular and the coterie drama of the Chinese and Japanese theatre were also essentially operatic, with a lyrical dialogue accompanied by music and chanted rhythmically. The effect of such rhythmical delivery of the words was to lift the mood of the whole theatre onto the level of religious worship. Verse is employed in other drama that is conventionally elevated, like the Christian drama of the Middle Ages, the tragedy of the English Renaissance, the heroic Neoclassical tragedies of 17th-century France by Pierre Corneille and Jean Racine, the Romantic lyricism of Goethe and Schiller, and modern attempts at a revival of a religious theatre like those of T.S. Eliot. Indeed, plays written in prose dialogue were at one time comparatively rare, and then associated essentially with the comic stage. Only at the end of the 19th century, when naturalistic realism became the mode, were characters in dramas expected to speak as well as behave as in real life.

The role of verse

Elevation is not the whole rationale behind the use of verse in drama. Some critics maintain that a playwright can exercise better control both over the speech and movement of his actors and over the responses of his audience by using the more subtle tones and rhythms of good poetry. The loose, idiomatic rhythms of ordinary conversation, it is argued, give both actor and spectator too much freedom of interpretation and response. Certainly, the aural, kinetic, and emotive directives in verse are more direct than prose, though, in the hands of a master of prose dialogue like Shaw or Chekhov, prose can also share these qualities. Even more certain, the "aesthetic distance" of the stage, or the degree of unreality and make-believe required to release the imagination, is considerably assisted if the play uses elements of verse, like rhythm and rhyme, not found in ordinary speech. Thus, verse drama may embrace a wide variety of nonrealistic aural and visual devices: Greek tragic choric speech provided a philosophical commentary upon the action, which at the same time drew the audience lyrically into the mood of the play. In the drama of India, a verse accompaniment made the actors' highly stylized system of symbolic gestures of head and eyes, arms and fingers a harmonious whole. The tragic soliloquy in Shakespeare permitted the hero, alone on the stage with his audience, to review his thoughts aloud in the persuasive terms of poetry; thus, the soliloquy was not a stopping place in the action but rather an engrossing moment of drama when the spectator's mind could leap forward.

Dramatic structure. The elements of a play do not combine naturally to create a dramatic experience but, rather, are made to work together through the structure of a play, a major factor in the total impact of the experience. A playwright will determine the shape of a play in part according to the conditions in which it will be performed: how long should it take to engage an audience's interest and sustain it? How long can an audience remain in their seats? Is the audience sitting in one place for the duration of performance, or is it moving from one pageant stage to the next, as in some medieval festivals? Structure is also dictated by the particular demands of the material to be dramatized: a revue sketch that turns on a single joke will differ in shape from a religious cycle, which may portray the whole history of mankind from the Creation to the Last Judgment. A realistic drama may require a good deal of exposition of the backgrounds and memories of the characters, while in a chronicle play the playwright may tell the whole story episodically from its beginning to the end. There is one general rule, as Aristotle originally suggested in his *Poetics:* a play must be long enough to supply the information an audience needs to be interested and to generate the experience of tragedy, or comedy, on the senses and imagination.

In the majority of plays it is necessary to establish a conventional code of place and time. In a play in which the stage must closely approximate reality, the location of the action will be precisely identified, and the scenic representation on stage must confirm the illusion. In such a play, stage time will follow chronological time almost exactly; and if the drama is broken into three, four, or five acts, the spectator will expect each change of scene to adjust the clock or the calendar. But the theatre has rarely expected realism, and by its nature it allows an extraordinary freedom to the playwright in symbolizing location and duration: as Dr. Samuel Johnson observed in his discussion of this freedom in Shakespeare, the spectators always allow the play to manipulate the imagination. It is sufficient for the witches in *Macbeth* to remark their "heath" with its "fog and filthy air" for their location to be accepted on a stage without scenery; and when Lady Macbeth later is seen alone reading a letter, she is without hesitation understood to be in surroundings appropriate to the wife of a Scottish nobleman. Simple stage symbolism may assist the imagination, whether the altar of the gods situated in the centre of the Greek *orchēstra*, a strip of red cloth to represent the Red Sea in a medieval miracle play, or a chair on which the Tibetan performer stands to represent a mountain. With this degree of fantasy, it is no wonder that the theatre can manipulate time as freely, passing from the past to the future, from this world to the next, and from reality to dream.

Conventions of time and place

It is questionable, therefore, whether the notion of "action" in a play describes what happens on the stage or what is recreated in the mind of the audience. Certainly it has little to do with merely physical activity by the players. Rather, anything that urges forward the audience's image of the play and encourages the growth of its imagination is a valid part of the play's action. Thus, it was sufficient for the ancient Greek dramatist Aeschylus to have only two speaking male actors who wore various masks, typed for sex, age, class, and facial expression. In the Italian 16th- and 17th-century commedia dell'arte, the standard characters Pantalone and Arlecchino, each wearing his traditional costume and mask, appeared in play after play and were immediately recognized, so that an audience could anticipate the behaviour of the grasping old merchant and his rascally servant. On a less obvious level, a speech that in reading seems to contribute nothing to the action of the play can provide in performance a striking stimulus to the audience's sense of the action, its direction and meaning. Thus, both the Greek chorus and the Elizabethan actor in soliloquy might be seen to "do" nothing, but their intimate speeches of evaluation and reassessment teach the spectator how to think and feel about the action of the main stage and lend great weight to the events of the play. For drama is a reactive art, moving constantly in time, and any convention that promotes a deep response while conserving precious time is of immeasurable value.

Conventions of action

DRAMA AS AN EXPRESSION OF A CULTURE

In spite of the wide divergencies in purpose and convention of plays as diverse as the popular kabuki of Japan and the coterie comedies of the Restoration in England, a Javanese puppet play and a modern social drama by the contemporary American dramatist Arthur Miller, all forms of dramatic literature have some points in common. Differences between plays arise from differences in conditions of performance, in local conventions, in the purpose of theatre within the community, and in cultural history. Of these, the cultural background is the most important, if the most elusive. It is cultural difference that makes the drama of the East immediately distinguishable from that of the West.

East–West differences. Oriental drama consists chiefly of the classical theatre of Hindu India and its derivatives in Malaya and of Burma, Thailand, China, Japan, Java, and Bali. It was at its peak during the period known in

the West as the Middle Ages and the Renaissance. Stable and conservative, perpetuating its customs with reverence, Oriental culture showed little of the interest in chronology and advancement shown by the West and placed little emphasis on authors and their individual achievements. Thus the origins of the drama of the Orient are lost in time, although its themes and characteristic styles probably remain much the same as before records were kept. The slow-paced, self-contained civilizations of the East have only recently been affected by Western theatre, just as the West has only recently become conscious of the theatrical wealth of the East and what it could do to fertilize the modern theatre (as in the 20th-century experimental drama of William Butler Yeats and Thornton Wilder in English, of Paul Claudel and Antonin Artaud in French, and of Bertolt Brecht in German).

Stylization of Oriental theatre

In its representation of life, classical Oriental drama is the most conventional and nonrealistic in world theatre. Performed over the centuries by actors devoted selflessly to the profession of a traditional art, conventions of performance became highly stylized, and traditions of characterization and play structure became formalized to a point of exceptional finesse, subtlety, and sophistication. In Oriental drama all the elements of the performing arts are made by usage to combine to perfection: dance and mime, speech and song, narrative and poetry. The display and studied gestures of the actors, their refined dance patterns, and the all-pervasive instrumental accompaniment to the voices of the players and the action of the play, suggest to Western eyes an exquisite combination of ballet with opera, in which the written text assumes a subordinate role. In this drama, place could be shifted with a license that would have astonished the most romantic of Elizabethan dramatists, the action could leap back in time in a way reminiscent of the "flashback" of the modern cinema, and events could be telescoped with the abandon of modern expressionism. This extreme theatricality lent an imaginative freedom to its artists and audiences upon which great theatre could thrive. Significantly, most Oriental cultures also nourished a puppet theatre, in which stylization of character, action, and staging were particularly suitable to marionettes. In the classical puppet theatre of Japan, the *bunraku*, the elocutionary art of a chanted narration and the manipulative skill with the dolls diminished the emphasis on the script except in the work of the 17th-century master Chikamatsu, who enjoyed a creative freedom in writing for puppets rather than for the actors of the Kabuki. By contrast, Western drama during and after the Renaissance has offered increasing realism, not only in decor and costume but also in the treatment of character and situation.

It is generally thought that Oriental drama, like that of the West, had its beginnings in religious festivals. Dramatists retained the moral tone of religious drama while using popular legendary stories to imbue their plays with a romantic and sometimes sensational quality. This was never the sensationalism of novelty that Western dramatists sometimes used: Eastern invention is merely a variation on what is already familiar, so that the slightest changes of emphasis could give pleasure to the cognoscenti. This kind of subtlety is not unlike that found in the repeatedly depicted myths of Greek tragedy. What is always missing in Oriental drama is that restlessness for change characteristic of modern Western drama. In the West, religious questioning, spiritual disunity, and a belief in the individual vision combined finally with commercial pressures to produce comparatively rapid changes. None of the moral probing of Greek tragedy, the character psychology of Shakespeare and Racine, the social and spiritual criticism of Ibsen and Strindberg, nor the contemporary drama of shock and argument, is imaginable in the classical drama of the East.

Drama in Western cultures. The form and style of ancient Greek tragedy, which flowered in the 5th century BC in Athens, was dictated by its ritual origins and by its performance in the great dramatic competitions of the spring and winter festivals of Dionysus. Participation in ritual requires that the audience largely knows what to expect. Ritual dramas were written on the same legendary stories of Greek heroes in festival after festival. Each new drama provided the spectators with a reassessment of the meaning of the legend along with a corporate religious exercise. Thus, the chorus of Greek tragedy played an important part in conveying the dramatist's intention. The chorus not only provided a commentary on the action but also guided the moral and religious thought and emotion of the audience throughout the play: for Aeschylus (*c.* 525–456 BC) and Sophocles (*c.* 496–406 BC) it might be said that the chorus *was* the play, and even for Euripides (*c.* 480–406 BC) it remained lyrically powerful. Other elements of performance also controlled the dramatist in the form and style he could use in these plays: in particular, the great size of the Greek arena demanded that the players make grand but simple gestures and intone a poetry that could never approach modern conversational dialogue. Today, the superhuman characters of these plays, Agamemnon and Clytemnestra, Orestes and Electra, Oedipus and Antigone, seem unreal, for they display little "characterization" in the modern sense and their fates are sealed. Nevertheless, these great operatic tableaux, built, as one critic has said, for weight and not speed, were evidently able to carry their huge audiences to a catharsis of feeling. It is a mark of the piety of those audiences that the same reverent festivals supported a leavening of satyr-plays and comedies, bawdy and irreverent comments on the themes of the tragedies, culminating in the wildly inventive satires of Aristophanes (*c.* 445–*c.* 385 BC.)

Use of familiar legends

The study of Greek drama demonstrates how the ritual function of theatre shapes both play and performance. This ritual aspect was lost when the Romans assimilated Greek tragedy and comedy. The Roman comedies of Plautus (*c.* 254–184 BC) and Terence (*c.* 186/185–159 BC) were brilliant but inoffensive entertainments, while the oratorical tragedies of Seneca (*c.* 4 BC–AD 65) on themes from the Greek were written probably only to be read by the ruling caste. Nevertheless, some of the dramatic techniques of these playwrights influenced the shape and content of plays of later times. The bold prototype characters of Plautus (the boasting soldier, the old miser, the rascally parasite), with the intricacies of his farcical plotting, and the sensational content and stoical attitudes of Seneca's drama reappeared centuries later when classical literature was rediscovered.

Establishment of prototypes of characters

Western drama had a new beginning in the medieval church, and, again, the texts reflect the ritual function of the theatre in society. The Easter liturgy, the climax of the Christian calendar, explains much of the form of medieval drama as it developed into the giant mystery cycles. From at least the 10th century the clerics of the church enacted the simple Latin liturgy of the *Quem quaeritis?* (literally "whom do you seek?"), the account of the visit to Jesus Christ's tomb by the three Marys, who are asked this question by an angel. The liturgical form of Lent and the Passion, indeed, embodies the drama of the Resurrection to be shared mutually by actor-priest and audience-congregation. When the Feast of Corpus Christi was instituted in 1246, the great lay cycles of Biblical plays (the mystery or miracle cycles) developed rapidly, eventually treating the whole story of man from the Creation to the Last Judgment, with the Crucifixion still the climax of the experience. The other influence controlling their form and style was their manner of performance. The vast quantity of material that made up the story was broken into many short plays, and each was played on its own stage in the vernacular by members of the craft guilds. Thus, the authors of these dramas gave their audience not a mass communal experience, as the Greek dramatists had done, but rather many small and intimate dramatizations of the Bible story. In stylized and alliterative poetry, they mixed awesome events with moments of extraordinary simplicity, embodying local details, familiar touches of behaviour, and the comedy and the cruelty of medieval life. Their drama consists of strong and broad contrasts, huge in perspective but meaningful in human terms, religious and ap-

propriately didactic in content and yet popular in its manner of reaching its simple audiences.

In an account of dramatic literature, the ebullient but unscripted farces and romances of the commedia dell'arte properly have no place, but much in it became the basis of succeeding comedy. Two elements are worth noting. First, the improvisational spirit of the commedia troupes, in which the actor would invent words and comic business (*lazzi*) to meet the occasion of the play and the audience he faced, encouraged a spontaneity in the action that has affected the writing and playing of Western comedy ever since. Second, basic types of comic character derived from the central characters, who reappeared in the same masks in play after play. As these characters became well known everywhere, dramatists could rely on their audience to respond to them in predictable fashion. Their masks stylized the whole play and allowed the spectator freedom to laugh at the unreality of the action. An understanding of the commedia illuminates a great deal in the written comedies of Shakespeare in England, of Molière and Marivaux in France, and of Goldoni and Gozzi in Italy.

The influence of improvisation

In the 16th century, England and Spain provided all the conditions necessary for a drama that could rival ancient Greek drama in scope and subtlety. In both nations, there were public as well as private playhouses, audiences of avid imagination, a developing language that invited its poetic expansion, a rapid growth of professional acting companies, and a simple but flexible stage. All these factors combined to provide the dramatist with an opportunity to create a varied and exploratory new drama of outstanding interest. In Elizabethan London, dramatists wrote in an extraordinary range of dramatic genres, from native comedy and farce to Senecan tragedy, from didactic morality plays to popular chronicle plays and tragicomedies, all before the advent of Shakespeare (1564–1616). Although Shakespeare developed certain genres, such as the chronicle play and the tragedy, to a high degree, Elizabethan dramatists characteristically used a medley of styles. With the exception of Ben Jonson (1572/73–1637) and a few others, playwrights mixed their ingredients without regard for classical rule. The result was a rich body of drama, exciting and experimental in character. A host of new devices were tested, mixing laughter and passion; shifting focus and perspective by slipping from verse to prose and back again; extending the use of the popular clown; exploiting the double values implicit in boy actors playing the parts of girls; exploring the role of the actor in and out of character; but, above all, developing an extraordinarily flexible dramatic poetry. These dramatists produced a visually and aurally exciting hybrid drama that could stress every subtlety of thought and feeling. It is not surprising that they selected their themes from every Renaissance problem of order and authority, of passion and reason, of good and evil and explored every comic attitude to people and society with unsurpassed vigour and vision.

Quite independently in Spain, dramatists embarked upon a parallel development of genres ranging from popular farce to chivalric tragedy. The hundreds of plays of Spain's greatest playwright, Lope de Vega (1562–1635), cover every subject from social satire to religion with equal exuberance. The drama of Paris of the 17th century, however, was determined by two extremes of dramatic influence. On the one hand, some playwrights developed a tragedy rigidly based in form upon Neoclassical notions of Aristotelian unity, controlled by verse that is more regular than that of the Spanish or English dramatists. On the other hand, the French theatre developed a comedy strongly reflecting the work of the itinerant troupes of the commedia dell'arte. The Aristotelian influence resulted in the plays of Pierre Corneille (1606–1684) and Jean Racine (1639–1699), tragedies of honour using classical themes, highly sophisticated theatrical instruments capable of searching deeply into character and motive, and capable of creating the powerful tension of a tightly controlled plot. The other influence produced the brilliant plays of Molière (1622–1673), whose training as an actor in the masked and balletic commedia tradition supplied him with a perfect mode for a more sophisticated comedy. Molière's work established the norm of French comedy, bold in plotting, exquisite in style, irresistible in comic suggestion. Soon after, upon the return of Charles II to the throne of England in 1660, a revival of theatre started the English drama on a new course. Wits such as William Wycherley (1640–1716) and William Congreve (1670–1729) wrote for the intimate playhouses of the Restoration and an unusually homogeneous coterie audience of the court circle. They developed a "comedy of manners," replete with social jokes that the actor, author, and spectator could share—a unique phase in the history of drama. These plays started a characteristic style of English domestic comedy still recognizable in London comedy today.

French drama in the 17th century

German dramatists of the later part of the 18th century achieved stature through a quite different type of play: Johann Wolfgang von Goethe (1749–1832), Johann Christoph Friedrich von Schiller (1759–1805), and others of the passionate, poetic Sturm und Drang ("storm and stress") movement tried to echo the more romantic tendencies in Shakespeare's plays. Dramatists of the 19th century, however, lacking the discipline of classical form, wrote derivative melodramas that varied widely in quality, often degenerating into mere sensationalism. Melodrama rapidly became the staple of the theatre across Europe and America. Bold in plotting and characterization, simple in its evangelical belief that virtue will triumph and providence always intervene, it pleased vast popular audiences and was arguably the most prolific and successful drama in the history of the theatre. Certainly, melodrama's elements of essential theatre should not be ignored by those interested in drama as a social phenomenon. At least melodramas encouraged an expansion of theatre audiences ready for the most recent phase in dramatic history.

Melodrama

The time grew ripe for a new and more adult drama at the end of the 19th century. As novelists developed greater naturalism in both content and style, dramatists too looked to new and more realistic departures: the dialectical comedies of ideas of George Bernard Shaw (1856–1950); the problem plays associated with Henrik Ibsen (1828–1906); the more lyrical social portraits of Anton Chekhov (1860–1904); the fiercely personal, social, and spiritual visions of August Strindberg (1849–1912). These dramatists began by staging the speech and behaviour of real life, in devoted detail, but became more interested in the symbolic and poetic revelation of the human condition. Where Ibsen began by modelling his tightly structured dramas of man in society upon the formula for the "well made" play, which carefully controlled the audience's interest to the final curtain, Strindberg, a generation later, developed a free psychological and religious dream play that bordered on Expressionism. As sophisticated audiences grew interested more in causes rather than in effects, the great European playwrights of the turn of the century mixed their realism increasingly with symbolism. Thus the Naturalistic movement in drama, though still not dead, had a short but vigorous life. Its leaders freed the drama of the 20th century to pursue every kind of style, and subsequent dramatists have been wildly experimental. The playwright today can adopt any dramatic mode, mixing his effects to shock the spectator into an awareness of himself, his beliefs, and his environment.

Drama in Eastern cultures. Because of its inborn conservatism, the dramatic literature of the East does not show such diversity, despite its variety of cultures and subcultures. The major features of Oriental drama may be seen in the three great classical sources of India, China, and Japan. The simplicity of the Indian stage, a platform erected for the occasion in a palace or a courtyard, like the simplicity of the Elizabethan stage, lent great freedom to the imagination of the playwright. In the plays of India's greatest playwright, Kālidāsa (probably 4th century AD), there is an exquisite refinement of detail in presentation. His delicate romantic tales leap time and place by simple suggestion and mingle courtly humour and light-hearted wit with charming sentiment

Indian theatre

and religious piety. Quite untrammelled by realism, lyrical in tone and refined in feeling, his fanciful love and adventure stories completely justify their function as pure entertainment. His plots are without the pain of reality, and his characters never descend from the ideal: such poetic drama is entirely appropriate to the Hindu aesthetic of blissful idealism in art.

Some contrast may be felt between the idealistic style of the Sanskrit drama and the broader, less courtly manner of the Chinese and its derivatives in Southeast Asia. These plays cover a large variety of subjects and styles, but all combine music, speech, song, and dance, as does all Oriental drama. Heroic legends, pathetic moral stories, and brilliant farces all blended spectacle and lyricism and were as acceptable to a sophisticated court audience as to a popular street audience. The most important Chinese plays stem from the Yüan dynasty (1279–1368), in which an episodic narrative is carefully structured and unified. Each scene introduces a song whose lines have a single rhyme, usually performed by one singer, with a code of symbolic gestures and intonations that has been refined to an extreme. The plays have strongly typed heroes and villains, simple plots, scenes of bold emotion, and moments of pure mime. Chinese drama avoided both the crudity of European melodrama and the esotericism of Western coterie drama.

Nō and Kabuki

The drama of Japan may be said to embrace both. There, the exquisite artistry of gesture and mime, and the symbolism of setting and costume, took two major directions. The Nō drama, emerging from religious ritual, maintained a special refinement appropriate to its origins and its aristocratic audiences; the Kabuki (its name suggesting its composition: *ka*, "singing"; *bu*, "dancing"; *ki*, "acting") in the 17th century became Japan's popular drama. Nō theatre is reminiscent of the religious tragedy of the Greeks in the remoteness of its legendary content, in its masked heroic characters, in its limit of two actors and a chorus, and in the static, oratorical majesty of its style. The Kabuki, on the other hand, finds its material in domestic stories and in popular history, and the actors, without masks, move and speak more freely, without seeming to be realistic. The Kabuki plays are less rarefied and are often fiercely energetic and wildly emotional as befitting their presentation before a broader audience. The written text of the Nō play is highly poetic and pious in tone, compressed in its imaginative ideas, fastidious and restrained in verbal expression, and formal in its sparse plotting; the text of a Kabuki play lends plentiful opportunities for spectacle, sensation, and melodrama. In the Kabuki there can be moments of realism, but also whole episodes of mime and acrobatics; there can be moments of slapstick, but also moments of violent passion. In all, the words are subordinate to performance in the Kabuki.

Drama and communal belief. The drama that is most meaningful and pertinent to its society is that which arises from it and is not imposed upon it. The religious drama of ancient Greece, the temple drama of early India and Japan, the mystery cycles of medieval Europe, all have in common more than their religious content: when the theatre is a place of worship, its drama goes to the roots of belief in a particular community. The dramatic experience becomes a natural extension of man's life both as an individual and as a social being. The content of the mystery cycles speaks formally for the orthodox dogma of the church, thus seeming to place the plays at the centre of medieval life, like the church itself. Within such a comprehensive scheme, particular needs could be satisfied by comic or pathetic demonstration; for example, such a crucial belief as that of the Immaculate Conception was presented in the York (England) cycle of mystery plays, of the 14th–16th centuries, with a nicely balanced didacticism when Joseph wonders how a man of his age could have got Mary with child and an Angel explains what has happened; the humour reflects the simplicity of the audience and at the same time indicates the perfect faith that permitted the near-blasphemy of the joke. In the tragedies Shakespeare wrote for the Elizabethan theatre, he had the same gift of satisfying deep communal needs while meeting a whole range of individual interests present in his audience.

Didactic drama

When the whole community shares a common heritage, patriotic drama and drama commemorating national heroes, as are seen almost universally in the Orient, is of this kind. Modern Western attempts at a religious didactic drama, or indeed at any drama of "ideas," have had to reckon with the disparate nature of the audience. Thus the impact of Ibsen's social drama both encouraged and divided the development of the theatre in the last years of the 19th century. Plays like *A Doll's House* (1879) and *Ghosts* (published 1881), which challenged the sanctity of marriage and questioned the loyalty a wife owed to her husband, took their audiences by storm: some violently rejected the criticism of their cherished social beliefs, and thus such plays may be said to have failed to persuade general audiences to examine their moral position; on the other hand, there were sufficient numbers of enthusiasts (so-called Ibsenites) to stimulate a new drama of ideas. "Problem" plays appeared all over Europe and undoubtedly rejuvenated the theatre for the 20th century. Shaw's early Ibsenite plays in London, attacking a negative drawing-room comedy with themes of slum landlordism (*Widowers' Houses*, 1892) and prostitution (*Mrs. Warren's Profession*, 1902) resulted only in failure, but Shaw quickly found a comic style that was more disarming. In his attack on false patriotism (*Arms and the Man*, 1894) and the motives for middle class marriage (*Candida*, 1897), he does not affront his audiences before leading them by gentle laughter and surprise to review their own positions.

INFLUENCES ON THE DRAMATIST

The author of a play is affected, consciously or unconsciously, by the conditions under which he conceives and writes, by his social and economic status as a playwright, by his personal background, by his religious or political position, by his purpose in writing. The literary form of the play and its stylistic elements will be influenced by tradition, a received body of theory and dramatic criticism, as well as by the author's innovative energy. Auxiliary theatre arts such as music and design also have their own controlling traditions and conventions, which the playwright must respect. The size and shape of the playhouse, the nature of its stage and equipment, and the actor–audience relationship it encourages also determine the character of the writing. Not least, the audience's cultural assumptions, holy or profane, local or international, social or political, may override all else in deciding the form and content of the drama. These are large considerations that can take the student of drama into areas of sociology, politics, social history, religion, literary criticism, philosophy and aesthetics, and beyond.

The role of theory. It is difficult to assess the influence of theory since theory usually is based on existing drama, rather than drama on theory. Philosophers, critics, and dramatists have attempted both to describe what happens and to prescribe what should happen in drama, but all their theories are affected by what they have seen and read.

Western theory. In Europe, the earliest extant work of dramatic theory, the fragmentary *Poetics* of Aristotle (384–322 BC), chiefly reflecting his views on Greek tragedy and his favorite dramatist, Sophocles, is still relevant to an understanding of the elements of drama. Aristotle's elliptical way of writing, however, encouraged different ages to place their own interpretation upon his statements and to take as prescriptive what many believe to have been meant only to be descriptive. There has been endless discussion of his concepts *mimēsis* ("imitation"), the impulse behind all the arts, and *katharsis* ("purgation," "purification of emotion"), the proper end of tragedy, though these notions were conceived, in part, in answer to Plato's attack on *poiēsis* as an appeal to the irrational. That "character" is second in importance to "plot" is another of Aristotle's concepts that may be understood with reference to the practice of the Greeks, but not more realistic drama, in which character psychology has a dominant importance. The concept in the *Poetics* that has

most affected the composition of plays in later ages has been that of the so-called unities—that is, of time, place, and action. Aristotle was evidently describing what he observed—that a typical Greek tragedy had a single plot and action that lasts one day; he made no mention at all of unity of place. Neoclassical critics of the 17th century, however, codified these discussions into rules.

"Rules" of drama

Considering the inconvenience of such rules and their final unimportance, one wonders at the extent of their influence. The Renaissance desire to follow the ancients and its enthusiasm for decorum and classification may explain it in part. Happily, the other classical work recognized at this time was Horace's *Art of Poetry* (c. 24 BC), with its basic precept that poetry should offer pleasure and profit and teach by pleasing, a notion that has general validity to this day. Happily, too, the popular drama, which followed the tastes of its patrons, also exerted a liberating influence. Nevertheless, discussion about the supposed need for the unities continued throughout the 17th century (culminating in the French critic Nicolas Boileau's *Art of Poetry*, originally published in 1674), particularly in France, where a master like Racine could translate the rules into a taut, intense theatrical experience. Only in Spain, where Lope de Vega published his *New Art of Writing Plays* (1609), written out of his experience with popular audiences, was a commonsense voice raised against the classical rules, particularly on behalf of the importance of comedy and its natural mixture with tragedy. In England both Sir Philip Sidney in his *Apologie for Poetry* (1595) and Ben Jonson in *Timber* (1640) merely attacked contemporary stage practice. Jonson, in certain prefaces, however, also developed a tested theory of comic characterization (the "humours") that was to affect English comedy for a hundred years. The best of Neoclassical criticism in English is John Dryden's *Of Dramatic Poesie, an Essay* (1668). Dryden approached the rules with a refreshing honesty and argued all sides of the question; thus he questioned the function of the unities and accepted Shakespeare's practice of mixing comedy and tragedy.

The lively imitation of nature came to be acknowledged as the primary business of the playwright and was confirmed by the authoritative voices of Dr. Samuel Johnson, who said in his *Preface to Shakespeare* (1765) "there is always an appeal open from criticism to nature," and the German dramatist and critic Gotthold Ephraim Lessing, who in his *Hamburgische Dramaturgie* (or *Hamburg Dramaturgy;* 1767–69) sought to accommodate Shakespeare to a new view of Aristotle. With the classical straitjacket removed, there was a release of dramatic energies in new directions. There were still local critical skirmishes, such as Jeremy Collier's attack on the "immorality and profaneness of the English stage" in 1698; Goldoni's attacks upon the already dying Italian commedia on behalf of greater realism; and Voltaire's reactionary wish to return to the unities and to rhymed verse in French tragedy, which was challenged in turn by Diderot's call for a return to nature. But the way was open for the development of the middle class *drame* and the excursions of romanticism. Victor Hugo, in his Preface to his play *Cromwell* (1827), capitalized on the new psychological romanticism of Goethe and Schiller as well as the popularity of the sentimental *drame* in France and the growing admiration for Shakespeare; Hugo advocated truth to nature and a dramatic diversity that could yoke together the sublime and the grotesque. This view of what drama should be received support from Émile Zola in the preface to his play *Thérèse Raquin* (1873), in which he argued a theory of naturalism that called for the accurate observation of people controlled by their heredity and environment. From such sources came the subsequent intellectual approach of Ibsen and Chekhov and a new freedom for such seminal innovators of the 20th century as Luigi Pirandello, with his teasing mixtures of absurdist laughter and psychological shock; Bertolt Brecht (1898–1956), deliberately breaking the illusion of the stage; and Antonin Artaud (1896–1948), advocating a theatre that should be "cruel" to its audience, employing all and any devices that lie to hand. The modern dramatist may be grateful that he is no longer hidebound by theory and yet also regret, paradoxically, that the theatre of his time lacks those artificial limits within which an artifact of more certain efficiency can be wrought.

Twentieth-century views

Eastern theory. The Oriental theatre has always had such limits, but with neither the body of theory nor the pattern of rebellion and reaction found in the West. The Sanskrit drama of India, however, throughout its recorded existence has had the supreme authority of the *Nāṭya-śāstra,* ascribed to Bharata (*c.* 1st century AD), an exhaustive compendium of rules for all the performing arts, but particularly for the sacred art of drama with its auxiliary arts of dance and music. Not only does the *Nāṭya-śāstra* identify many varieties of gesture and movement but it also describes the multiple patterns that drama can assume, similar to a modern treatise on musical form. Every conceivable aspect of a play is treated, from the choice of metre in poetry to the range of moods a play can achieve; but perhaps its primary importance lies in its justification of the aesthetic of Indian drama as a vehicle of religious enlightenment.

In Japan, the most celebrated of early Nō writers, Zeami Motokiyo (1363–1443), left an influential collection of essays and notes to his son about his practice, and his deep knowledge of Zen Buddhism infused the Nō drama with ideals for the art that have persisted. Religious serenity of mind (*yūgen*), conveyed through an exquisite elegance in a performance of high seriousness, is at the heart of Zeami's theory of dramatic art. Three centuries later, the outstanding dramatist Chikamatsu (1653–1725) built equally substantial foundations for the Japanese puppet theatre, later known as the *bunraku.* His heroic plays for this theatre established an unassailable dramatic tradition of depicting an idealized life inspired by a rigid code of honour and expressed with extravagant ceremony and fervent lyricism. At the same time, in another vein, his pathetic "domestic" plays of middle class life and the suicides of lovers established a comparatively realistic mode for Japanese drama, which strikingly extended the range of both the *bunraku* and the Kabuki. Today, these forms, together with the more aristocratic and intellectual Nō, constitute a classical theatre based on practice rather than on theory. They may be superseded as a result of the recent invasion of Western drama, but in their perfection they are unlikely to change. The Yüan drama of China was similarly based upon a slowly evolved body of laws and conventions derived from practice, for, like the Kabuki of Japan, this too was essentially an actors' theatre, and practice rather than theory accounts for its development.

The role of music and dance. The Sanskrit treatise *Nāṭya-śāstra* suggests that drama had its origin in the art of dance, and any survey of Western theatre, too, must recognize a comparable debt to music in the classical Greek drama, which is believed to have sprung from celebratory singing to Dionysus. Similarly, the drama of the medieval church began with the chanted liturgies of the Roman mass. In the professional playhouses of the Renaissance and after, only rarely is music absent: Shakespeare's plays, particularly the comedies, are rich with song, and the skill with which he pursues dramatic ends with musical help is a study in itself. Molière conceived most of his plays as comedy-ballets, and much of his verbal style derives directly from the balletic qualities of the commedia. The popularity of opera in the 18th century led variously to John Gay's prototype for satirical ballad-opera, *The Beggar's Opera* (1728), the opera buffa in Italy, and the opéra comique in France. The development of these forms, however, resulted in the belittling of the written drama, with the notable exception of the parodistic wit of W.S. Gilbert (1836–1911). It is worth noting, however, that the most successful modern "musicals" lean heavily on their literary sources. Today, two of the strongest influences on contemporary theatre are those of Bertolt Brecht, who believed that a dialectical theatre should employ music not merely as a background embellishment but as an equal voice with the actor's, and of Antonin Artaud, who argued that the the-

atre experience should subordinate the literary text to mime, music, and spectacle. Since it is evident that drama often involves a balance of the arts, an understanding of their interrelationships is proper to a study of dramatic literature.

The influence of theatre design. Though apparently an elementary matter, the shape of the stage and auditorium probably offers the greatest single control over the text of the play that can be measured and tested. Moreover, it is arguable that the playhouse architecture dictates more than any other single factor the style of a play, the conventions of its acting, an the quality of dramatic effect felt by its audience. The shape of the theatre is always changing, so that to investigate its function is both to understand the past and to anticipate the future. Today, Western theatre is in the process of breaking away from the dominance of the Victorian picture-frame theatre, and therefore from the kind of experience this produced.

The contemporary English critic John Wain has called the difference between Victorian and Elizabethan theatre a difference between "consumer" and "participation" art. The difference resulted from the physical relationship between the audience and the actor in the two periods, a relationship that determined the kind of communication open to the playwright and the role the drama could play in society. Three basic playhouse shapes have emerged in the history of the theatre: the arena stage, the open stage, and the picture-frame.

Three basic playhouse types

The arena stage. To the arena, or theatre-in-the-round, belongs the excitement of the circus, the bullring, and such sports as boxing and wrestling. Arena performance was the basis for all early forms of theatre—the Druid ceremonies at Stonehenge, the Tibetan harvest-festival drama, probably early Greek ritual dancing in the *orchēstra*, the medieval rounds in 14th-century England and France, the medieval street plays on pageant wagons, the early Nō drama of Japan, the royal theatre of Cambodia. Characteristic of all these theatres is the bringing together of whole communities for a ritual experience; therefore, a sense of ritualistic intimacy and involvement is common to the content of the drama, and only the size of the audience changes the scale of the sung or spoken poetry. Clearly, the idiom of realistic dialogue would have been inappropriate both to the occasion and the manner of such theatre.

The open stage. When more narrative forms of action appeared in drama and particular singers or speakers needed to control the attention of their audience by facing them, the open, "thrust," or platform stage, with the audience on three sides of the actor, quickly developed its versatility. Intimate and ritualistic qualities in the drama could be combined with a new focus on the players as individual characters. The open stage and its variants were used by the majority of great national theatres, particularly those of China and Japan, the booths of the Italian commedia, the Elizabethan public and private playhouses, and the Spanish *corrales* (*i.e.*, the areas between town houses) of the Renaissance. While open-stage performance discouraged scenic elaboration, it stressed the actor and his role, his playing to and away from the spectators, with the consequent subtleties of empathy and alienation. It permitted high style in speech and behaviour, yet it could also accommodate moments of the colloquial and the realistic. It encouraged a drama of range and versatility, with rapid changes of mood and great flexibility of tone. It is not surprising that in the 20th century the West has seen a return to the open stage and that recent plays of Brechtian theatre and the theatre of the absurd seem composed for open staging.

The proscenium stage. The third basic theatre form is that of the proscenium-arch or picture-frame stage, which reached its highest achievements in the late 19th century. Not until public theatres were roofed, the actors withdrawn into the scene, and the stage artificially illuminated were conditions ripe in Western theatre for a new development of spectacle and illusion. This development had a revolutionary effect upon the literary drama. In the 18th and 19th centuries, plays were shaped into a new structure of acts and scenes, with intermissions to permit scene changes. Only recently has the development of lighting techniques encouraged a return to a more flexible episodic drama. Of more importance, the actor increasingly withdrew into the created illusion of the play, and his character became part of it. In the mid-19th century, when it was possible to dim the house lights, the illusion could be made virtually complete. At its best, stage illusion could produce the delicate naturalism of a Chekhovian family scene, into which the spectator was drawn by understanding, sympathy, and recognition; at its worst, the magic of spectacle and the necessary projection of the speech and acting in the largest picture-frame theatres produced a crude drama of sensation in which literary values had no place.

Audience expectations. It may be that the primary influence upon the conception and creation of a play is that of the audience. An audience allows a play to have only the emotion and meaning it chooses, or else it defends itself either by protest or by a closed mind. From the time the spectator began paying for his playgoing, during the Renaissance, the audience more and more entered into the choice of the drama's subjects and their treatment. This is not to say that the audience was given no consideration earlier; even in medieval plays there were popular non-biblical roles such as Noah's wife, or Mak the sheepthief among the three shepherds, and the antic devils of the Harrowing of Hell in the English mystery cycles. Nor, in later times, did a good playwright always give the audience only what it expected—Shakespeare's *King Lear* (*c.* 1605), for example, in the view of many the world's greatest play, had its popular elements of folktale, intrigue, disguise, madness, clowning, blood, and horror; but each was turned by the playwright to the advantage of his theme.

Any examination of the society an audience represents must illuminate not only the cultural role of its theatre but also the content, genre, and style of its plays. The exceptionally aristocratic composition of the English Restoration audience, for example, illuminates the social game its comedy represented, and the middle class composition of the subsequent Georgian audience sheds light on the moralistic elements of its "sentimental" comedy. Not unrelated is the study of received ideas in the theatre. The widespread knowledge of simple Freudian psychology has undoubtedly granted a contemporary playwright like Tennessee Williams (1911–) the license to invoke it for character motivation; and Brecht increasingly informed his comedies with Marxist thinking on the assumption that the audiences he wrote for would appreciate his dramatized argument. Things go wrong when the intellectual or religious background of the audience does not permit a shared experience, as when Jean-Paul Sartre (1905–) could not persuade a predominantly Christian audience with an existentialist explanation for the action of his plays, or when T.S. Eliot (1888–1965) failed to persuade an audience accustomed to the conventions of drawing-room comedy that *The Cocktail Party* (1949) was a possible setting for Christian martyrdom. Good drama persuades before it preaches, but it can only begin where the audience begins.

Special audiences. A great variety of drama has been written for special audiences. Plays have been written for children, largely in the 20th century, though Nativity plays have always been associated with children both as performers and as spectators. These plays tend to be fanciful in conception, broad in characterization, and moralistic in intention. Nevertheless, the most famous of children's plays, James Barrie's *Peter Pan* (1904), implied that the young are no fools and celebrated children in their own right. Barrie submerged his point subtly beneath the fantasy, and his play is still regularly performed, while Maurice Maeterlinck's *Blue Bird* (1908) has disappeared from the repertory because of its weighty moral tone.

Children's drama

In the wider field of adult drama, the social class of the audience often accounts for a play's form and style. Court or aristocratic drama is readily distinguished from that of the popular theatre. The veneration in which the Nō drama was held in Japan derived in large part from

the feudal ceremony of its presentation, and its courtly elements ensured its survival for an upper class and intellectual elite. Although much of it derived from the Nō, the flourishing of the Kabuki at the end of the 17th century is related to the rise of a new merchant and middle class audience, which encouraged the development of less esoteric drama. The popular plays of the Elizabethan public theatres, with their broader, more romantic subjects liberally spiced with comedy, are similarly to be contrasted with those of the private theatres. The boys' companies of the private theatres of Elizabethan London played for a better paying and more sophisticated audience, which favoured the satirical or philosophical plays of Thomas Middleton (1570?–1627), John Marston (1576–1634), and George Chapman (1559?–1634). Similarly today, in all Western dramatic media—stage, film, radio, and television—popular and "commercial" forms run alongside more "cultural" and avant-garde forms, so that the drama, which in its origins brought people together, now divides them. Whether the esoteric influences the popular theatre, or vice versa, is not clear, and research remains to be done on whether this dichotomy is good or bad for dramatic literature or the people it is written for.

THE RANGE OF DRAMATIC FORMS AND STYLES

Dramatic literature has a remarkable facility in bringing together elements from other performing and nonperforming arts: design and mime, dance and music, poetry and narrative. It may be that the dramatic impulse itself, the desire to recreate a picture of life for others through impersonation, is at the root of all the arts. Certainly, the performing arts continually have need of dramatic literature to support them. A common way of describing an

Opera

opera, for example, is to say that it is a play set to music. In Wagner the music is continuous; in Verdi the music is broken into songs; in Mozart the songs are separated by recitative, a mixture of speech and song; while operettas and musical comedy consist of speech that breaks into song from time to time. All forms of opera, however, essentially dramatize a plot, even if the plot must be simplified on the operatic stage. This is because, in opera, musical conventions dominate the dramatic conventions, and the spectator who finds that the music spoils the play, or who finds that the play spoils the music, is one who has not accepted the special conventions of opera. Music is drama's natural sister; proof may be seen in the early religious music-drama of the Dionysiac festivals of Greece and the *mystères* of 14th-century France, as well as in the remarkable development of opera in 17th-century Italy spreading to the rest of the world. The librettist who writes the text of an opera, however, must usually subserve the composer, unless he is able to embellish his play with popular lyrics, as John Gay did in *The Beggar's Opera* (1728), or to work in exceptionally close collaboration with the composer, as Brecht did with Kurt Weill for his *Die Dreigroschenoper* (*The Threepenny Opera*, 1928).

Dance, with its modern, sophisticated forms of ballet, has also been traditionally associated with dramatic representation and has similarly changed its purpose from religious to secular. In ballet, the music is usually central, and the performance is conceived visually and aurally; hence, the writer does not play a dominant role. The scenario is prepared for dance and mime by the choreographer. The contemporary Irish writer Samuel Beckett, trying to reduce his dramatic statement to the barest essentials, "composed" two mimes entitled *Act Without Words I* and *II* (1957 and 1966), but this is exceptional.

In motion pictures, the script writer has a more important but still not dominant role. He usually provides a loose outline of dialogue, business, and camera work on which the director, his cameramen, and the cutting editor build the finished product. The director is usually the final artistic authority and the central creative mind in the process, and words are usually subordinate to the dynamic visual imagery. (This subject is developed at length in the article MOTION PICTURES, ART OF.)

The media of radio and television both depend upon words in their drama to an extent that is not characteristic of the motion picture. Though these mass media have been dominated by commercial interests and other economic factors, they also have developed dramatic forms from the special nature of their medium. The writer of a radio play must acknowledge that the listener cannot see the actors but hears them in conditions of great intimacy. A radio script that stresses the suggestive, imaginative, or poetic quality of words and permits a more than conventional freedom with time and place can produce a truly poetic drama, perhaps making unobtrusive use of earlier devices like the chorus, the narrator, and the soliloquy: the outstanding example of radio drama is *Under Milk Wood* (1953), by the Welsh poet Dylan Thomas.

A similar kind of dramatic writing is the so-called readers' theatre, in which actors read or recite without decor before an audience. (This is not to be confused with "closet drama," often a dramatic poem that assumes dialogue form; *e.g.*, Milton's *Samson Agonistes*, 1671, written without the intention of stage performance.) The essential discipline of the circuit of communication with an audience is what distinguishes drama as a genre, however many forms it has taken in its long history.

BIBLIOGRAPHY. ALLARDYCE NICOLL, *World Drama* (1949), offers the best survey of the whole field, but should be supplemented by JOHN GASSNER and EDWARD QUINN, *The Reader's Encyclopaedia of World Drama* (1969); and PHYLLIS HARTNOLL, *The Oxford Companion to the Theatre*, 3rd ed. (1967). The classical texts of dramatic theory and criticism may be found in a collection by B.H. CLARK, *European Theories of the Drama*, rev. ed. (1965), which also contains an extensive bibliography; and a useful commentary on critics writing after 1750 is supplied in RENE WELLEK, *A History of Modern Criticism*, 4 vol. (1955–65). The *Natya Shastra* of BHARATA, the classic source for Indian dramatic theory, was translated by M.M. GHOSE in 1951.

Books of importance in the development of modern theory on drama are: BERNARD BECKERMAN, *Dynamics of Drama* (1970); E.R. BENTLEY, *The Life of the Drama* (1964); KENNETH BURKE, *A Grammar of Motives* (1945); FRANCIS FERGUSSON, *The Idea of a Theatre* (1949); S.K. LANGER, *Feeling and Form* (1953); ELDER OLSON, *Tragedy and the Theory of Drama* (1961); RONALD PEACOCK, *The Art of Drama* (1957); and J.L. STYAN, *The Elements of Drama* (1960).

The finest study of the classical drama of Greece is probably H.D.F. KITTO, *Greek Tragedy*, 3rd ed. (1961); and for the medieval drama are recommended: KARL YOUNG, *The Drama of the Medieval Church*, 2 vol. (1933); HARDIN CRAIG, *English Religious Drama of the Middle Ages* (1955); and O. B. HARDISON, *Christian Rite and Christian Drama in the Middle Ages* (1965). Oriental theatre is surveyed in FAUBION BOWERS, *Japanese Theatre* (1952); F.A. LOMBARD, *An Outline History of the Japanese Drama* (1928), which should be read in conjunction with ARTHUR WALEY'S classic *The Noh Plays of Japan* (1922); A.C. SCOTT, *The Classical Theatre of China* (1957); A.B. KEITH, *The Sanskrit Drama* (1924); with H.W. WELLS' comparative studies, *The Classical Drama of India* (1963), and *The Classical Drama of the Orient* (1965).

M.C. BRADBROOK, *Themes and Conventions of Elizabethan Tragedy* (1935); and U.M. ELLIS-FERMOR, *Jacobean Drama* (1936), are standard surveys of the English Renaissance drama; and for standard Shakespearian criticism the reader should consult A.M. EASTMAN, *A Short History of Shakespearean Criticism* (1968). The classic source books for the *commedia dell'arte* are P.L. DUCHARTRE, *La Comédie italienne* (Eng. trans., *The Italian Comedy*, 1929, reprinted 1966); and ALLARDYCE NICOLL, *Masks, Mimes and Miracles* (1931). On the French classical drama H. C. LANCASTER, *A History of French Dramatic Literature in the Seventeenth Century*, 9 vol. (1929–42), is standard; but MARTIN TURNELL, *The Classical Moment* (1947), deals more briefly with Corneille, Racine, and Molière. On Restoration comedy J.L. PALMER, *The Comedy of Manners* (1913); and BONAMY DOBREE, *Restoration Comedy* (1924), remain the best.

American drama is surveyed briefly in W.J. MESERVE, *An Outline History of American Drama* (1965); and A.S. DOWNER, *Fifty Years of American Drama, 1900–1950* (1951). U.M. ELLIS-FERMOR, *The Irish Dramatic Movement*, 2nd ed. (1954), is a comprehensive study of the early years at Dublin's Abbey Theatre; and on Western drama after Ibsen the reader should begin by consulting E.R. BENTLEY, *The Playwright as Thinker* (1946); ROBERT BRUSTEIN, *The Theatre of Revolt* (1964); and J.L. STYAN, *The Dark Comedy*, 2nd ed. (1968).

(J.L.S.)

Dravidian Languages

The Dravidian language family, as known to date, consists of 23 languages spoken by more than 110,000,000 people in South Asia. In terms of population figures the major languages of the family may be listed in the following order: Telugu, 37,600,000; Tamil, 30,560,000; Kannada (Kannaḍa), also called Kanarese, 17,420,000; Malayalam (Malayālam), 17,020,000; Gondi, 1,500,000; Kurukh (Kurukh), 1,140,000; and Tulu (Tuḷu), 940,000. The Dravidian languages are spoken in the Republic of India (mainly in its southern, eastern, and central parts), in Sri Lanka (Ceylon), and by settlers in areas of Southeastern Asia, southern and eastern Africa, and elsewhere. Brahui (Brāhuī), with 300,000 speakers in Pakistan, is isolated from all of the other members of the family. The four major languages—Telugu, Tamil, Kannada, and Malayalam—possess independent scripts and literary histories dating from the pre-Christian Era. Now recognized by the constitution of India, they form the basis of the linguistic states of Andhra Pradesh (established as the first Indian linguistic state in 1953), Tamil Nadu, Mysore, and Kerala.

Official use of Dravidian languages

Tamil is the language with the greatest geographical extension and the richest and most ancient literature, paralleled in India only by that of Sanskrit. Its phonological and grammatical systems correspond in many points to the ancestral parent language, called Proto-Dravidian.

Nothing definite is known about the origin of the Dravidian family. There are vague indigenous traditions about a migration from the south, from a submerged continent in the Indian Ocean. According to some scholars, Dravidian languages are indigenous to India. In recent years, a hypothesis has been gaining ground about a movement of Dravidian speakers from the northwest to the south and east of the Indian Peninsula, possibly from as far away as Central Asia. Another theory connects the Dravidian speakers with the peoples of the Indus Valley civilization. The Dravidian languages have remained an isolated family until now and have defied all of the attempts to show a connection with the Indo-European tongues, Mitanni, Basque, Sumerian, or Korean. The most promising and plausible hypothesis is that of a linguistic relationship with the Uralic (Hungarian, Finnish) and Altaic (Turkish, Mongol) language groups.

As an independent family, the Dravidian languages were first recognized in 1816 by Francis W. Ellis, a British civil servant. The actual term Dravidian was first employed by Robert A. Caldwell, who introduced the Sanskrit word *drāvida* (which, in a 7th-century text, obviously meant Tamil) into his epoch-making *A Comparative Grammar of the Dravidian or South Indian Family of Languages* (1st ed., 1856).

Languages of the family. Tamil is spoken by 30,562,698 people (1961) in the state of Tamil Nadu, by another 2,500,000 in Sri Lanka (Ceylon), and in Burma, Malaysia, Indonesia, and Vietnam (about 1,000,000), in East and South Africa (almost 250,000), and by small numbers in British Guiana and on the islands of Fiji, Mauritius, Réunion, Madagascar, Trinidad, and Martinique. The earliest literary monuments of the language belong roughly to the 3rd and 2nd centuries BC. There exist a number of local dialects, the major dialect regions being the northern and eastern areas combined, the western area, the southern area (split into at least four major dialects of Madurai, Tirunelveli, Nanjiland, and Ramnad), and Sri Lanka (Ceylon). Correlated with the social position of the speaker are a number of speech forms; a major division occurs between the Brahmin and the non-Brahmin varieties. In addition, there is a sharp dichotomy between the formal language and informal speech.

Dialects and forms of Tamil

Malayalam, closely related to Tamil, is spoken in the state of Kerala by 17,015,782 people. Possessing an independent script, it also has a rich modern literature. There are at least three main regional dialects (North, Central, South) and a number of communal dialects.

In the Nīlgiris and adjacent regions, several minor tribes speak the following languages: Kota (862), Toda (765), Badaga (85,463), Irula (Iruḷa) (4,124). The languages of a number of other tribes may yet be established as independent members of the family (*e.g.*, Kurumba, Paṇiya).

Kodagu (Koḍagu), a non-literary language of a mountainous region called Coorg, has 78,202 speakers.

Kannada (Kanarese), spoken by 17,415,827 people in the state of Mysore, has a dichotomy between educated speech and colloquial Kannada; in the latter at least three social dialects are recognizable—Brahmin, non-Brahmin, and Harijan ("untouchable"). A number of regional dialects (among them, Dharwar, Bangalore, and Mangalore) also exist. Kannada has an orthography of its own and an important ancient and modern literature.

To the south of the Kannada territory, more than 900,000 people speak Tulu (Tuḷu), a South Dravidian language having no developed written literature.

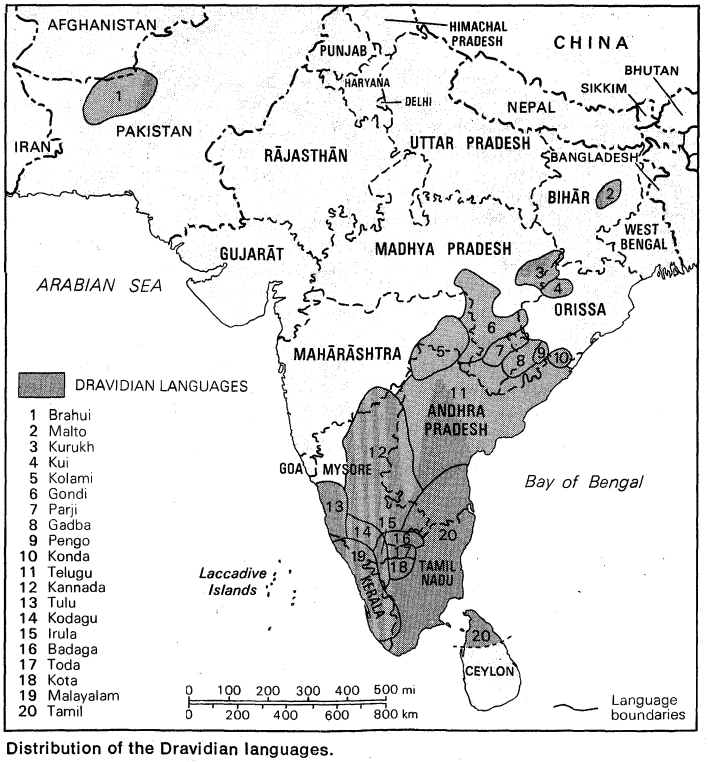
Telugu (37,668,132), the official language of Andhra Pradesh, exhibits a dichotomy between the written and the spoken styles, in addition to a number of sharply distinct local and regional dialects (Telangana, coastal area, Rayalaseema, and a "transitional" zone) and divisions between Brahmin, non-Brahmin, and Harijan speech. The language has its own script, closely akin to that of Kannada, and an important literary tradition.

In the northern parts of Andhra Pradesh, two tribes speak Kolami (46,065 persons) with its dialect Naikri (Naikṛi), and Naiki (1,000), whereas Parji (94,607) is spoken in Bastar, Madhya Pradesh. In Orissa, Konda (Koṇḍa) is spoken by about 12,000 Konda Doras, and about 2,000 Gadbas speak the three closely related dialects of Ollari, Pottangi, and Poya; Pengo (1,254) and Manda (Maṇḍa) were discovered only recently, and Naiki of Chānda is also a newly investigated language (since 1961). The Khond tribes of Orissa (600,000) speak two closely related languages, Kui and Kuvi.

In the vast regions of Madhya Pradesh, many groups of Gonds (including about 1,500,000 persons) speak a number of Gondi dialects. Still further to the north, in Bihār, Orissa, and Madhya Pradesh, the Oraon tribe speaks Kurukh (1,132,931), and near the borders of Bihār and West Bengal, 88,645 tribals speak Malto.

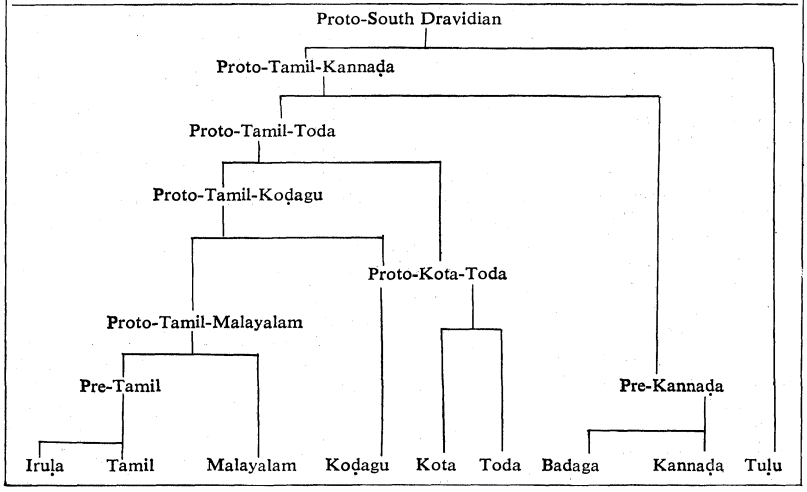
The only Dravidian language spoken entirely outside India is Brahui, with about 300,000 speakers in the Qualat, Hairpur, and Hyderābād districts of Pakistan.

Adapted from Ramanujan and Masica, "Toward a Phonological Typology of the Indian Linguistic Area," *Current Trends in Linguistics*, vol. 5 (1969); Mouton & Co., Publishers, The Hague



Distribution of the Dravidian languages.

Table 1: South Dravidian Subfamily



Historical survey of the Dravidian languages. Although in modern times Dravidian has mainly occupied the southern portion of India, while the Indo-Aryan (Indic) tongues have predominated in north India, nothing definite is known about the ancient domain of the Dravidian parent speech. It is, however, a well-established hypothesis that Dravidian speakers must have been widespread throughout India, including the northwest region.

Dravidian features in the Ṛgveda

This is clear because a number of features of the Dravidian languages appear in the Ṛgveda, the earliest known Indo-Aryan literary work, thus showing that the Dravidian languages must have been present in the area of the Indo-Aryan ones. The Indo-Aryan languages were not, however, originally native to India; they were introduced by Aryan invaders from the north. Several scholars have demonstrated that pre-Indo-Aryan and pre-Dravidian bilingualism in India provided conditions for the far-reaching influence of Dravidian on the Indo-Aryan tongues in the spheres of phonology (*e.g.*, the retroflex consonants, made with the tongue curled upward toward the palate), syntax (*e.g.*, the frequent use of gerunds, which are nonfinite verb forms of nominal character, as in "by the falling of the rain"), and vocabulary (a number of Dravidian loanwords apparently appearing in the Ṛgveda itself).

Thus a form of Proto-Dravidian, or perhaps Proto-North Dravidian, must have been extensive in north India before the advent of the Aryans. Apart from some islands of Dravidian speech, however, the process of replacement of the Dravidian languages by the Aryan tongues was entirely completed before the beginning of the Christian Era, after a period of bilingualism that must have lasted many centuries. Finally, the almost universal adoption of Indo-Aryan in the north and Dravidian in the south has covered up the original linguistic diversity of India.

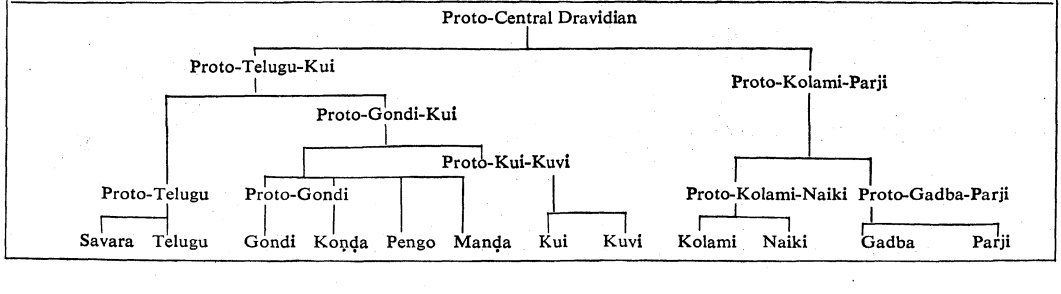
The advent of Dravidian speakers in India is shrouded in mystery. There are vague linguistic and cultural ties with the Urals, with the Mediterranean area, and with Iran. It is possible that a Dravidian-speaking people that can be described as dolichocephalic (longheaded from front to back) Mediterraneans mixed with brachycephalic (short-headed from front to back) Armenoids and established themselves in northwest India during the 4th millennium BC. Along their route, these immigrants may have possibly come into an intimate, prolonged contact with the Ural-Altaic speakers, thus explaining the striking affinities between the Dravidian and Ural-Altaic language groups. Between 2000 and 1500 BC, Dravidian speakers moved constantly from the northwest to the southeast of India, and in about 1500 BC three distinct dialect groups probably existed: Proto-North Dravidian, Proto-Central Dravidian, and Proto-South Dravidian. The beginnings of the splits in the parent speech are obviously earlier. It is possible that Proto-Brahui was the first language to split off from Proto-Dravidian, probably during the immigration movement into India sometime in the 4th millennium BC, and that the next subgroup to split off was Proto-Kurukh-Malto, sometime in the 3rd millennium BC (see the family tree diagrams, Tables 1–3).

Possible Dravidian and Ural-Altaic contact

Compared to the work done on other language families, the progress in comparative Dravidian studies is still meagre. Considerable knowledge has been acquired in comparative phonology (sound systems), but correspondences have been worked out only for the sounds in the roots of words. Very little comparative work has been done on grammatical processes and complete historical grammars of the literary languages are still lacking. Hence the reconstruction of any feature of the Dravidian protolanguage, with the possible exception of some parts of the phonology, must necessarily be very tentative.

The vowel system of Proto-Dravidian consisted of five vowels—**i*, **u*, **e*, **o*, **a* (an asterisk denotes an unattested, reconstructed, hypothetical form)—each having two quantities, short and long. Relative stability of root vowels seems to have been the rule. The Proto-Dravidian consonant system consisted of obstruants (stops) **p*, **t*, **ṯ*, **ṭ*, **c*, **k*; nasals **m*, **n*, **ṉ*, **ñ*; laterals **l*, **ḷ*; the flap **r*; the voiced retroflex continuant **ṛ*; and the semivowels **y* and **v*. The most characteristic feature of the consonantal system was the six positions of articulation for obstruants: labial (with the lips), dental (tongue

Table 2: Central Dravidian Subfamily



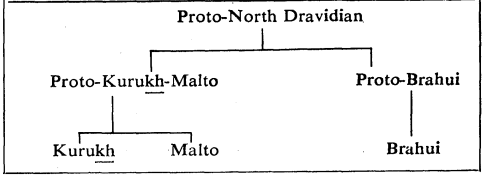
touching the back of the upper teeth), alveolar (tongue touching the upper gum ridge), retroflex (tip of tongue curled upward toward the palate and back), palatal (body of tongue touching the palate, or roof of the mouth), and velar (back of tongue touching the velum, or soft palate). The retroflex series was very distinctive and important and comprised an obstruant *ṭ, a nasal *ṇ, a lateral *ḷ, and a continuant *ṛ. No consonant of the alveolar or retroflex series began a word. In the final position all of the consonants occurred, but all of the obstruants were followed by an automatic release sound, the vowel *-u. Initial consonant clusters did not occur. There was only one series of obstruant phonemes (distinctive sounds); these sounds were voiceless (produced without vibration of the vocal cords) initially and voiced (with vocal cord vibration) between vowels. All Proto-Dravidian roots were monosyllables.

Proto-Dravidian used only suffixes, never prefixes or infixes. Hence, the roots of words always occurred at the beginning. Nouns, verbs, and indeclinable words constituted the original word classes.

Process of Dravidian acculturation in India

During the 1st millennium BC, while Aryanization steadily progressed in north India, the Dravidian-speaking newcomers began to mix with the Negritos and Proto-Australoids in the south; this process of acculturation continued during the period from approximately 1200 to 600 BC. A movement of the Aryans into the south of India began sometime about 1000 BC. Before the 5th century BC, Proto-South Dravidian was probably still one language, with two strongly marked dialects. Within Proto-Central Dravidian, a deep two-way division also occurred, and North Dravidian must have already been split into the Kurukh–Malto and Brahui subgroups (see the family tree diagrams, Tables 1–3).

Table 3: North Dravidian Subfamily



Apart from a possible Dravidian word in the Hebrew text of the Bible (*tukkhiyīm* "peacocks"; *cf.* Tamil *tōkai* "tail of a peacock"), the Dravidian languages enter history in Sanskrit and Greco-Roman texts. The *Cēras*, a south Indian dynasty, are possibly mentioned in the early Sanskrit text *Aitareya Āraṇyaka*. Kātyāyana, a grammarian of the 4th century BC, mentions the countries of Pāṇḍya (Tamil *pāṇṭiya*), Cōla (Tamil *cōla*), and Kerala, or Cēra (Tamil *cēra*); these lands were well known to Kauṭilya (4th century BC), the author of the earliest treatise on statecraft, and appear in the edicts of the great Buddhist leader Aśoka (3rd century BC). The term *drāviḍa* is almost certainly a Sanskritization (with an inserted "hypercorrect" *r*) of the earlier Pāli and Prākrit terms *dāmiḷo, damiḷa, dāviḍa*, which must have been derived from the Tamil name of the language, *tamil*. A number of South Dravidian words, almost all of the geographic and dynastic names, occur in such Greco-Roman sources as the *Periplus maris Erythraei* ("Circumnavigation of the Erythraean Sea") of about AD 89 and in the writing of Ptolemaeus of Naukratis of the 2nd century AD; it is also very probable that Western-language terms for rice (compare Italian *riso*, Latin *oryza*, Greek *oryza*) and ginger (compare Italian *zenzero*, German *Ingwer*, Greek *zingiberis*) are cultural loans from Old Tamil *arici* and *iñcivēr*, respectively.

Sometime during the reign of Aśoka (3rd century BC), the two South Dravidian languages, Tamil and Kannada, developed into distinct idioms and the two cultures emerged as separate entities; a third major Dravidian linguistic and cultural unit, Telugu, appeared in the Andhra country. In the period from 300 to 100 BC, one of the pre-Tamil dialects (probably that of Madurai) gained prestige and became the standard literary language (*centamil*), the written form of early Old Tamil, established in poetic texts and in its earliest grammar, *Tolkāppiyam*. At the same time, about 250 BC, the Aśokan Southern Brāhmī script was adapted for Tamil and used in short cave inscriptions by Jain monks during several centuries, dating approximately from the 2nd century BC to the 5th century AD.

Early inscriptions and writings

The earliest inscriptions in Kannada may be dated at AD 450; Kannada literature begins with Nṛpatuṅga's *Kavirājamārga*, about AD 850. The oldest Telugu inscription is from AD 633, and the literature begins with the grammarian Nannaya's 11th-century translation of the Sanskrit classic the *Mahābhārata*. In Malayalam, the earliest writings are from the close of the 9th century, and the first literary text is probably the *Bhāṣākauṭalīyam*, AD 1125–1250.

Since these attested beginnings, the four languages—Tamil, Malayalam, Kannada, and Telugu—have been used continuously in administration and literature up to the present day. In addition to possessing an immense wealth of epigraphic and literary texts, they all developed pronounced features of diglossia, a dichotomy between the standardized, formal language and the informal, colloquial speech, which is divided into regional as well as social dialects. In modern times, all of the four cultivated languages have adapted quickly to new conditions resulting from economic, social, and political changes. All of them are used for basic courses in science and the arts; and new technological terminology is coined, based either on English or Sanskrit models, but often on exclusively indigenous linguistic material (in Tamil).

To date, nothing is known about the history of the nonliterary Dravidian languages before their "discovery," which began at the end of the 18th century. The Gonds, however, are mentioned (as Gondaloi) by Ptolemy of Naukratis, writing in the 2nd century AD.

A tendency toward structural and systemic balance and stability is characteristic of the Dravidian group. Nevertheless, there is no doubt about the influence of the other languages of India. Dravidian languages show extensive lexical (vocabulary) borrowing, but only a few traits of structural (either phonological or grammatical) borrowing, from the Indo-Aryan tongues. On the other hand, Indo-Aryan shows rather large-scale structural borrowing from Dravidian, but relatively few loanwords. There is indeed a possibility of Dravidian and Indo-Aryan drawing even closer together in the future; but it is highly doubtful that a new family of languages will develop with the bases of the contributing groups (*i.e.*, Dravidian and Indo-Aryan) completely eliminated through the phenomena of borrowing.

Characteristics of the Dravidian languages. Dravidian languages would probably be called agglutinative in the categorization of the 19th-century philologists. An agglutinative language incorporates separate formal units of distinct meaning into a single word. There are some elements of "internal flexion" (*e.g.*, the alternation of short–long root vowels in derived words), however, as well as regular alternations in vowel and consonant quantities within the root. Relatively low receptivity to change results in a slower rate of change than is found in the Indo-European language family. The degree of phonetic divergence among the Dravidian languages is not very great; hence, etymologies are not too difficult to discover. The territory occupied by Dravidian speakers in India may be characterized as a large dialect area resembling the area of the Romance languages, with numerous boundaries marked by bundles of isoglosses (an isogloss is a boundary line that separates the areas of two differing features of language usage), but also with many isoglosses enclosing more than one language. In any study of Dravidian, therefore, both evolution and diffusion must be taken into account.

Agglutinating character of Dravidian

Sounds of Dravidian. Compared to the reconstructed system of Proto-Dravidian phonemes (distinctive sounds), the most striking developments in vowels are the gradual elimination of the contrast between *e* and *ē* (long *e*) and *o* and *ō* (long *o*) in Brahui, as a result of the influence of Indo-

Aryan languages or Iranian or both; the raising of Proto-Dravidian *e and *o to i and u and the lowering of these protolanguage sounds in Brahui; and the merger of Proto-Dravidian *i and *u with *e and *o in the South Dravidian languages before a consonant plus the vowel a. Also noteworthy are the emergence of retroflex vowels (i.e., centralized vowels "coloured" by neighbouring retroflex consonants) in Kodagu and Irula; the nasalization of vowels, as in colloquial Tamil; the loss of vowels in unaccented noninitial syllables in Toda, Kota, some dialects of Kannada, and Tamil, and the resulting consonant clusters (e.g., Kota *anjrčgčgvdk*, "because of the fact that [someone] will cause [someone] to terrify [someone]"). Metathesis (the transposition of sounds, as in "aks" from "ask") and vowel contraction resulted in initial consonant clusters in Telugu and other Central Dravidian languages—e.g., Tamil *koḷu*, but Kui *krōga*, both meaning "fat."

Consonant changes

Among the most important consonantal developments are the loss of *c-, a typical South Dravidian phenomenon that seems to be still in progress (e.g., Proto-Dravidian *caṛ-, but Tamil *aḷal* "to burn," and *talal* "to glow"); the velarization of *c- to k- in North Dravidian when the sound is followed by ŭ (e.g., Tamil *cuṭu* "be hot," but Malto *kut-* "burn"); the palatalization of Proto-Dravidian *k- to c- before front vowels in Tamil, Malayalam, and Telugu (e.g., *ke- "red," but Tamil *ce-*); and the replacement of *k- in North Dravidian by x before ă, ŏ, and ŭ (e.g., Tamil *kal*, but Brahui *xal*, "stone"). The retroflex voiced continuant *ṛ has been preserved only in the old stages of the cultivated languages and partly in modern Tamil and Malayalam; elsewhere, it merged with ḷ, ḍ, and other sounds. Some languages, notably Kannada, developed a secondary h-, not inherited from the parent speech (e.g., Tamil *peyar*, Old Kannada *pesar*, but Modern Kannada *hesru*, "name"). According to the Dravidian scholar Bhadriraju Krishnamurti, a laryngeal (or h- type of sound) should be reconstructed for some items in Proto-Dravidian.

Problems of accent and intonation still remain to be worked out. Word stress is predictable, always occurring on the radical (initial) syllable and therefore being nondistinctive. The rules of sandhi (change of a sound or sounds as a result of adjacent sounds) are as complicated and delicate as in Sanskrit.

Grammatical features of Dravidian. In grammar, the absolutely prevailing process is suffixation, the addition of suffixes. Grammatical functions are, however, also expressed by composition (the compounding of word elements) and by word order. There are no prefixes or infixes. Suffixes agglutinate (are attached to one another); e.g., Tamil *coṉṉatilēyiruntu* "from what was said" is composed of *col* "say" + *n* "past" + *atu* "3rd person singular neuter" + *il* "locative" + *ē* "emphatic" + *y* (an automatic insertion resulting from a sound rule) + *iruntu* "ablative" (*iruntu* comes from *iru* "be" + *nt/u* "past").

The major word classes are nouns (substantives, numerals, pronouns), adjectives, verbs, and indeclinables (particles, enclitics, adverbs, interjections, onomatopoetic words, echo words). There are two numbers and four different gender systems, the "original" probably having "male:non-male" in the singular and "person:non-person" in the plural. The pronoun has a category "inclusive:exclusive" in the 1st person plural. A characteristic derivation is that of "pronominalized" or "personal" nouns and adjectives; e.g., Old Tamil *iḷai* "youth," *iḷai-y-am* "young-we," *iḷai-y-ar* "young-they." Finite forms of the verb (forms showing person and number) are, ultimately, "pronominalized" verb stems; e.g., Tamil *aṭi-(y)-ēn* ("slave"—1st person singular) "I am a slave"; *nal-(l)-ēn* ("good"—1st person singular) "I am good"; *pō-v-ēn* ("go"—future—1st person singular) "I shall go." The most characteristic feature of the Dravidian verb is a full-fledged negative system: all of the positive verb forms have their corresponding negative counterparts. Verbs are intransitive, transitive, and causative; there are also active and passive forms. The main (and probably original) dichotomy in tense is past:non-past. Present tense developed later and independently in each language or subgroup.

Negative verbs

In a sentence, however complex, only one finite verb occurs, normally at the end, preceded if necessary by a number of gerunds. Gerunds and participles, as well as verb-nouns, play an important role. The determining member always precedes the determined; e.g., Tamil *pon* "gold" + *nakaram* "city" becomes *poṉṉakaram* "city of gold, golden city." Word order follows certain basic rules but is relatively free.

Vocabulary. In vocabulary, different languages were receptive to loanwords in differing degrees. Among the cultivated languages, Tamil has the relatively lowest number of Indo-Aryan loanwords (18–25 percent, according to the style), whereas in Malayalam and Telugu the percentage is much higher. The most important sources of loanwords have been Sanskrit, Pāli, and Prākrit (with varying impact in different periods); in modern times Urdu, Portuguese, and English have made significant contributions. There was not much lexical borrowing from one Dravidian language into another in historical times. Among all the Dravidian languages, Brahui, in Pakistan, is inevitably the one most influenced by Indo-Aryan and Iranian; in contrast, Toda is probably the one language least influenced by any other idiom. In Tamil, there is currently a very notable and active purifying movement; it aims at removing as many "Sanskritic" (but not English) vocabulary items as possible. Such purism has not yet occurred in any other of the cultivated languages.

Writing. Writing was first developed in Tamil Nadu, sometime about 250 BC, when the Aśokan Southern Brāhmī script was adapted for Tamil. The earliest inscriptions in Tamil script proper are the Pallava copperplates of about AD 550. The Kannada–Telugu script is based on Cālukya (6th century) inscriptions; the Grantha script, used in Tamil Nadu for Sanskrit since the 6th century, was accommodated for Malayalam and Tulu. Apart from these, Tamil has an old cursive script called Vaṭṭeḻuttu, "round script," and Malayalam possesses its own modern cursive form, Koleḻuttu, "rod-script."

BIBLIOGRAPHY. R. CALDWELL, *A Comparative Grammar of the Dravidian or South-Indian Family of Languages*, 3rd ed. by J.L. WYATT and T.R. PILLAI (1913, reprinted 1956 and 1961), the classic work that laid the foundations of Dravidian linguistics; G.A. GRIERSON (ed.), *Linguistic Survey of India*, vol. 4, *Muṇḍā and Dravidian Languages*, by S. KONOW (1906); T. BURROW and M.B. EMENEAU, *A Dravidian Etymological Dictionary* (1961, reprinted 1966; *Supplement*, 1968), the first etymological dictionary of the family, marking a new era in Dravidian studies (indispensable point of departure for any further work in the field); B. KRISHNAMURTI, *Telugu Verbal Bases: A Comparative and Descriptive Study* (1961), an indispensable study of the phonology and derivational morphology of Dravidian, with a much wider coverage of problems than the title suggests; "Comparative Dravidian Studies," *Linguistics in South Asia*, pp. 309–333, vol. 5 of *Current Trends in Linguistics* (1969), a summary treatment of the latest developments in the field; K. ZVELEBIL, *Comparative Dravidian Phonology* (1970), the first systematic compendium of the comparative phonology of Dravidian; J. BLOCH, *Structure grammaticale des langues dravidiennes* (1946; Eng. trans., *The Grammatical Structure of Dravidian Languages*, 1954), an excellent description of the main morphological and syntactic features of the family that ignores phonology totally; F.B.J. KUIPER, "The Genesis of a Linguistic Area," *Indo-Iranian Journal*, 10:81–102 (1967), a brief and brilliant treatment of the problems of Aryan and Dravidian convergence; M.S. ANDRONOV, *Materials for a Bibliography of Dravidian Linguistics* (1966); M. ISRAEL, "Additional Materials for a Bibliography of Dravidian Languages," *Tamil Culture*, 12:69–74 (1966); S.E. MONTGOMERY, "Supplemental Materials for a Bibliography of Dravidian Linguistics," *Studies in Indian Linguistics*, pp. 234–246 (1968), three bibliographies that provide fairly complete coverage.

(K.V.Z.)

Drawing

Drawing as formal artistic creation might be defined as the primarily linear rendition of objects in the visible world, as well as of concepts, thoughts, attitudes, emo-

tions, and fantasies given visual form, of symbols and even of abstract forms. This definition, however, applies to all graphic arts and techniques that are characterized by an emphasis on form or shape rather than mass and colour, as in painting. Drawing as such differs from graphic printing processes in that a direct relationship exists between production and result. Drawing, in short, is the end product of a successive effort applied directly to the carrier, which is usually paper. Whereas a drawing may form the basis for reproduction or copying, it is nonetheless unique by its very nature.

Function of drawing in the visual arts

Although not every art work has been preceded by a drawing in the form of a preliminary sketch, drawing is in effect the basis of all visual arts. Often the drawing is absorbed by the completed work or destroyed in the course of completion. Thus, the usefulness of a ground-plan drawing of a building that is to be erected decreases as the building goes up. Similarly, points and lines marked on a raw stone block represent auxiliary drawings for the sculpture that will be hewn out of the material. Essentially, every painting is built up of lines and pre-sketched in its main contours; only as the work proceeds is it consolidated into coloured surfaces. As shown by an increasing number of findings and investigations, drawings form the material basis of mural, panel, and book paintings. Such preliminary sketches may merely indicate the main contours or may predetermine the final execution down to exact details. They may also be mere probing sketches. Long before the appearance of actual small-scale drawing, this procedure was much used for monumental murals. With sinopia—the preliminary sketch found on a layer of its own on the wall underneath the fresco, or painting on freshly spread, moist plaster—one reaches the point at which a work that merely served as technical preparation becomes a formal drawing expressing an artistic intention.

Not until the late 14th century, however, did drawing come into its own—no longer necessarily subordinate, conceptually or materially, to another art form. Autonomous, or independent, drawings, as the name implies, are themselves the ultimate aim of an artistic effort; therefore, they are usually characterized by a pictorial structure and by precise execution down to details.

Formally, drawing offers the widest possible scope for the expression of artistic intentions. Bodies, space, depth, substantiality, and even motion can be made visible through drawing. Furthermore, because of the immediacy of its statement, drawing expresses the draftsman's personality spontaneously in the flow of the line; it is, in fact, the most personal of all artistic statements. It is thus plausible that the esteem in which drawing was held should have developed parallel to the value placed on individual artistic talent. Ever since the Renaissance, drawing has gradually been losing its anonymous and utilitarian status in the eyes of artists and the public, and its documents have been increasingly valued and collected as evidence of human creativity.

This article will deal with the aesthetic characteristics, the mediums of expression, the subject matter, and the history of drawing.

This article is divided into the following sections:

I. General considerations
  Elements and principles of design
    Line and linear techniques
    Plane techniques
    The drawing surface
    Relationship between drawing and other art forms
  Surfaces, mediums, and techniques
    Types of ground
    Tools and techniques
  Applied drawings
  Subject matter of drawing
    Portraits
    Landscapes
    Figure compositions and still lifes
    Fanciful and nonrepresentational drawings
    Artistic architectural drawings
II. History of drawing
  Western
    14th, 15th, and 16th centuries
    17th, 18th, and 19th centuries
    Modern
  Eastern

I. General considerations

ELEMENTS AND PRINCIPLES OF DESIGN

Line and linear techniques. The principal element of drawing is the line. Through practically the entire development of Western drawing, this figure, essentially abstract, not present in nature, and appearing only as a border setting of bodies, colours, or planes, has been the vehicle of a representational more or less illusionist rendition of objects. Only in very recent times has the line been conceived of as an autonomous element of form, independent of an object to be represented.

Conscious and purposeful drawing represents a considerable mental achievement, for the ability to reduce the spatial objects in the world around one to lines drawn on a plane presupposes a great gift for abstraction. The identification of the motif of a drawing by the viewer is no less an achievement, although it is mastered by practically all human beings. The visual interpretation of a line as a representation of a given object is made possible through certain forms of that line that call forth associations. The angular meeting of two lines, for example, may be considered as representing the borders of a plane; the addition of a third line can suggest the idea of a cubic body. Vaulting lines stand for arches, convergent lines for depth.

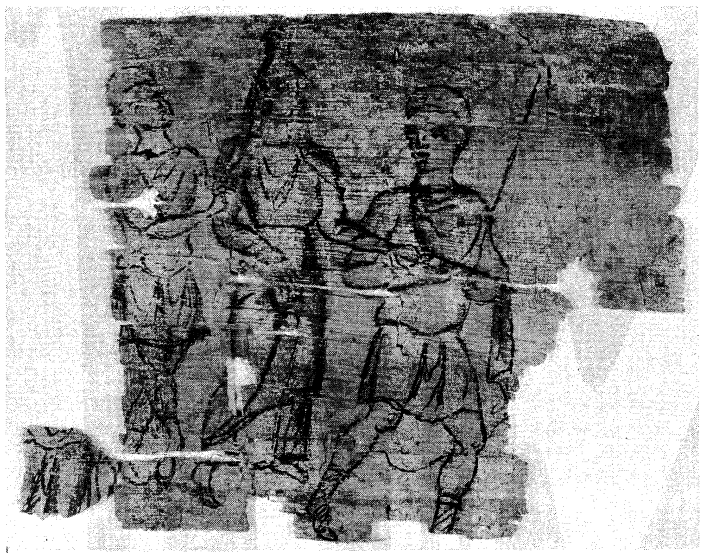
With the aid of this modest basic vocabulary, one can distill comprehensible images from a variety of linear phenomena. The simple outline sketch—Greek legend has it that the first "picture" originated from copying the shadows on the sand—represents one of the oldest and most popular possibilities of graphic rendition. After decisively characterizing the form of Egyptian drawing and the archaic art of Greece, the outline sketch became the chief vehicle of artistic communication in late antiquity and the Middle Ages. Used in a variety of ways in the early Renaissance, it became dominant once again in Neoclassicism, as it is, for that matter, in the classicist period of a given artist's total work.

The outline sketch

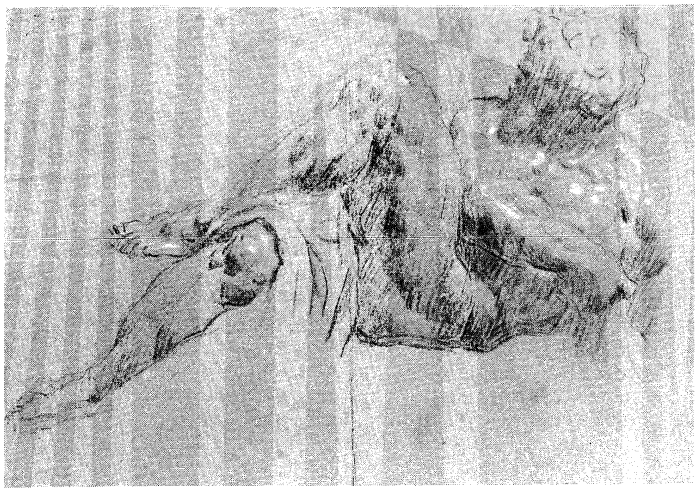
The outline sketch is elaborated into the detailed drawing by means of the line, which differentiates between the plastic and the spatial values of the object. Borders of individual objects, changes in the spatial plane, and varying intensities of colour applied within an outline sketch all tend to enrich and clarify the relationship between the whole and its component parts.

The free beginning, the disappearance, or the interruption of a line provides opportunities for gradually slurring an edge until it becomes a plane, for letting colour

By courtesy of the Bayerische Staatsbibliothek, Munich



Late antique outline sketch, "The Abduction of Briseis," pen drawing on papyrus, Greek, 4th century AD. In the Bayerische Staatsbibliothek, Munich. 13.2 × 14.4 cm.



Loose use of line emphasizing pictorial effect, study after Michelangelo's "Day," by Jacopo Tintoretto (*c.* 1518–94), black and white chalk on blue paper. In the Metropolitan Museum of Art, New York. 34.9 × 50.5 cm.

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1954

transitions fade away, for having the line vanish in the depth.

The thickening or thinning of a line can also be used to indicate, spatially or by means of colour, a change in the object designated by that line. Even light-and-shadow values may be rendered by differences in stroke strength.

While the chopping up of a line into several brief segments, and, even more, the drawing of individual lines running parallel in one direction, makes the outlined form appear less corporeal and firm, it reproduces the visual impact of the form in a more pictorial manner. Slight shifts in the flow of the line are intended to represent smooth curves and transitions; they also reinforce the effect of light striking a surface and thus give the corporeal appearance. Finally, short, curving segments of a line that do not stand in a clearly angular relationship to one another but are arranged on the sheet in loose formation allow the pictorial and colour component to dominate, as in the work of the 16th-century Italian artist Jacopo Tintoretto. An extreme case is the complete dissolution of the linear stroke into dots and spots, as, for example, in the drawings of the 19th-century pointillist painter Georges Seurat.

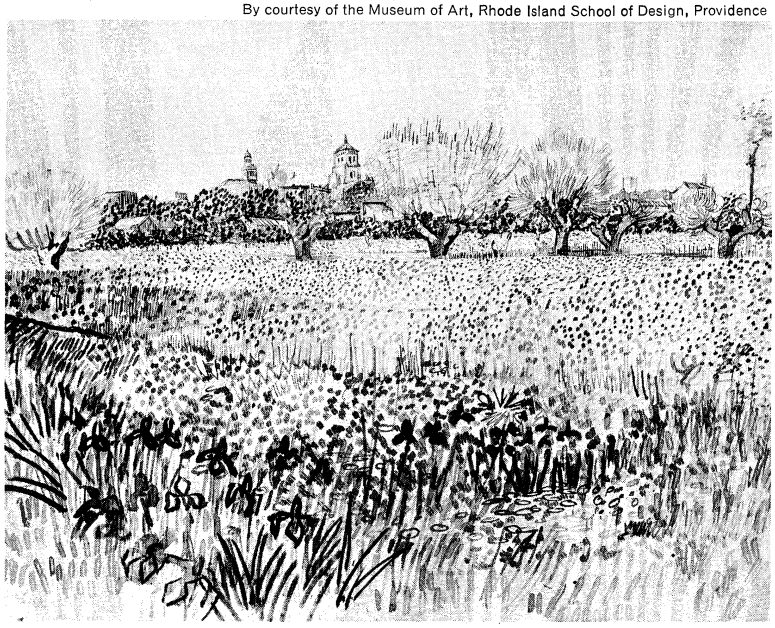
A mere combination of these varied shapes of the line, without reference to the mediums in which the lines are drawn, provides the artist with a plethora of subjective opportunities for the expression both of general stylistic traits and of personal characteristics. An arrangement of forceful, mainly straight strokes in accentuated, sharp angles lends the drawing an austere character emphasizing dramatic and expressive traits. This method of drawing, in fact, is characteristic of stylistic epochs and artistic regions (not to mention individual artists) that prefer these qualities: in the rather sober city of Florence, in German Expressionism, where it is used to convey mood, but also in the drawings of Rembrandt and Vincent van Gogh. Soft lines, on the other hand, running in drawn-out, smoothly rounded forms and stressing graphic regularity above any statement of content, constitute the formal equivalent to elegant, courtly, and lyric qualities of expression. Accordingly, they are often found in drawings of the Soft style; in the early Renaissance, particularly in the work of artists from the Italian province of Umbria and in young Raphael's sketches; in the work of Nazarenes, a 19th-century group of Romantic painters whose subjects were mainly religious; in the Jugendstil, a late-19th- and early-20th-century German decorative style parallel to Art Nouveau in its organic foliate forms, sinuous lines, and non-geometric curves; and in a very pure form in one of the classic draftsmen, the 19th-century French painter Jean-Auguste-Dominique Ingres. A markedly even-stroke texture, with waxing and waning strokes in regular proportions and evenly distributed within the page, brings drawing close to calligraphic writing and is found in all stylistic epochs that value ornamentation.

Hatching and cross-hatching

The technique of hatching gives the line an additional potential for the clarification of plastic relationships and of light phenomena. In hatching, parallel, short, equidistant, more or less straight lines create static and tectonic (structural) values by marking individual body planes. Gently curved hatching stresses the roundness of the body and can also accentuate, as tone value, shaded parts of the representation.

Cross-hatching, in which two layers of hatching intersect at right angles, reinforces the body-and-shadow effect. Known since the days of Michelangelo and Dürer in the 15th and 16th centuries, this artistic technique is often used with slanted or even curved hachures for the linear rendition of rounded parts. In rigorously monotone drawings, this method is the most suitable for the depiction of spherical bodies. The human body, with its highly articulated surface, can be modelled in this fashion very clearly and precisely. For 17th- and 18th-cen-

By courtesy of the Museum of Art, Rhode Island School of Design, Providence



Hard style, "View of Arles," by Vincent van Gogh, China Ink with reed pen and wash, 1888. In the Museum of Art, Rhode Island School of Design, Providence. 43.3 × 54.9 cm.

tury engravers, this process became the most important means of drawing. All of these different possibilities of linear rendition can be achieved with pen and crayon as well as with the brush.

Plane techniques. Linear techniques of drawing are supplemented by plane methods, which can also be carried out with crayon. For example, evenly applied dotting, which is better done with soft mediums (see *Surfaces, mediums, and techniques* below), results in an areal effect in uniform tone. Various values of the chiaroscuro (pictorial representation in terms of light and shade without regard to colour) scale can also be rendered by means of dry or moist rubbing. Pulverized drawing materials that are rubbed into the drawing surface result in evenly toned areas that serve both as a closed foundation for linear drawing and as indication of colour values for individual sections.

More significant for plane phenomena, however, is brushwork, which, to be sure, can adopt all linear drawing methods but the particular strength of which lies in stroke width and tone intensity, a medium that allows for extensive differentiation in colour tone and value. Emphases created by the repeated application of the same tone provide illusionistic indentations that can be conceived of spatially and corporeally. Colour differences result from the use of various mediums. Brushwork also lends itself to spatial and plastic representation, just as it can constitute an autonomous value in nonrepresentational drawings.

All of these effects of monochrome drawing are accentuated with the use of varicoloured mediums of a basic material; for example, coloured chalks, drawing inks, or watercolour. While these mediums enrich the art of drawing, they do not widen its basic range.

The drawing surface. To these graphic elements must be added another phenomenon the formal significance of which is restricted to drawing: the effect of the unmarked drawing surface, usually paper. Almost all studies (drawings of details), many autonomous sheets, most portrait drawings, as well as figure compositions, still lifes, and even landscapes stand free on the sheet instead of being closed off with a frame-line. Thus, the empty surface, suggesting by itself a spatial background to the drawing on it, contributes actively to the artistic effect.

Details conveyed by empty space

Even within line composition, the surface left blank fulfills an essential role as a representational value defined by the drawn statements surrounding it as body of a given substantiality or space of a given expanse. Among the details conveyed by the empty space may be the planes of a face, the smooth width of a garment, the mass of a figure or object, the substance the borders and nuances of which are indicated by the drawing. Even the space around individual objects, the spatial distance between them and their environment, the width of a river and the depth of a landscape may be merely signalled by the drawing and filled by the void.

This void can itself become the dominant form enclosed by lines or contours—for example, in decorative sketches and in many ornamental drawings that make use of the negative form, an effect attainable also by tinting the blank planes.

Relationship between drawing and other art forms. The bond between drawing and other art forms is of course very close, because the preliminary sketch was for a long time the chief purpose of the drawing. A state of mutual dependence exists in particular between painting and drawing, above all, in the case of sketches and studies for the composition of a picture. The relationship is closest with preliminary sketches of the same size as the original, the so-called cartoons whose contours were pressed through or perforated for dyeing with charcoal dust. Once transferred to the painting surface, the sketch had served its purpose.

On autonomous sheets, too, the close connection between drawing and painting is evidenced by the stylistic features that are common to both. Drawing and painting agree in many details of content and form. Measurements; proportions of figures; relationship of figure to surrounding space; the distribution of the theme within the composition according to static order, symmetry, and equilibrium of the masses or according to dynamic contrasts, eccentric vanishing points, and overaccentuation of individual elements; rhythmic order in separate pictorial units in contrast to continuous flow of lines—all of these formal criteria apply to both art forms. The uniform stylistic character shared by drawing and painting is often less severely expressed in the former because of the spontaneous flow of the unfettered artist's stroke, or "handwriting," and of the struggle for form as recorded in the pentimenti (indications in the drawing that the artist had changed his mind and drawn over his original formulation). Furthermore drawing can stimulate certain aspects of movement more easily than painting can through the rhythmic repetition of a contour or the blended rubbing of a sharp borderline.

Still closer, perhaps, is the bond between drawing and engraving, which works with the same artistic means, with monochrome linearity as its main formal element and with various tone and plane methods closely related to those of drawing.

Drawing is more independent than sculpture because sculpture uses a three-dimensional model. As a result, sculptors' drawings can always claim a greater degree of autonomy. (For the special position of the architectural sketch, see *Subject matter of drawing* below.)

SURFACES, MEDIUMS, AND TECHNIQUES

Types of ground. One can draw on practically anything that has a plane surface (it does not have to be level); for example, papyrus and parchment, cloth, wood, metals, ceramics, and even walls, glass, and sand. (With some of these, to be sure, another dimension is introduced through indentations that give the visual effect of lines.) Ever since the 15th century, however, paper has been by far the most popular ground.

Paper manufacturing

The technique of paper manufacturing, introduced from East Asia by the Arabs, has remained virtually unchanged for the past 2,000 years. A fibrous pulp of mulberry bark, hemp, bast, and linen rags is drained, pressed, and dried in flat molds. The introduction of wood pulp in the mid-19th century, which enabled manufacturers to satisfy the enormously increased demand for bulk paper, did not affect art paper because paper of large wood content yellows quickly and is therefore ill-suited for art drawing. The essential preparation of the paper to give it a smooth and even surface for writing or drawing was once done by rubbing it with bone meal, gypsum chalk, or zinc and titanium white in a very thin solution of glue and gum arabic. The proper priming, achieved through repeated rubbing and polishing, was of the utmost importance, especially for metalpoint drawings. If such preparation is too weak, the paper accepts the stroke badly; if it is too strong, the coating cracks and chips under the pressure of the hand. Since the early 15th century, however, the sheets have been given the desired smooth and nonabsorbent consistency by dipping them in a glue or alum bath. The addition of glue also made it possible to impart to the pulp paper a quality that permitted pen drawings. Pigments, too, could of course be added to the pulp, and the so-called natural papers—chiefly blue and called Venetian papers after the centre of the retail trade in this commodity—became more and more popular. While the 17th century liked half tints of blue, grey, brown, and green, the 18th preferred warm colours such as ivory and beige, along with blue. Since the 18th century, paper has been manufactured in all conceivable colours and half tones.

The range of quality has also greatly increased since the end of the 18th century to give more painstakingly produced drawing papers. Even in earlier times, the absorbent Japan paper made of mulberry bark enjoyed great popularity. Handmade paper, stronger and free of wood; with an irregular edge, has remained to this day a favourite surface for drawings. Vellum, delicate and without veins, resembles parchment in its smooth surface. Modern watercolour paper is a pure linen paper glued in bulk and absolutely free of fat and alum; its two surfaces are of different grain. For pastel drawings, a firm, slight-

ly rough surface is indicated, whereas pen drawings are best done on a very smooth paper.

Granulated and softer drawing tools, such as charcoal, chalk, and graphite are not as dependent on a particular type of paper; but, because of their slight adhesiveness, they often require a stronger bond with the foundation as well as some form of surface protection. This process of fixing was formerly done through repeated varnishing with gum-arabic solution and even with glue or egg-white emulsion. Modern siccatives (drying substances) inhibit discoloration but cannot prevent the living surface from appearing sealed, as it were, under a skin. In pastels especially, the manifold prismatic effects of finely powdered coloured crayons are thus lost, and the bright and airy surface is turned into an amorphous, heavy layer. Pastels, which brush off easily, are therefore best preserved under glass.

Tools and techniques. Such varied tools as slate pencils, charcoal, metal styli, and chalks may be used for drawing as well as all writing utensils, including pens, pencils, and brushes; indeed, even chisels and diamonds are used for drawing. Dry drawing tools differ in effectiveness from liquid ones because it is not irrelevant from the artistic point of view whether one uses a self-drawing medium that permits an evenly flowing line dependent only on hand pressure or a transferring tool that must be put down periodically and refilled with resultant differences in the strength and concentration of the line. Modern drawing mediums that combine both possibilities, such as fountain pens, ball-point pens, and felt pencils, are recent inventions.

Dry and liquid drawing tools

No less varied than the nature and composition of these drawing mediums is their aesthetic effect. It would nevertheless be wrong to systematize the art of drawing on the basis of the techniques applied; not only does almost every technique have several applications but it can also be combined with other techniques, and the draftsman's temperament inevitably plays a role as well. Even if certain techniques predominate in certain periods, the selection of drawing mediums depends on the intended effect and not vice versa. Artists have always been able to attain the desired effect with a variety of techniques. Dry mediums, for example, are predestined for clear lines, liquid ones for plane application. Yet extremely fine strokes can also be made by brush, and broad fields can be marked in with pencil or crayon. Some mediums, including charcoal, one of the oldest, if not the oldest of all, allow both extremes.

Charcoal. In every hearth or fireplace, partially consumed pieces of wood remain that can be used as a convenient tool for drawing. Evidence of charcoal sketches for mural, panel, and even miniature paintings can still occasionally be seen under the pigment. Drawing charcoal produced from wood that is as homogeneous as possible gives a porous and not very adhesive stroke. The pointed charcoal pencil permits hair-thin lines; if used broadside on the surface, it creates evenly toned planes. Rubbing and pulverizing the charcoal line results in dimmed intermediate shades and delicate transitions. Because of its slight adhesiveness, charcoal is eminently suited to corrective sketching; but if the drawing is to be preserved, it must be protected by a fixative.

As a medium for quick, probing sketches and practice in studying models, charcoal was once much used in all academies and workshops. The rapid notation of difficult poses, such as Tintoretto demanded of his models, could be done quickly and easily with the adaptable charcoal pencil. While some of these sheets were deemed worthy of preservation, hundreds have surely been lost.

Charcoal used for portraits

Charcoal has often been used for portrait drawings to preserve for the eventual painting pictorial tints that were already present in the preliminary sketch. When destined to be autonomous portraits, charcoal drawings are executed in detail; with their sharp accents and delicate modelling, such portraits cover the whole range of the medium. In "Portrait of a Lady," by the 19th century French painter Édouard Manet, the grain of the wood in the chair, the fur trimming on the dress, the compactness of the coiffure, and the softness of the flesh are all rendered in the same material: charcoal. Popular as that material was for studies and sketches, it has been used for independent drawings destined for preservation by only a few artists; for example, the 17th-century Dutch painter Paulus Potter. It is somewhat more frequent among the great draftsmen of the 19th and 20th centuries, such as Edgar Degas, Henri de Toulouse-Lautrec, Käthe Kollwitz, and Ernst Barlach.



"Portrait of a Lady," charcoal drawing by Édouard Manet (1832–83). In the National Museum Vincent van Gogh, Amsterdam. 54 ×45 cm.
By courtesy of the National Museum Vincent van Gogh, Amsterdam

Oiled charcoal, with the charcoal pencils dipped in linseed oil, provides better adhesion and a deeper black. Used in the 16th century by Tintoretto, this technique was applied above all by the Dutch draftsmen of the 17th century in order to set deep-black accents. The advantage of better adhesion in the indentations of the paper in contrast to dry charcoal, which sticks to the elevations, has to be paid for, however, by "incorrigibility"; *i.e.*, correction cannot be made. In addition, charcoal crayons that have been deeply dipped in oil show a brownish streak left by the oil alongside the lines.

Chalks. The chalks, which resemble charcoal pencils in outward appearance, are an equally important drawing medium. If charcoal was primarily a medium for quick sketching that could be corrected and for the search for artistic form, chalk drawing, which can also fulfill all of these functions, has steadily gained in importance as an autonomous vehicle of expression. Since the end of the 15th century, stone chalk, as found in nature, has become increasingly more significant in art drawing. As a basic material, alumina chalk has various degrees of hardness, so that the stroke varies from slightly granular to homogeneously dense and smooth.

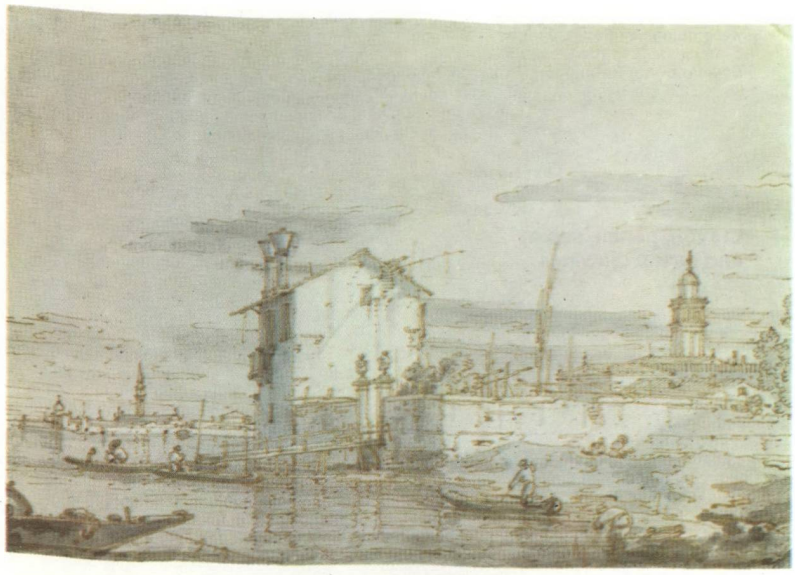
The attempt to produce a crayon or pencil of the greatest possible uniformity has led to the production of special chalks for drawing; that is, chalks, which, after being pulverized, washed, and molded into convenient sticks, allow a softer and more regular stroke and are also free of sandy particles. The admixture of pigments (carbons in the case of black chalks) creates various tints from a rich black to a brownish gray; compared to the much-used black chalk, the brown variety is of little significance. White chalk, also found in nature, is rarely employed as an independent medium for drawing, although it is frequently used in combination with other mediums in order to achieve reflections of light as individual accents of plastic modelling.

Beginning with the 15th century, chalk has been used

Pen and ink, chalk, silverpoint, and wash drawings



"Head of Girl" (study for "The Virgin of the Rocks," 1483), silverpoint on light brown paper by Leonardo da Vinci. In the Biblioteca Reale, Turin, Italy. 18.2 × 15.9 cm.



"An Island in the Lagoon," pen, brown ink, and carbon ink wash over ruled pencil lines by Canaletto (1697–1768). In the Ashmolean Museum, Oxford, England. 18.3 × 27.8 cm.





Study probably for "L'Indifférent," black, red, and white chalk on yellowish-gray paper by Jean-Antoine Watteau (1684–1721). In the Museum Boymans-van Beuningen, Rotterdam. 27.2 × 19 cm.

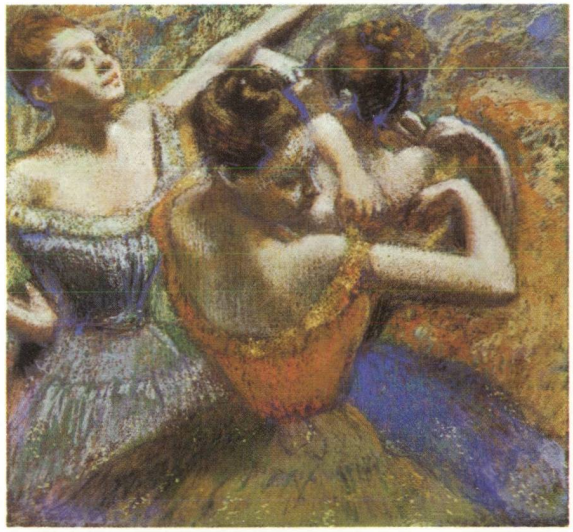
"Madonna with Many Animals," pen, ink, and watercolour by Albrecht Dürer, *c.* 1503. In the Albertina, Vienna. 32.1 × 24.3 cm.

Plate 1: By courtesy of (top left) Biblioteca Reale, Turin, Italy, (top right) the Ashmolean Museum, Oxford, England, (bottom left) the Albertina, Vienna, (bottom right) the Museum Boymans-van Beuningen, Rotterdam



"Le Château Noir," pencil and watercolour by Paul Cézanne, *c.* 1895–1900. In the Museum Boymans-van Beuningen, Rotterdam. 36 × 52.6 cm.

Crayon, pencil, sepia, and pastel drawings



"The Dancers," pastel on paper by Edgar Degas, 1899. In the Toledo Museum of Art, Ohio. 62 × 64.5 cm.



"Trees and a Stretch of Water on the Stour," pencil and sepia wash by John Constable (1776–1837). In the Victoria and Albert Museum, London. 16.2 × 20.3 cm.



"Two Women III," crayon on paper by Willem de Kooning, 1952. In the Allen Memorial Art Museum, Oberlin College, Ohio. 37.5 × 47 cm.

Plate 2: By courtesy of (top) the Museum Boymans-van Beuningen, Rotterdam, (centre right) Victoria and Albert Museum, London, (bottom) Allen Memorial Art Museum, Oberlin College, Ohio; photograph, (centre left) The Toledo Museum of Art, Ohio, gift of Edward Drummond Libbey

increasingly for studies and sketches. Its suitability for drawing exact lines of any given width and also for laying on finely shaded tints makes it particularly appropriate for modelling studies. Accents that stress plastic phenomena are applied by varying the pressure of the hand. Characteristic details in portrait drawings in particular can be brought out in this manner. Pictorial values as well as light and shadow effects can be rendered with chalk without losing their firm, plastic form. For the same reason, chalk is also most valuable in sketching out paintings and indicating their values.

All of these qualities explain why chalk is such a good medium for autonomous drawings. Indeed, there is scarcely a draftsman who has not worked in chalk, often in combination with other mediums. Aside from portrait drawings done all over the world, landscapes have formed the main theme of chalk drawings, especially with the Dutch, in whose art landscape drawings have played a large role. Ever since the invention of artificial chalk made of lampblack (a fine, bulky, dull-black soot deposited in incomplete combustion of carbonaceous materials), which possesses a precisely measurable consistency—an invention ascribed to Leonardo da Vinci—the pictorial qualities of chalk drawing have been fully utilized. Chalks range from those that are dry and charcoal-like to the fatty ones used by lithographers.

Use of sanguine

Another very important drawing pencil is similarly a chalk product: the red pencil, or sanguine, which contains ferric oxide, which occurs in nature in shadings from dark brown to strong red and can also be manufactured from the same aluminum-oxide base with ferric oxide or rust added. Besides the stronger pictorial effect possible because of its chromatic value, sanguine also possesses a greater suppleness and solubility in water. Thus, a homogeneous plane can be created through moist rubbing, a compact stroke through liquid linear application, a very delicate tone through light wiping. Although this oxide was used for red tints in prehistoric painting, sanguine does not seem to have acquired artistic dignity until the 15th century, when it became customary to fix drawings by painting them over with a gum solution, for sanguine has no more adhesiveness than charcoal. In the 15th century, sanguine was a popular drawing medium because of its wealth of pictorial possibilities. Those inclined to be colorists—such as the portraitists Jean Clouet and Hans Holbein, the Flemish painters around Peter Paul Rubens, and, above all, the French artists of the 18th century—particularly favoured it. The possibilities of sanguine range from suggestive forms with markedly plastic values to a very pictorial, soft rendition of visual surface stimuli.

A combination of various chalks offers still richer coloristic possibilities. Black chalk and sanguine have been widely used since the 16th century to achieve colour differentiation between flesh tones, hair, and the material of garments. The combination of black and white chalk serves plastic modelling, as does that of the softer sanguine with white chalk; in the former case, the accentuation rests with the black, in the latter, with the more suggestive delineation in white.

A decidedly coloristic method lies in the combination of various chalk colours with one another and with tinted paper. Such pictorially executed sheets, called *à deux crayons* (with two colours) and *à trois crayons* (with three colours), respectively, were especially popular in 17th- and 18th-century France. Antoine Watteau reached a previously unheard of harmony of different chalks on natural paper. With the three colours, Nicolas Lancret, Jean-Étienne Liotard, Jacques-André Portail, François Boucher—to name but a few such artists—achieved sensitive drawings that are very appealing coloristically.

Pastel crayons

An additional colour refinement is made possible with pastel crayons. An ample selection of dry colour pigments in pastel crayons, prepared with a minimum of agglutinants and compounded with different shades of white for the articulation of tints, is commercially available. The colours can be laid on in linear technique directly with the crayons, but an area application made with a piece of soft suede or directly, with the fingers, is more frequent. Although this technique was known to the Accademia degli Incamminati (to the painter Guido Rent, for example) as early as the 17th century, it did not reach its flowering until the 18th century, especially in France (with Jean-Marc Nattier and Jean-Baptiste-Siméon Chardin) and in Venice (with Rosalba Carriera). Pastel chalks are particularly favoured for portraits; their effect approximates that of colour-and-area painting rather than line drawing.

In the 19th and 20th centuries, Degas reverted to a stronger accentuation of the delineatory aspects of drawing. With intermediate varnishes he achieved an overlay drawing with different colours and thus an increased emphasis on individual strokes. This technique, fundamentally different from the older one, was imitated with minor variations by Odilon Redon, Gustave Moreau, Jean-Édouard Vuillard, Pierre Bonnard, and others. It has also been borrowed by such Expressionist artists as Edvard Munch and Ernst Ludwig Kirchner.

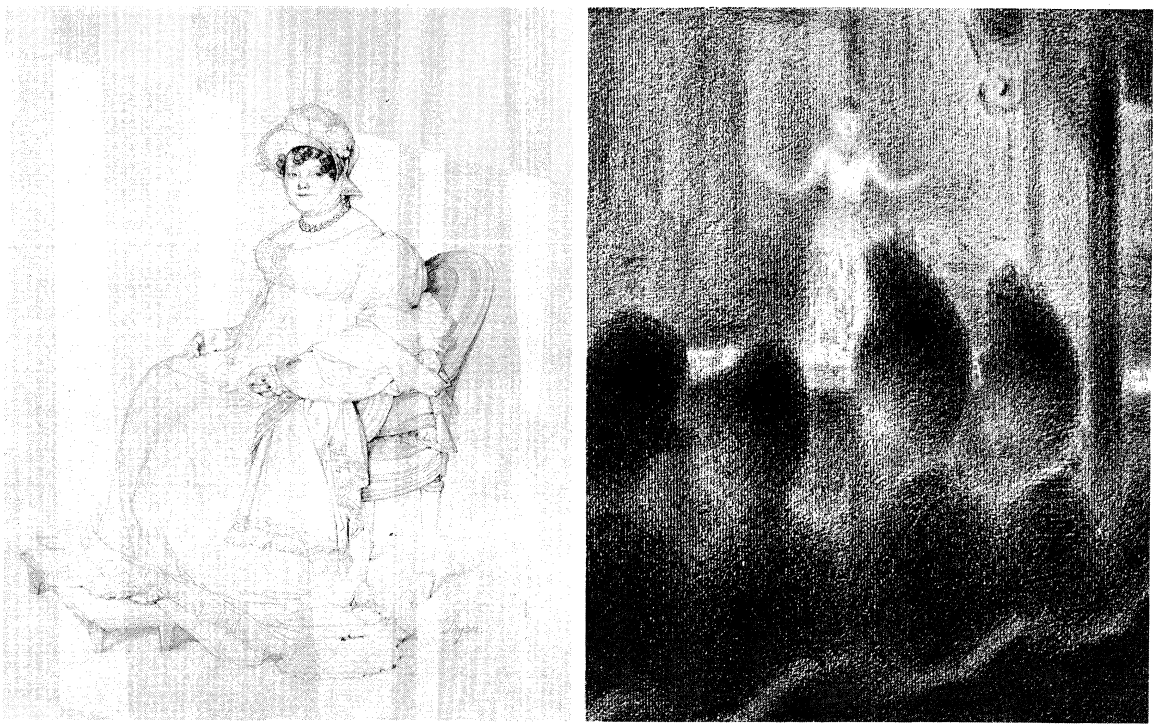
Modern grease chalks offer a chromatic scale of similar range. Developed originally for such technical purposes as the lettering of very smooth surfaces, such as metal or glass, they can be applied in the same flat manner as pastels, although with the opposite aesthetic effect: that of compact colours. It was the 20th-century English sculptor Henry Moore who first and convincingly exploited the feasibility of continuing, with other mediums, such as pen or watercolour, work on the firm surface that had been led out with grease chalks.

Metalpoints. Metalpoints have been used for writing and delineation ever since the scriptoria of antiquity. It required little imagination to employ them also in drawing. The most frequently used material was soft lead, which on a smooth surface comes out pale gray, not very strong in colour, and easily erasable but very suitable for preliminary sketches. Aside from lead, tin and copper were also used, as well as sundry lead-and-pewter alloys. The 15th-century Venetian painter Jacopo Bellini's book of sketches in London with leadpoint drawings on tinted paper, is a particularly valuable example of this technique, even if individual portions and, indeed, entire pages that had become effected were drawn over long ago. One can see little more than the traces left by the pencil because, as in many other metalpoint drawings, the sketches were redrawn in another medium. Botticelli, for example, sketched with a leadpoint the outline of his illustrations to Dante's *Divine Comedy*, retracing them afterward with the pen. Metalpoints were used into the 18th century for perspectivist constructions and auxiliary delineation, especially in architectural drawings.

Silverpoint drawing

More suited to permanent drawing is the silverpoint, which requires special preparation of the foundation and, once applied, cannot be corrected. Its stroke, also pale gray and mostly oxidized by now into brown, adheres unerasably. Silverpoint drawings accordingly require a clearer concept of form and a steady hand because corrections remain visible. Because too much pressure can bring about cracks in the foundation, the strokes must be even; emphases, modelling, and light phenomena must be rendered either by means of dense hachures, repetitions, and blanks or else supplemented by other mediums. Despite these difficulties, silverpoint was much used in the 15th and 16th centuries. Dürer's notebook on a journey to Holland shows landscapes, portraits, and various objects that interested him drawn in this demanding technique. Silverpoint was much in demand for portrait drawings from the 15th into the 17th century; revived in the 18th-century Romantic era, it is still occasionally used by modern artists.

Graphite point. Toward the end of the 16th century, a new drawing medium was introduced and soon completely displaced metalpoint in sketching and preliminary drawing: the graphite point. Also called Spanish lead after its chief place of origin, this drawing medium was quickly and widely adopted; but because of its soft and smeary consistency it was used for autonomous drawings only by some Dutch painters, and even they employed it mostly in conjunction with other points. (It might be added that the graphite point was originally taken for a



Graphite mediums.
(Left) "Portrait of Mme. Guillaume Guillon Lethière," lead pencil drawing by Jean-Auguste-Dominique Ingres (1780–1867). In the Fogg Art Museum, Harvard University. 27 × 16.4 cm. (Right) "Au concert européen," conté crayon drawing by Georges Seurat, *c.* 1887. In the Museum of Modern Art, New York. 31.1 × 24.8 cm.
By courtesy of (left) the Fogg Art Museum, Harvard University, Grenville L. Winthrop bequest; (right) the Museum of Modern Art, New York, Lillie P. Bliss Collection

metal because its texture shines metallically in slanting light.) The lead pencil, or more properly *crayons Conté*, became established in art drawing after Nicolas-Jacques Conté invented, around 1790, a manufacturing process similar to that used in the production of artificial chalk. Purified and washed, graphite could henceforth be made with varying admixtures of clay and in any desired degree of hardness. The hard points, with their durable, clear, and thin stroke layers, were especially suited to the purposes of Neoclassicist and Romantic draftsmen. The Germans working in Rome, in particular, took advantage of the chance to sketch rapidly and to reproduce, in one and the same medium, subtle differentiations as well as clear proportions of plasticity and light. Among the most masterful pencil artists of all was Ingres, who presketched systematically in pencil the well-thought-out structure of his paintings.

The more pictorially inclined artists of the late 19th century, such as Ferdinand Delacroix, preferred softer pencils in order to throw into plastic relief certain areas within the drawing. Seurat, on the other hand, reached back to graphite in his drawings from the concert cafés, among them "Au concert européen" (Museum of Modern Art, New York) in which he translated the pointillistic technique (applying dots of colour to a surface so that from a distance they blend together) into the monochrome element of drawing. Pencil frottage (rubbing made on paper laid over a rough surface), first executed by the Surrealist artist Max Ernst, represents a marginal kind of drawing.

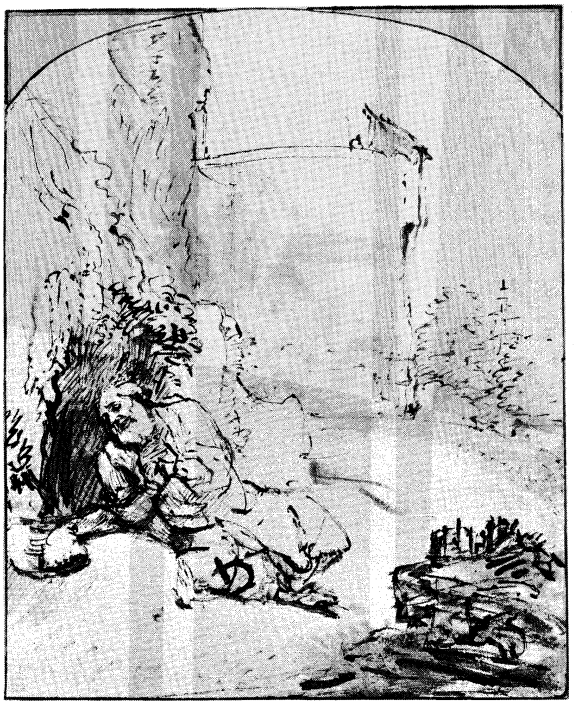
Pencil frottage

Coloured crayons. Coloured crayons, in circulation since the later 19th century, offer all the possibilities of black graphite points; and, in combinations, they attain a stronger colour value than chalks because they do not merge with one another. Every line preserves its original and characteristic colour, a form of independence that Gustav Klimt and Picasso exploited to the full.

Incised drawing. A role apart is that played by incised drawings. Their pronounced linearity gives them the visual appearance of other drawings; materially, however, they represent the opposite principle, that of subtracting from a surface rather than adding to it. Incised drawings are among the oldest documents of human activity. In primitive African cultures, the methods and forms of prehistoric bone and rock drawings have survived into the present. In a decorative and conceivably also symbolic form, incised decorations on pottery have existed for thousands of years; insofar as the comparison is valid, they correspond in every formal respect to applied drawings of the same period. A formal equivalent may also be observed in later times: in the decorative details of implements, especially metal—from the drawings on Greek mirrors, through the jewelry made at the end of the Roman Empire, to the scenes on medieval weapons and, above all, on Renaissance dress armour. More often than not these are drawings that follow certain models rather than free drawings in the sense of sketches.

Logically, one would also have to consider all niello work under the heading of drawing, because the picture in this case is cut out of the metal and filled with a deep black-coloured paste so that it appears to the eye as a linear projection on a plane. In like manner, work with the graver or burin (cutting tools) and with the etching needle on the engraving plate may be considered to parallel in its execution that gradual effort applied directly to the carrier that was defined earlier as the art of drawing. The difference lies in the fact that this work is not a goal in itself but the prerequisite for a printing process that is intended to be repetitive.

Brush, pen, and dyestuffs. Of the many possibilities of transferring liquid dyestuffs onto a plane, two have become particularly significant for art drawing: brush and pen. To be sure, finger painting, as found in prehistoric cave paintings, has occasionally been practiced since the late Renaissance and increasingly so in more recent times. For drawing as such, however, the method is irrelevant. Similarly, the use of pieces of fur, frayed pieces of wood, bundles of straw, and the like is more significant as a first step toward the camel's-hair brush than as indication that these objects were ever drawing mediums in their own right. Although it is antedated by the brush, which in some cultures (East Asia, for example) has remained in continued use, the pen has been the favorite writing and drawing tool ever since classical antiquity.



"The Prophet Jonah Before the Walls of Nineveh," by Rembrandt, reed pen in bistre with wash, c. 1654–55. In the Albertina, Vienna. 21.7 × 17.3 cm.
By courtesy of the Albertina, Vienna

The principle of transferring dyestuffs with the pen has remained virtually unchanged for thousands of years. The capillary effect of the split tip, cut at a slant, applies the drawing fluid to the surface (parchment, papyrus, and, since the late Middle Ages, almost exclusively paper) in amounts varying with the saturation of the pen and the pressure exerted by the drawing hand. The oldest form is that of the reed pen; cut from papyrus plants, sedge, or bamboo, it stores a reservoir of fluid in its hollow interior. Its stroke—characteristically powerful, hard, and occasionally forked as a result of stronger pressure being applied to the split tip—became a popular medium of artistic expression only with the rise of a subjective view of the artist's personality during the Renaissance. Rembrandt made superb use of the strong, plastic accents of the reed pen, supplementing it as a rule with other pens or brushes. Beginning in the 19th century with the Dutch artist van Gogh, pure reed-pen drawings with a certain forcefulness of expression have been created by all artists. Expressionists such as George Grosz used the reed pen frequently.

The reed pen

If the selection of the reed pen already implies a formal statement of sorts, that of the quill pen opens up a far wider range of possibilities. Ever since the rise of drawing in Western art—that is, since the late Middle Ages—the quill has been the most frequently used instrument for applying liquid mediums to the drawing surface. The importance accorded to this tool is attested by the detailed instructions in painters' manuals about the fashioning of the pen from wing shafts of geese, swans, and even ravens. The supple tip of the quill, available in varying strengths, permits a relatively wide scale of individual strokes—from soft, thin lines, such as those used in preliminary sketches for illustrations in illuminated books, through waxing and waning lines that allow differentiation within the stroke, to energetic, broad lines. It was only when metal pens began to be made of high-grade steel and in different strengths that they became a drawing implement able to satisfy the demands made by the individual artist's hand.

Although all dyestuffs of low viscosity lend themselves to pen drawing, the various inks are most often employed. The manufacture of gallnut ink had been known from the medieval scriptoria, copying rooms set apart for scribes in monasteries. An extract of gallnuts mixed with iron vitriol and thickened with gum-arabic solution produces a writing fluid that comes from the pen black, with a strong hint of purple violet, and dries almost black. In the course of time it turns a darkish brown, so that the writing fluid in old manuscripts and drawings cannot always be identified by the colour alone. In contrast to other brown writing fluids, the more strongly coloured parts of gallnut ink remain markedly darker; and because inks of especially great vitriol content decompose the paper, the drawing, particularly in its more coloured portions, tends to shine through on the reverse side of the sheet. Only industrially produced chemical inks possess the necessary ion balance to forestall this undesirable effect.

Another ink, one that seems to have found no favour as a writing fluid but has nonetheless had a certain popularity in drawing, is bister, an easily dissolved, light-to-dark-brown transparent pigment obtained from the soot of the lampblack that coats wood-burning chimneys. Its shade depends both on the concentration and on the kind of wood from which it is derived, hardwoods (especially oaks) producing a darker shade than conifers, such as pine. During the pictorially oriented Baroque period, in the 17th and early 18th centuries the warm tone that can be thinned at will made bister a popular medium with which to supplement the planes of a pen drawing.

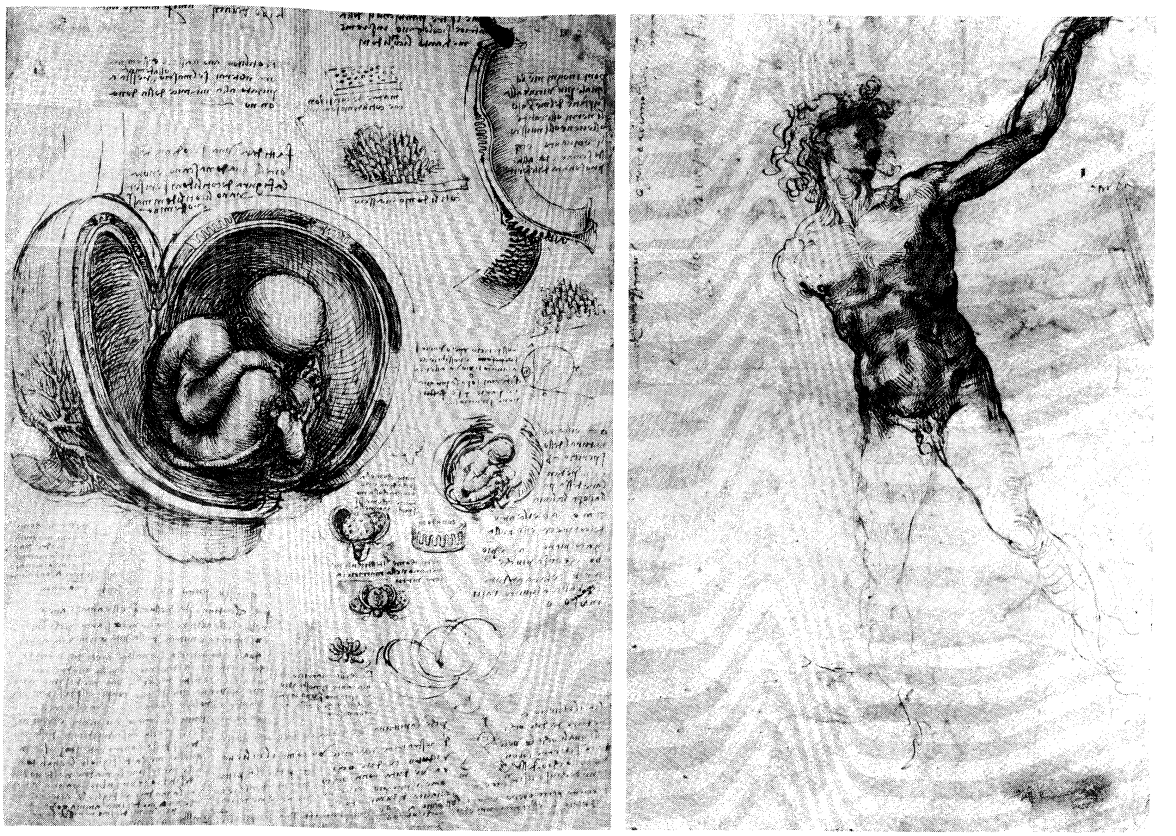
Also derived from a carbon base is India ink, made from the soot of exceptionally hard woods, such as olive or grape vines, or from the fatty lampblack of the oil flame, with gum-arabic mixed in as a binding agent. This deep-black, thick fluid preserves its dark tone for a long time and can be thinned with water until it becomes a light gray. Pressed into sticks or bars, it was sold under the name of Chinese ink or India ink. This writing fluid, known already in Egypt and used to this day in China and India, has been manufactured in Europe since the 15th century. Favoured in particular by German and Dutch draftsmen because of its strong colour, it lent itself above all to drawing on tinted paper. Since the 19th century, India ink has been the most popular drawing ink for pen drawings, replacing all other dyestuffs in technical sketches. Only very recently have writing inks gained some significance in art drawing—in connection with the practical fountain pen.

Chinese or India ink

For a relatively short time, a dyestuff of animal origin, sepia, obtained from the pigment of the cuttlefish, was used for drawing. Known ever since Roman times, it did not come into more general use until the 18th century. Compared to yellowish bister, it has a cooler and darker tone, and is brown with a trace of violet. Until the 18th century, it was employed by such amateur painters as the German poet Johann Wolfgang von Goethe because of its effectiveness in depth; as a primary pigment, however, it has been completely replaced by industrially manufactured watercolours.

Other dyestuffs are of only minor importance compared with these inks, which are primarily used for pen drawings. Minium (red lead) was used in the medieval scriptoria for the decoration of initial letters and also in illustrated pen drawings. Chinese white is easier to apply with a pointed brush because of its thickness; other pigments, among them indigo and green copper sulphate, are rarely found in drawings. For them, too, the brush is a better tool than the pen. The systematically produced watercolours of various shades are almost wholly restricted to technical drawings.

In combination with written texts, pen drawings are among the oldest artistic documents. Already in classical times, texts were illustrated with firm contours and sparse interior details. During the Middle Ages, marginal drawings and book illustrations were time and again presketched, if not definitively executed, with the pen. In book painting, decidedly delineatory styles developed in which the brush was also employed in the manner of a pen drawing: for example, in the Carolingian school of Reims, which produced the Utrecht Psalter in the 9th century, and also in southern Germany, where a separate illustrative form with line drawings was widespread with the *Biblia Pauperum* ("Poor People's Bibles," biblical picture books used to instruct large numbers of people



Individual expression in pen drawing of the Renaissance.
(Left) "Foetus in Utero," scientific drawing by Leonardo, pen and ink with red chalk, *c.* 1510–12. In the Royal Library, Windsor Castle. 30.1 × 21.4 cm. (Right) "Running Youth with Left Arm Extended," study for a sculpture by Michelangelo, pen and brown ink, *c.* 1504. In the British Museum. 37.5 × 22.9 cm.
(Left) Copyright reserved; (right) by courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

in the Christian faith). The thin-lined outline sketch is also characteristic of the earliest individual drawings of the late Middle Ages and early Renaissance. Sketches after ancient sculptures or after nature as well as compositions dealing with familiar motifs form the main themes of these drawings. Such sheets were primarily used as models for paintings; gathered in sketchbooks, they were often handed on from one generation to the next. The practical usefulness of these drawings is attested by the supplements added to them by younger artists and by the fact that many metalpoint drawings that had become hard to decipher were redrawn with the pen, as shown by the sketchbooks of the 15th-century Italian artist Antonio Pisanello, now broken up and preserved in several different collections.

Use of sketchbooks

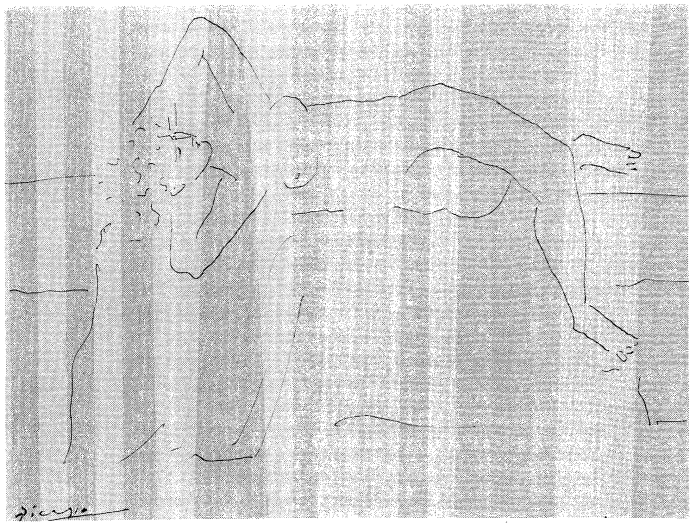
In the 16th century, the artistic range of the pen drawing reached an individual articulation that it hardly ever attained again. Every artist was free to exploit with the pen the formal possibilities that corresponded to his talents. Thus Leonardo used a precise stroke for his scientific drawings; Raphael produced relaxed sketches, in which he probed for forms and variations of form; Michelangelo drew with short strokes reminiscent of chisel work; Titian contrasted light and dark by means of hachures laid broadly over the completed figures. Among the Northerners, Dürer mastered all the possibilities of pen drawing, from quick notation to the painstakingly executed autonomous drawing, ranging from a purely graphic and delineatory technique to a spatial and plastic modelling one; it is no wonder that he stimulated so many other artists. The subjective attitude of the later 16th century is often expressed more clearly in Mannerist drawings—characterized by spacial incongruity and excessive elongation of the human figures, which are as revelatory of the artist's personality as handwriting—than it is in completed works of painting and sculpture. A special form of exact drawing is found in models for engravings; some of these were directly mounted on the wood block; some anticipate the style of the copperplate engraving in the pen-drawing stage, with waxing and waning lines, delicate stroke layers, and cross-hatching for spatial and plastic effects.

In the 17th century, the pen drawing took second place to combined techniques, especially wash, a sweep or splash of colour, applied with the brush. An open style of drawing that merely hints at contours, along with contrasting thin and powerful strokes, endowed the line itself with expressive qualities. In his numerous drawings, Rembrandt in particular achieved an exceedingly subtle plastic characterization and even light values through the differentiation of stroke layers and the combination of various pens and brushes.

Additional techniques came to the fore in the 18th century, with the pen sketch providing the scaffold for the drawing that was carried out in a pictorial style. Only decorative sketches and practical studies were laid out more often as linear drawings.

The closed, thin-contour drawing regained its importance with Neoclassicism at the end of the 18th century. The Nazarenes (the nickname of the Lucas Brotherhood —later Guild of St. Luke, who lived in monastic style) and Romanticists consciously referred to the early-Renaissance manner of drawing, modelling with thin lines. With closed contours, carefully set hair-and-shadow strokes, and precise parallel hachures, they attained plastic values by purely graphic means.

This technique was again followed by a more pictorially oriented phase, culminating in the late 19th century in the recognition of drawing as the most immediate and personal expression of the artist's hand. The pure pen drawing took its place by the side of other highly esteemed art forms. The English Art Nouveau artist Aubrey Beardsley at the end of the 19th century applied the direct black–white contrast to planes, while in the 20th



Pure line pen and ink drawing of the 20th century.
"Reclining Nude," by Pablo Picasso (1881–1973). In the Fogg Art Museum, Harvard University. 26 × 35 cm.
By courtesy of the Fogg Art Museum, Harvard University, the Meta and Paul Sachs Collection

century the French masters Henri Matisse and Picasso reduced the object to a mere line that makes no claim to corporeal illusion. A large number of illustrators, as well as the artists who draw the comic strips, prefer the clear pen stroke. In the Russian artist Wassily Kandinsky's nonrepresentational compositions, finally, the independence of the line as an autonomous formal value became a new theme in drawing. In the hair-thin automatist seismograms (so-called because of their resemblance to the records of earthquakes) of the 20th-century German artist Wols (Alfred Otto Wolfgang Schulze), which are sensitive to the slightest stirring of the hand, this theme leads to a new dimension transcending all traditional concepts of a representational art of drawing.

The use of a brush in drawing considerably antedates that of the pen; in some cultures, notably in East Asia, the brush is still the most popular writing tool. Although it is best suited to the flat application of pigments—in other words, to painting—its use in a clearly delineatory function, with the line dominating and (a crucial property of brush drawing) in monochrome fashion, can be traced back to prehistoric times.

All of the above-mentioned drawing inks have been used as dyes in brush drawings, often with one and the same pigment employed in combined pen-and-brush work. Still greater differentiation in tone is often obtained through concentrated or thinned mediums and with the addition of supplementary ones. To the latter belong chiefly distemper, a paint in which the pigments are mixed with an emulsion of egg or size or both, and watercolours, which can be used along with bister and drawing ink. Even oils can sometimes be used for individual effects in drawing, as in the works of the 17th-century Flemish painter Jacob Jordaens.

Technique of sinopia

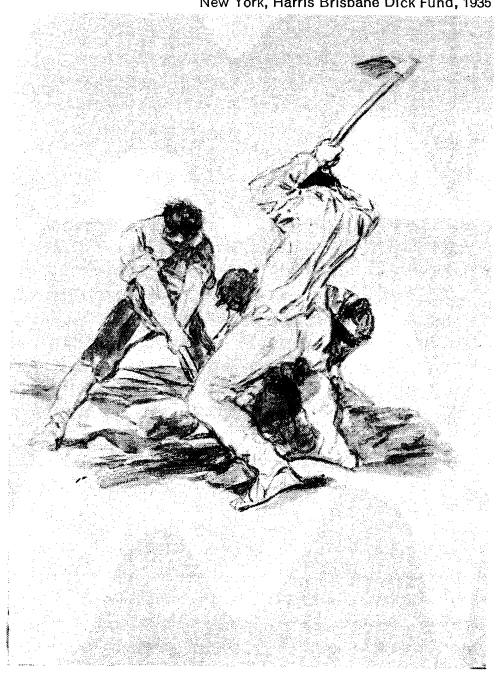
Sinopia, the preliminary sketch for a monumental wall painting, was done with the brush and has all the characteristics of a preparatory, form-probing drawing. The sketch was carried out directly on the appropriate spot and covered over with a thin layer of plaster, on which the pictorial representation was then painted.

The brush drawing differs from the pen drawing by its greater variation in stroke width, and by the stroke itself, which sets in more smoothly and is altogether less severely bordered. Early brush drawings nonetheless show a striking connection with the technique of the pen drawing. The early examples of the 15th century completely follow the flow of contemporaneous pen drawings. Leonardo's or Dürer's pen drawings, with their short, waxing and waning stroke layers, refine the system of pen drawing; many 16th-century artists used a comparable technique. The brush drawing for chiaroscuro sheets on tinted paper was popular because Chinese white, the main vehicle of delineation in this method, is more easily applied with the brush than the pen and because the intended pictorial effect is more easily attained, thanks to the possibility of changing abruptly to a plane representation.

Such representations are particularly distinctive as done by Vittore Carpaccio and Palma il Giovane in Venice and in a Mannerist spotting technique used by Parmigianino. In the 16th century, the brush nevertheless played a greater role as a supporting than as independently form-giving instrument. Pure brush drawings were rare even in the 17th century, although the brush played a major role in landscapes, in which, by tinting of varying intensity, it ideally fulfilled the need to provide for all desired degrees of spatial depth and strength of lighting. Some Dutchmen, such as Adriaen Brouwer, Adriaen van Ostade, and Jan Steen, transcended the limits of drawing in the narrower meaning of the term by doing brushwork limited to a few tones within a monochrome scale, giving the impression of a pictorial watercolour.

Although the coloristically inclined 18th century was little interested in the restriction to a few shadings within one colour value, Jean-Honoré Fragonard raised this technique to perfection, with all its possibilities of sharply accented contours, soft delineation, delicate tones, and deep shadows. The brush drawings of the Spanish painter Francisco Goya must also be counted among the

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1935



"Three Men Digging," by Francisco de Goya (1746–1828), brush drawing in sepia. In the Metropolitan Museum of Art, New York. 21 × 14 cm.

great achievements of this technique. In his strong plastic effects, the English painter George Romney made the most of the contrast between the white foundation and the broad brushstrokes tinted in varying intensities. Other English artists, among them Alexander Cozens, John Constable, and J.M.W. Turner, took advantage of the delicately graded pictorial possibilities for their landscape studies.

In the 19th century, the French artists Théodore Géricault, Eugène Delacroix, and Constantin Guys still followed the character of the brush drawing, even though it was already being replaced by the variegated watercolour and gouache painting, a method of painting with opaque colours that have been ground in water and mingled with a preparation of gum. In modern drawing, the brush has regained some importance as an effective medium for contrasting planes and as carrier of the

theme; in this, the dry brush has proven itself a useful tool for the creation of a granular surface structure.

The combination of various techniques plays a greater role in drawing than in all other art forms. Yet it is necessary, in the numerous drawings in which two or more mediums are involved, to distinguish between those in which the mediums were changed in the course of artistic genesis and those in which an artistic effect based on a combination of mediums was intended from the beginning.

In the first case, one is confronted with a preliminary sketch, as it were, of the eventual drawing: the basic structure with some variations is tried out in charcoal, chalk, metalpoint, pencil, or some other (preferably dry and easily corrected) material and then carried out in a stronger and more durable medium. Most pen drawings are thus superimposed on a preliminary sketch. The different materials actually represent two separate stages of the same artistic process.

More relevant artistically is the planned combination of different techniques that are meant to complement each other. The most significant combination from the stylistic point of view is that of pen and brush, with the pen delineating the contours that denote the object and the brush providing spatial and plastic as well as pictorial—that is, colour—values. The simplest combined form is manuscript illumination, where the delineated close contours are filled in with colour. The drawing may actually be improved if this is done by a hand other than the draftsman's or at a later time.

Pen-and-brush combination

More important is brushwork that supplements linear drawing, in which entire segments may be given over to one technique or the other; for example, the considerable use of white (which is hard to apply with the pen) in drawings on tinted paper. In similar complementary fashion the brush may be used for plastic modelling as a way of highlighting, that is indicating the spots that receive the greatest illumination. The technique of combined pen-and-brush drawing was favored by the draftsmen of Germany and the Netherlands, especially in the circle around Dürer and the south German Danube School. Shadows, too, can be inserted in a drawing with dark paint. The illusion of depth can also be achieved with white and dark colours in a pure chalk technique.

In contrast to these methods, which still belong to a linear system of drawing, is the flat differentiation of individual segments of a work in (usually) the same medium: wash. Various bodies and objects are evenly tinted with the brush within or along the drawn contours. Planes are thus contrasted with lines, enhancing the illusionary effect of plasticity, space, and light and shadow. This modelling wash has been used again and again since the 16th century, sometimes in combination with charcoal, chalk, or pencil drawings. A further refinement, used particularly in landscape drawings, is wash in varying intensities; additional shadings in the sense of atmospheric phenomena, such as striking light and haze merging into fog and cloud, can be rendered through thinning of the colour or repeated covering over a particular spot. A chromatic element entered drawing with the introduction of diluted indigo, known in the Netherlands from the East India trade; it is not tied to objects but used in spatial and illusionist fashion, by Paul Brill and Hans Bol in the 16th and 17th centuries, for example. The mutual supplementation and correlation of pen and brush in the wash technique was developed most broadly and consistently in the 17th century, in which the scaffold, so to speak, of the pen drawing became lighter and more open, and brushwork integrated corporeal and spatial zones. The transition from one technique to the other—from wash pen drawings to brush drawings with pen accents—took place without a break. Claude Lorrain and Nicolas Poussin in 16th- and 17th-century France are major representatives of the latter technique, and Rembrandt once again utilized all its possibilities to the full.

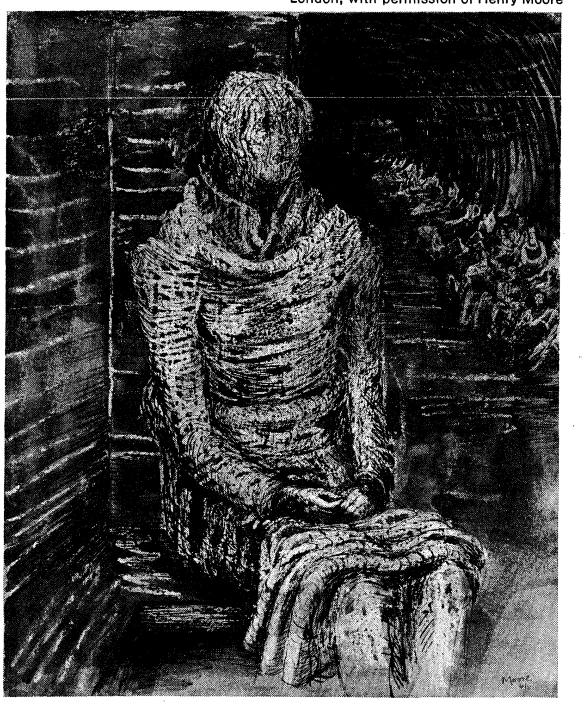
Whereas this method served—within the general stylistic intentions of the 17th century—primarily to elucidate spatial and corporeal proportions, the artists of the 18th century employed it to probe this situation visually with the aid of light. The unmarked area, the spot left empty, has as much representational meaning as the pen contours, the lighter or darker brush accent, and the tinted area.

The art of omission plays a still greater role, if possible, in the later 19th century and in the 20th. Paul Cézanne's late sheets, with their sparse use of the pencil and the carefully measured out colour nuances, may be considered the epitome of this technique. As the colouring becomes increasingly varied through the use of watercolours to supplement a pen or metalpoint drawing, one leaves the concept of drawing in the strict sense of the term. According to the quality and quantity of the mediums employed, one then speaks of "drawings with watercolour," "watercolourized drawings," and "watercolours on preliminary drawings." The predominant stroke character, rather than the fact that paper is the carrier, is the chief feature when deciding whether or not the work may legitimately be called a drawing.

Combination of dry and fluid drawing mediums

The combination of dry and fluid drawing mediums provides a genuine surface contrast that may be exploited for sensuous differentiation. Here again a distinction must be made between various ways of applying the identical medium—for example, charcoal and charcoal dust in a water solution or, more frequently, sanguine and sanguine rubbed in with a wet brush—and the stronger contrast brought about by the use of altogether different mediums. Chalk drawings are frequently washed with bister or watercolour, after the principle of the washed pen drawing. Stronger contrasts, however, can be obtained if the differing techniques are employed graphically, as the Flemish draftsmen of the 17th century liked to do. The Chinese ink wash of chalk drawings also contributes to the illusion of spatial depth. Along with such Dutch painters as Jan van Goyen and members of the family van de Velde, Claude Lorrain achieved great mastery in this technique. The differentiated treatment of the foreground with pen and brush and the background with chalk renders spatial depth plausible and plastic. In modern art, the use of different mediums—whether for plastic differentiation, such as Henry Moore carried out with unequalled mastery in his "Shelter Drawings," or only for the purpose of con-

By courtesy of the trustees of the Tate Gallery, London, with permission of Henry Moore



Use of different mediums to emphasize sculptural effects, "Women Seated in the Underground," crayon and wash drawing by Henry Moore, 1941. In the Tate Gallery, London. 48.3 ×38.1 cm.

trasting varied surface stimuli of nonrepresentational compositions, as well as the enrichment with colours and even with collage elements (the addition of paper, metal, or other actual objects) broadens the concept of the drawing so that it becomes an autonomous picture the mixed technique of which transcends the borderline between drawing and painting.

Mechanical devices. Mechanical aids are far less important for art drawing than for any other art form. Many draftsmen reject them altogether as unartistic and inimical to the creative aspect of drawing.

Apart from the crucial importance that mechanical aids have had and continue to have for all kinds of construction diagrams, plans, and other applied drawings, some mechanical aids have been used in varying but significant measure for artistic drawings. The ruler, triangle, and compass as basic geometric instruments have played a major role especially in periods in which artists created in a consciously constructionist and perspectivist manner. Marks for perspective constructions may be seen in many drawings of early and High Renaissance vintage.

For perspectively correct rendition, the graticulate frame, marked off in squares to facilitate proportionate enlargement or reduction, allowed the object to be drawn to be viewed in line with a screen on the drawing surface. Fixed points can be marked with relative ease on the resultant system of coordinates. For portrait drawings the glass board used into the 19th century, had contours and important interior reference points marked on it with grease crayons or soap sticks, so that they could be transferred onto paper by tracing or direct copying. Both processes are frequently used for preliminary sketches for engravings to be duplicated, as is the screened transmission of a preliminary sketch onto the engraving plate or, magnifying, the painting surface. In such cases the screen lies over the preparatory drawing.

Mirrors and mirror arrangements with reducing convex mirrors or concave lenses were likewise used (especially in the 17th and 18th centuries) as drawing aids in the preparation of reproductions. Even when it was a matter of the most exact rendition of topographical views, such apparatus, as well as the camera obscura (a darkened enclosure having an aperture usually provided with a lens through which light from external objects enters to form an image on the opposite surface), were frequently employed. In a darkened room the desired section is reflected through a lens onto a slanting mirror and from that inverse image is reflected again onto the horizontally positioned drawing surface. Lateral correction can be obtained by means of a second mirror.

Unless the proportions do not allow it, true-to-scale reducing or enlarging can also be carried out with the aid of the tracing instrument called the pantograph. When copying, the crayon or pencil inserted in the unequally long feet of the device reproduces the desired contours on the selected scale.

Most of these aids were thus used in normal studio practice and for the preparation of certain applied drawings. Equally practical, but useful only for closely circumscribed tasks, were elliptic compasses, curved rulers, and stencils, particularly for ornamental and decorative purposes. Only a few present-day artists use stencils or simple blocks with a given shape in larger scale composition, in order to obtain the effect of repetition, often in an arbitrary use, in "alienating" technique and colour.

Mechanically produced drawings

Mechanically produced drawings such as typewriter sketches, computer drawings, and oscillograms, all of which can bring forth unusual and attractive results, nevertheless do not belong to the topic because they lack the immediate creativity of the art drawing.

APPLIED DRAWINGS

Applied and technical drawings differ in principle from art drawings in that they record unequivocally an objective set of facts and on the whole disregard aesthetic considerations. The contrast to the art drawing is sharpest in the case of technical project drawings, the purpose of which is to convey not so much visual plausibility as to give exact information that makes possible the realization of an idea. Such plans for buildings, machines, and technical systems are not instantly readable because of the orthogonal (statically independent) projection, the division into separate planes of projection, and the use of symbols. Prepared as a rule with such technical aids as ruler and compass, they represent a specialized language of their own, which must be learned. For topographic (detailed delineation of the features of a place) and cartographic (map-making) drawings, too, a special terminology has developed that above all systematizes spatial representations, making them intelligible to the expert with the aid of emblems and symbols.

Equally far removed from any claim to artistic standing are most illustrations serving scientific purposes, the aim of which is to record as objectively as possible the characteristic and typical features of a given phenomenon. The systematic drawings, used especially in the natural sciences to explain a system or a function, resemble plans; descriptive and naturalistic illustrations, on the other hand, approach the illusionistic plausibility of visual experience and can attain an essentially artistic character. A good many artists have drawn scientific illustrations, and their works—the botanical and zoological drawings of the Swiss Merian family in the 17th and 18th centuries, for example—are today more esteemed for their artistic than for their documentary value.

Of a similarly ambivalent nature is the illustrative drawing that perhaps does not go beyond a simple pictorial rendition of a literary description but because of its specific formal execution may still satisfy the highest artistic demands. Great artists have again and again illustrated Bibles, prayerbooks, novels, and literature of all kinds. Some of the famous examples are Botticelli's illustrations for Dante's *Divine Comedy* and Dürer's marginal illustrations for the emperor Maximilian's prayerbook. Some artists have distinguished themselves more as illustrators than as autonomous draftsmen, as for example the 18th-century German engraver Daniel Chodowiecki, the 19th-century caricaturist Honoré Daumier, the 19th-century satiric artist Wilhelm Busch, and the 20th-century Austrian illustrator Alfred Kubin.

Clearly connected with illustrative drawing is caricature, which, by formally overemphasizing the characteristic traits of a person or situation, creates a suggestive picture that—precisely because of its distortion—engraves itself on the viewer's mind. This special kind of drawing was done by such great artists as Leonardo, Dürer, and the 17th-century artist Gian Lorenzo Bernini and by draftsmen who, often for purposes of social criticism, have devoted themselves wholly to caricaturing, such as the 18th-century Italian Pier Leone Ghezzi, the 19th-century Frenchman Grandville (professional name of Jean-Ignace-Isidore Gérard), and Daumier.

Modern cartoons

From such overdrawn types developed continuous picture stories that could dispense to a considerable extent with the explanatory text. Modern cartoons are based on these picture stories. Through the formally identical treatment of peculiar types, these drawings acquire an element of consecutiveness that, by telling a continuing story, adds a temporal dimension to two-dimensional drawing. This element is strongest in trick drawings that fix on paper, in brief segments of movement, invented creatures and phenomena that lack all logical plausibility; a rapid sequence of images (leafing through the pages, seeing it projected on the screen) turns the whole into apparent motion. The artistic achievement, if any, lies in the original invention; its actual realization is predetermined and sometimes carried out by a large and specialized staff of collaborators, often with the aid of stencils and traced designs. Moreover, since the final result is partially determined by the mechanical multiplication, an essential criterion of drawing—the unity of work and result—does not apply.

SUBJECT MATTER OF DRAWING

Anything in the visible or imagined universe may be the theme of a drawing. In practice, however, by far the greatest number of art drawings in the Western world deal with the human figure. This situation springs from



"Five Grotesque Heads," pen and ink drawing by Leonardo (1452–1519). In the Royal Library, Windsor Castle. 26 × 20.5 cm.
Copyright reserved

the close bond between drawing and painting: in sketches, studies, and compositions, drawing prepared the way for painting by providing preliminary clarification and some formal predetermination of the artist's concept of a given work. Many drawings now highly regarded as independent works were originally "bound," or "latent," in that they served the ends of painting or sculpture. Yet, so rounded, self-contained, and aesthetically satisfying are these drawings that their erstwhile role as handmaidens to the other pictorial arts can be reconstructed only from knowledge of the completed work, not from the drawing itself. This situation is especially true of a pictorial theme that acquired, at a relatively early stage, an autonomous rank in drawing itself: the portrait.

Portraits. Drawn 15th-century portraits—by Pisanello or Jan van Eyck, for example—may be considered completed pictorial works in their concentration, execution, and distribution of space. The clear, delicately delineated representation follows every detail of the surface, striving for realism. The profile, rich in detail, is preferred; resembling relief, it is akin to the medallion. Next in prominence to the pure profile, the three-quarter profile, with its more spatial effect, came to the fore, to remain for centuries the classic portrait stance.

The close relationship to painting applies to practically all portrait drawings of the 15th century. Even so forceful a work as Dürer's drawing of the emperor Maximilian originated as a portrait study for a painting. At the same time, however, some of Dürer's portrait drawings clearly embody the final stage of an artistic enterprise, an ambivalence that can also be observed in other 16th-century portraitists. The works of Jean and François Clouet in France and of the younger Hans Holbein in Switzerland and even more markedly in England in the same century bestowed an autonomy on portrait drawing especially when a drawing is completed in chalk of various colours. The choice of the softer medium, the contouring, which for all its exactitude is less severely self-contained, and the more delicate interior drawing with plane elements gives these drawings a livelier, more personal character and accentuates once more their proximity to painting.

Chalk and pastel portraits

In polychromatic chalk technique and pastel, portrait drawing maintained its independence into the 19th century. In the 18th century, Quentin de La Tour, François Boucher, and Jean-Baptiste Chardin—all of these artists from France—were among its chief practitioners, and even Ingres, living in the 19th century, still used its technique. In pastel painting, the portrait outweighed all other subjects.

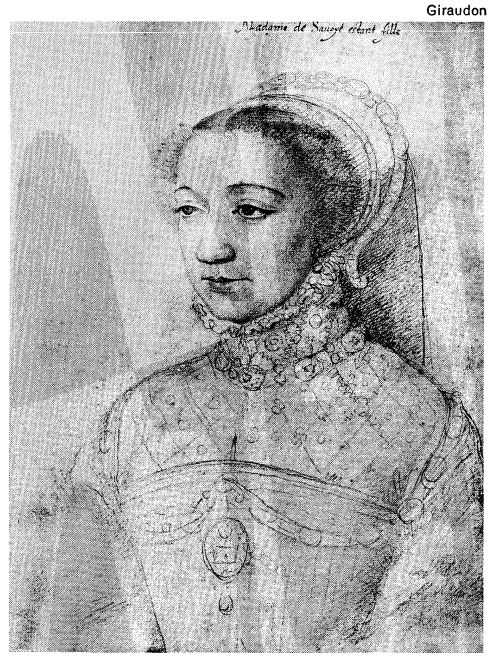
In the choice of pose, type, and execution, portrait painting, like other art forms, is influenced by the general stylistic features of an epoch. Thus, the extreme pictorial attitude of the late Baroque and Rococo was followed by a severer conception during Neoclassicism, which preferred monochrome techniques and cultivated as well the special form of the silhouette, a profile contour drawing with the area filled in in black. Unmistakably indebted to their 15th-century predecessors, the creators of portrait drawings of the early 19th century aimed once more at the exact rendition of detail and plastic effects gained through the most carefully chosen graphic mediums: the thin, hard pencil was their favorite instrument, and the silverpoint, too, was rediscovered by the Romantics.

More interested in the psychological aspects of portraiture, late 19th- and 20th-century draftsmen prefer the softer crayons that readily follow every artistic impulse. The seizing of characteristic elements and an adequate plane rendition weigh more heavily with them than realistic detail. Mood elements, intellectual tension, and personal engagement are typical features of the modern portrait and thus also of modern portrait drawing, an art that continues to document the artist's personal craftsmanship beyond the characteristics of various techniques.

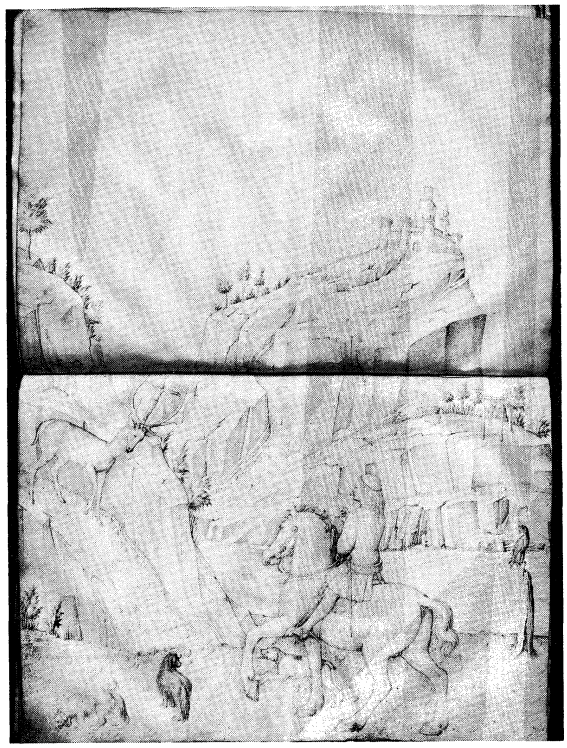
Landscapes. As early as the 15th century, landscape drawings, too, attained enough autonomy so that it is hard to distinguish between the finished study for the background of a particular painting and an independent, self-contained sketched landscape. Already in Jacopo Bellini's 15th-century sketchbooks (preserved in albums in the British Museum and the Louvre), there is an intimate connection between nature study and pictorial structure; in Titian's studio in the 16th century, landscape sketches must have been displayed as suggestions for pictorial backgrounds.

But it was Dürer who developed landscape as a recollected image and autonomous work of art, in short, as a theme of its own without reference to other works. His watercolours above all but also the drawings of his two Italian journeys, of the surroundings of Nürnberg and of the journey to the Netherlands, represent the earliest

Giraudon



"Portrait of Marguerite de Valois," chalk drawing by François Clouet, *c.* 1559. In the Musée Condé, Chantilly, France. 30.1 × 21.1 cm.



"St. Hubert," two pages from a sketchbook by Jacopo Bellini (c. 1400–c. 1470), pen and ink over chalk or lead point. In the British Museum. 67.2 × 41.5 cm.
Giraudon

pure landscape drawings. Centuries had to pass before such drawings occurred again in this absolute formulation.

Landscape elements were also very significant in 16th-century German and Dutch drawings and illustrations. The figurative representation, still extant in most cases, is formally quite integrated into the romantic forest-and-meadow landscape, particularly in the works of the Danube School—Albrecht Altdorfer and Wolf Huber, for example. More frequently than in other schools, one finds here carefully executed nature views. In the Netherlands, Pieter Bruegel drew topographical views as well as free landscape compositions, in both cases as autonomous works.

17th-century landscape drawing

In the 17th century, the nature study and the landscape drawing that grew out of it reached a new high. The landscape drawings of the Accademia degli Incamminati (those of Domenichino, for example) combined classical and mythological themes with heroic landscapes. The Frenchman Claude Lorrain, living in Rome, frequently worked under the open sky, creating landscape drawings with a hitherto unattained atmospheric quality. This type of cultivated and idealized landscape, depicted also by Poussin and other Northerners residing in Rome (they were called Dutch Romanists), is in contrast with the unheroic, close-to-nature concept of landscape held primarily by the Netherlanders. All landscape painters—a specialty that was strongly represented in the artistically specialized Low Countries—also created independent landscape drawings (Jan van Goyen and Jacob van Ruisdael and his uncle and cousin, for example), with Rembrandt again occupying a special position: capturing the characteristics of a region often with only a few strokes, he enhanced them in such manner that they acquire monumental expressive power even in the smallest format. In 18th-century Italy, the topographically faithful landscape drawing gained in importance with the advent of the *Vedutisti*, the purveyors of "views," forming a group by themselves (among them, Giambattista Piranesi and Canaletto [Giovanni Antonio Canal]) and often working with such optical aids as the graticulate frame and camera obscura. Landscape drawings of greater artistic freedom, as well as imaginary landscapes, were done most successfully by some French artists, among them Hubert Robert; pictorially and atmospherically, these themes reached a second flowering in the brush-drawn landscapes of such English artists as Turner and Alexander Cozens, whose influence extends well into the 20th century.

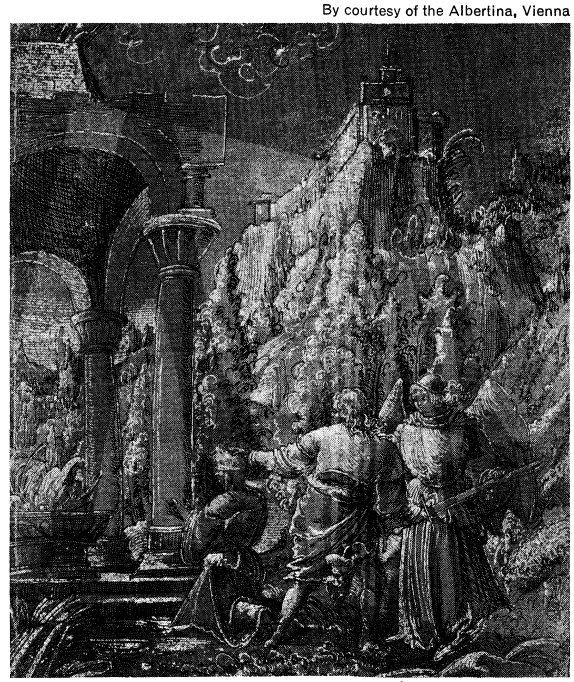
Given their strong interest in delineation, the 18th-century draftsmen of Neoclassicism and, even more, of Romanticism observed nature with topographical accuracy. As a new "discovery," the romantically and heroically exaggerated Alpine world now took its place in the artist's mind alongside the arcadian view of the Italian landscape.

Landscape drawings and even more, watercolours, formed an inexhaustible theme in the 19th century. The French artist Jean-Baptiste-Camille Corot and, toward the end of the century, Cézanne and van Gogh, were among the chief creators of landscape drawings. While landscapes form part of the work of many 20th-century draftsmen, the genre as such takes second place to general problems of form, in which the subject is merely treated as starting point.

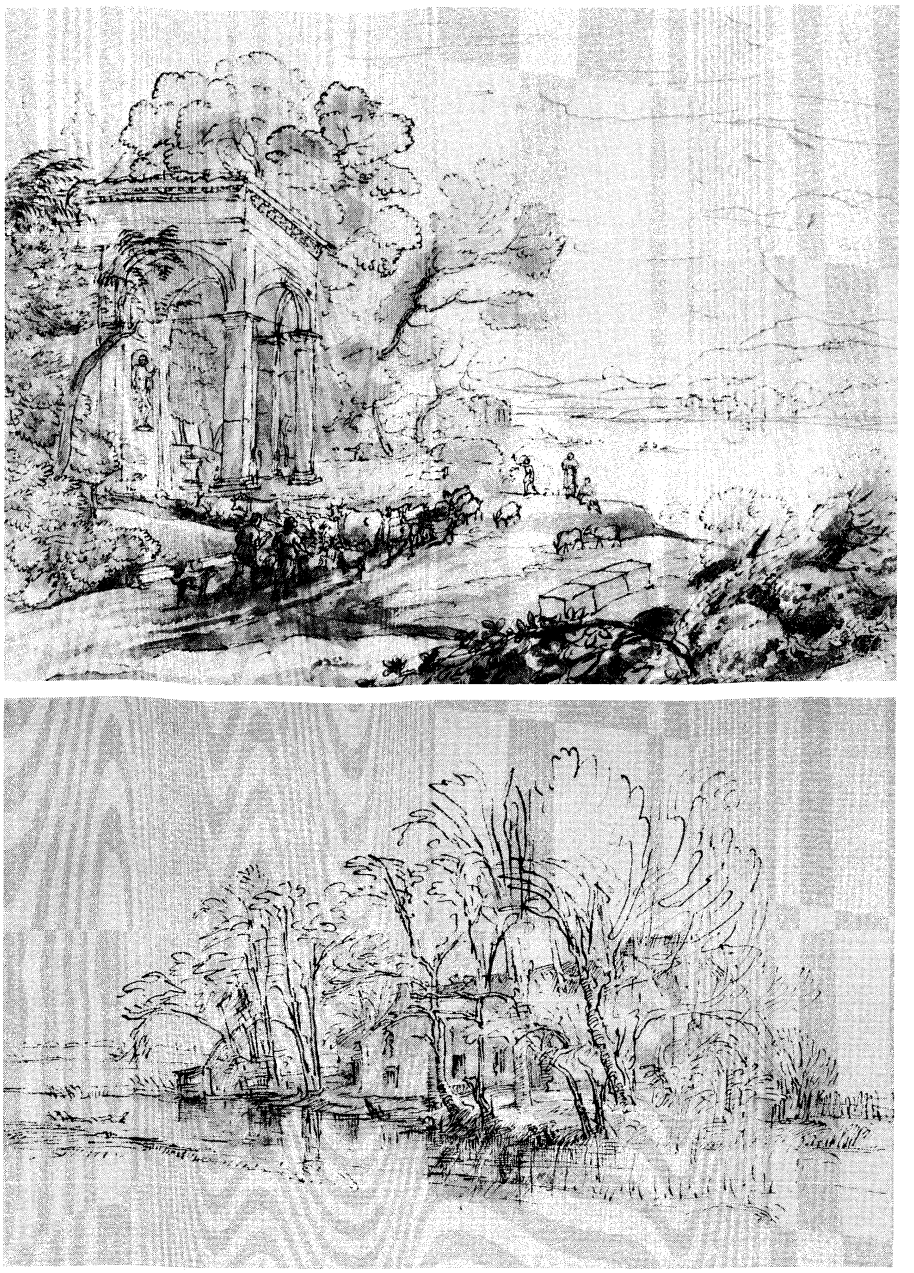
Figure compositions and still lifes. Compared to the main themes of autonomous drawing—portraiture and landscape—all others are of lesser importance. Figure compositions depend greatly on the painting of their time and are often directly connected with it. There were, to be sure, artists who dealt in their drawings with the themes of monumental painting, such as the 17th-century engraver and etcher Raymond de La Fage; in general, however, the artistic goal of figure composition is the picture, with the drawing representing but a useful aid and a way station. Genre scenes, especially popular in the 17th-century Low Countries (as done by Adriaen Brouwer, Adriaen von Ostade, and Jan Steen, for example) and in 18th-century France and England, did attain some independent standing. In the 19th century, too, there were drawings that told stories of everyday life; often illustrative in character, they may be called "small pictures," not only on account of the frequently multicoloured format but also in their artistic execution.

Still lifes can also lay claim to being autonomous drawings, especially the representations of flowers, such as those of the Dutch artist Jan van Huysum, which have been popular ever since the 17th century. Here, again, it is true that a well-designed arrangement transforms an

By courtesy of the Albertina, Vienna



"The Sacrifice of Isaac," by Albrecht Altdorfer (1480?–1538), pen and ink, heightened with white, on gray-green paper. In the Albertina, Vienna. 20.8 × 17.5 cm.



Contrasting approaches to the representation of landscape.
(Top) "Pastoral Landscape," by Claude Lorrain, pen with brown and gray brown wash, from the *Liber Veritatis* (78), 1644. In the British Museum. 18.5 × 26 cm. (Bottom) "House amid Trees on the Bank of a River," by Rembrandt (1606–69), pen and black ink, India ink wash, on brown coloured paper. In the British Museum. 40.6 × 59.2 cm.
By courtesy of the trustees of the British Museum; photographs, J.R. Freeman & Co. Ltd.

immediate nature study into a pictorial composition. In some of these compositions the similarity to painting is very strong; the pastels of the 19th- and 20th-century artist Odilon Redon, for instance, or the work of the 20th-century German Expressionist Emil Nolde, with its chromatic intensity, transcend altogether the dividing line between drawing and painting. In still lifes, as in landscapes, autonomous principles of form are more important to modern artists than the factual statement.

Fanciful and nonrepresentational drawings. Drawings with imaginary and fanciful themes are more independent of external reality. Dream apparitions, metamorphoses, and the entwining of separate levels and regions of reality have been traditional themes in drawing. Consider the late 15th-century phantasmagoric works of Hieronymus Bosch as an early example. Or there are allegorical peasant scenes by the 16th-century Flemish artist Pieter Bruegel along with the carnival etchings of the 17th-century French artist Jacques Callot. Other artists whose works illustrate what can be done with drawing outside of landscape and portraiture are: the 18th-century Italian engraver Giambattista Piranesi, the 18th- and 19th-century Anglo-Swiss artist Henry Fuseli, the 19th-century English illustrator Walter Crane, the 19th-century French symbolist artist Gustav Moreau, and the Surrealists of the 20th century.

Nonrepresentational art

Nonrepresentational art, with its reduction of the basic elements of drawing—point, line, plane—to pure form, offered new challenges to drawing. Through renunciation of associative corporeal and spatial relationships, the unfolding of the actual dimensions of drawing and the structure of the various mediums acquire new significance. The graphic qualities of the line in the plane, the unmarked area had already been emphasized in earlier times—for example, in the *grotteschi* of Raphael in the 16th century (the fanciful or fantastic representations of human and animal forms often combined with each other and interwoven with representations of fo-

liage, flowers, fruit, or the like) and in calligraphic exercises such as moresques (strongly stylized linear ornament, based on leaves and blossoms)—but mostly as printing or engraving models for the most disparate decorative tasks (interior decoration, furniture, utensils, jewelry, weapons, and the like).

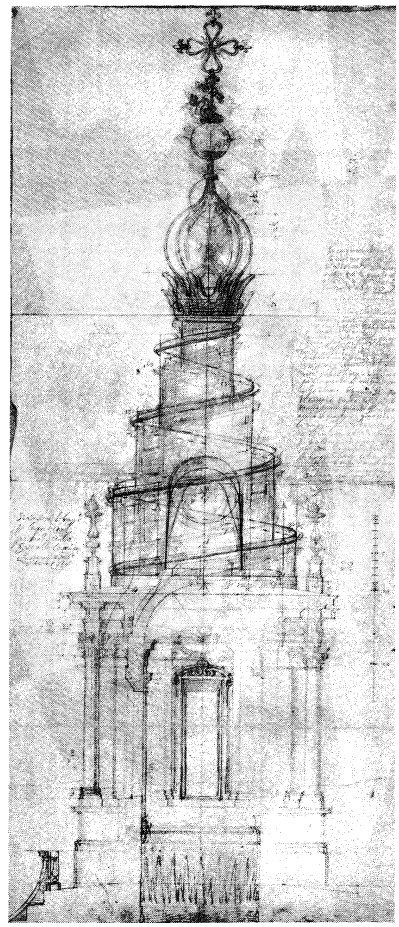
Artistic architectural drawings. There is one field in which drawing fulfills a distinct function: artistic architectural drawings are a final product in their form and quality as drawings, differing from the impersonal, exact plans and designs by precisely the same "handwriting" character that typifies art drawings. In many cases, no actual execution of these plans was envisaged; since the early Renaissance, such ideal plans have been drawn to symbolize, in execution and accessories, an abstract content. Despite the often considerable exactitude with which the plans are drawn, the personal statement predominates in the flow of the line. This personal note clearly identifies the drawings of such artists and architects as Albrecht Altdorfer, Leonardo, Michelangelo, Bernini, Francesco Borromini, and Piranesi. Also distinct from the ground-plan type of architectural drawing are the art drawings of autonomous character created by such 20th-century architects as Erich Mendelsohn and Le Corbusier.

II. History of drawing

WESTERN

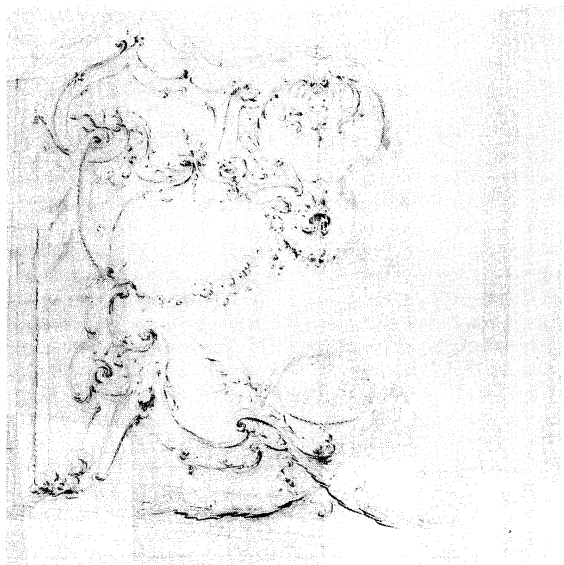
As an artistic endeavour, drawing is almost as old as mankind. In an instrumental, subordinate role, it developed along with the other arts in antiquity and the Middle Ages. Whether preliminary sketches for mosaics and murals or architectural drawings and designs for statues and reliefs within the variegated artistic production of the Gothic medieval building and artistic workshop, drawing as a nonautonomous auxiliary skill was subordinate to the other arts. Only in a very limited sense can one speak of centres of drawing in the early and High Middle Ages; that is, the scriptoria of the monasteries of Corbie and Reims in France, as well as those of Canterbury and Winchester in England, and also a few places in southern Germany, where various strongly delineatory (graphically illustrated) styles of book illumination were cultivated.

14th, 15th, and 16th centuries. In the West, the history of drawing as independent artistic document began toward the end of the 14th century. If its development was independent, however, it was not insular. Just as the greatest draftsmen have been for the most part also distinguished painters, illustrators, sculptors, or architects, so the centres and the high points of drawing have generally coincided with the leading localities and the major epochs of the other arts. Moreover, the same stylistic phenomena have been expressed in drawing as in other art forms. Indeed, drawing shares with other art forms the characteristics of individual style, period style, and regional features. Drawing differs, however, in that it interprets and renders these characteristics in terms of its own unique mediums.



"S. Ivo della Sapienza," architectural drawing by Francesco Borromini, *c.* 1642–60. In the Albertina, Vienna. 41 × 26.7 cm.
By courtesy of the Albertina, Vienna

By courtesy of the Pierpont Morgan Library, New York



"Design for a Wall Panel," by Giambattista Piranesi (1720–78), pen with brown ink, brown wash, over black chalk. In the Pierpont Morgan Library, New York. 28.8 × 28.2 cm.

Drawing as an independent art form

Drawing became an independent art form in northern Italy, at first quite within the framework of ordinary studio activity. But with nature studies, copies of antiques, and drafts in the various sketchbooks (those of Giovannino de'Grassi, Antonio Pisanello, and Jacopo Bellini, for example), the tradition of the Bauhütten studio workshop changed to individual work: the place of "exempla," models, reproduced in formalized fashion was now being taken by subjectively probing and partially creative drawings. In the early 15th century the international Soft Style of the period still largely predominated over the draftsman's individual "handwriting." At mid-century, however, the differentiation of drawing style according to region and the artist's personality set in. Essential criteria, destined to remain characteristic for generations, begin to strike the eye.

In drawing produced north of the Alps, the characteristic features lie in the tendency to pictorial compactness and precise execution of detail. Many painters produced individual drawings, but the most notable draftsmen are the otherwise unidentified 15th-century German Master of the Housebook and his contemporary Martin Schongauer. Both of these artists were also major copperplate engravers, so that it is not always easy to determine whether the work is a preliminary sketch or an independent drawing.

In Italian Renaissance drawings, of which there are a

great many, the diverging stylistic features of the various artistic regions were particularly evident. What they had in common was the overwhelming importance of the sketch and the study, in contrast to the far rarer finished drawings. The formal and thematic connection with painting is very close even when it was not a question of preliminary drawings. The draftsmen of Venice and northern Italy preferred an open form with loose and interrupted delineation in order to achieve even in drawing the pictorial effect that corresponded to their painters' imagination.

In central Italy, on the other hand, and especially in Florence, it was the clear contour that predominated, the closed and firmly circumscribed form, the static and plastic character. Corresponding to the functional purpose of drawing, the individual artists' studios (which, as was the case with the Medicis' Academy of St. Mark, also had to engage in general educational and humanistic investigations) formed the most significant centres of art drawing. In these large studios, drawing served not only for the probing realization of creative ideas, it was not only study and mediator between the conception and the master's finished work; it functioned also as teaching aid for the assistants who worked with the master and as a vehicle for the formation and preservation of an individual workshop tradition. Although Leonardo's scientific interests were expressed in a large number of drawings, his ideal concept of the human figure is much more frequently preserved in the drawings of his collaborators and successors than in his own. Raphael and Michelangelo were also outstanding draftsmen. Each of them used drawing in order to allow his thoughts about individual works to mature; each had a highly personal drawing style, the one with a soft and rounded stroke, the other with a sculptor's intermittent and powerful stroke. Probably a great deal of drawing was done in Raphael's studio, especially if only for the preparation of the engravings after Raphael's compositions. From Michelangelo's hand came the first so-called connoisseur drawings that are esteemed as a personal document. They are the precursors of the collector's drawings that began in the later 16th century (autonomous works, destined for collections).

Northern European drawing

North of the Alps the autonomy of drawing was championed in the first instance by Dürer, an indefatigable draftsman who mastered all techniques and exercised an enduring and widespread influence. The delineatory constituent clearly predominates even in his paintings. This corresponds to the general stylistic character of 16th-century German art, within which Matthias Grünewald, with his freer, broader, and therefore more pictorial style of drawing, and the painters of the Danube school, with their ornamentalizing and agitated stroke, represent significant exceptions. In their metamorphozing of the perceived reality into drawings, the landscapes of Altdorfer and Wolf Huber in particular are astonishing documents of a feeling for nature that might almost be called Romantic.

Soberer, incredibly compact in their pictorial concept and yet akin to the Renaissance in their objective viewing, were the portrait drawings of Hans Holbein, the Younger, whose sojourns in 16th-century England proved stimulating to other artists as well. Similar, if less personal than Holbein because of the stricter linearity of their work, were the drawings of the French portraitists Jean and François Clouet. In the Low Countries, where they were combined with the idealized image of Italy (as in the drawings of Lucas van Leyden), Dürer's methods gained lasting popularity in the landscape drawings and studies "after life" by Pieter Bruegel the Elder.

Drawing acquired a pivotal significance in the period of Mannerism (c. 1525–1600), both as a document of artistic invention and as a means of its realization. Jacopo Pontormo in Florence, Parmigianino in northern Italy, and Tintoretto in Venice used point and pen as essential and spontaneous vehicles of expression. Their drawings were clearly related to their painting, both in content and in the graphic method of sensitive contouring and daringly drawn foreshortening.

17th, 18th, and 19th centuries. In the early 17th century, Jacques Callot rose to prominence in French art: gifted as a draftsman above all, he recorded with the pen his clever inventions and great picture stories, primarily in bold abbreviations.

The importance of drawing for an artist's growth and the widening of his horizon is attested also by the work of Peter Paul Rubens, whose studies and sketches make up an integral part of his creative achievement. In order to disseminate his pictorial themes and concept of form, he maintained his own school for draftsmen and engravers. Among the circle of Flemings around him, Jacob Jordaens and Sir Anthony Van Dyck are notable as draftsmen with a style of their own.

Hercules Seghers was among the most fascinating artists of the 17th century, a creator of drawn and etched landscapes that he continued to rework while experimenting with printing processes. From the point of view of technique and form, he was important for the greatest artist of Holland, Rembrandt. Seghers combined great inventiveness, especially in his interpretations of Old Testament motifs, and broad mastery of all the techniques of drawing. In his studio, too, drawing was emphasized as a teaching aid and a means of formal experimentation.

Most Dutch painters of the 17th century, such as the van de Velde family, Brouwer, van Ostade, Pieter Saenredam, and Paulus Potter, were also industrious draftsmen who recorded their special thematic concerns in drawings that were largely completed. Beyond serving as preparation for paintings, these were regarded as autonomous works representing the final stage of the creative process.

In 17th-century Italy, drawing by way of artistic practice and experimentation became established in the academies, especially in Bologna. More significant, however, was the continuing development of landscape drawing, as initiated by the brothers Agostino and Annibale Carracci and articulated further by Domenichino and Salvator Rosa. The Frenchman Claude Lorrain so developed the landscape drawing of the Roman countryside that it became almost a genre of its own; in his works, which were often intended for sale, nature study and an idealized pictorial concept are uniquely merged. In detailed studies directly before the object, he achieved a timeless validity. Like Lorrain, Poussin also drew under the open sky. Using various techniques, he combined realistic experiences and humanistic concepts in idealizing compositions the figures and scenes of which are harmoniously integrated into a spacious landscape. This open-air painting and drawing was practiced also by some other artists who spent a considerable time in Rome—the Dutch artists Jan Asselijn, Nicholaes Berchem, Karel Dujardin, and Adam Pijnacker, for example. For most southern European artists of the 17th century, however, drawing was a mere stage in the creation of a painting.

17th-century Italian landscape drawing

Antoine Watteau, too, did drawings to "keep his hand in" for his painting, although he did so with an independence that led him far beyond the immediate occasion. Most figures in the paintings from various periods of his career were based on earlier drawings. In the grand scale of his form and the attention paid to pictorial elements, he carried on in the manner of Rubens, combining it with the light esprit of the 18th century. The leading position of French art in the first half of that century was confirmed by the achievements of Boucher, Fragonard, Hubert Robert, and Gabriel de Saint-Aubin, whose drawings include figure studies, genre-like works, and landscapes.

In contrast to the French draftsmen who brought about a flowering of the *à trois crayons* method on tinted paper, some artists created similar landscapes with pen and brush but with greater objective abbreviation. Mention must here be made of Venice, with the Giovanni Battista Tiepolo family, whose expansively conceived pen drawings, washed with a broad brush, call forth the kind of luminaristic effect that Francesco Guardi also used for landscape studies and imaginary scenes. These had been preceded by Canaletto's views of Venice, composed more

severely as far as tectonic (constructional) detail is concerned but nonetheless the first examples of this form of the landscape capriccio, or fantasy. The architect Giambattista Piranesi made his name primarily as a draftsman who recorded views of Rome; above all, in his drawings of architecture and eerie vaults ("Carceri"), he left behind a body of work of great intellectual and formal forcefulness.

The Spanish painter Goya, at the very end of the 18th and in the beginning of the 19th century, was in advance of his time in the way in which he handled his themes. Forming an odd contrast to the court-painter's pictures, his brush-and-sanguine drawings are rather more closely tied to his cycles of etchings. He combined the luminaristic effects of Tiepolo's drawings, with the dramatic impact of a Rembrandt chiaroscuro.

Also at the turn of the 19th century is an artist whose main work was that of a draftsman: the English caricaturist and social satirist Thomas Rowlandson, who produced colourful and distinctive watercolours. The late 18th and, even more, the early 19th century produced a drawing style that, in accordance with both the Neoclassical and the Romantic ideal, emphasized once more the linear element. In Ingres, idealistic Neoclassicism found an exemplary expression of strict linearity, and the pencil drawing became a downright classical form. The Nazarenes and Romantics in Rome and the Alpine region (Joseph Anton Koch, the brothers Friedrich and Ferdinand Olivier and Julius Schnorr von Carolsfeld) as well as those in north Germany (Philipp Otto Runge and Caspar David Friedrich) were more lyrical but equally rigorous in the use of the hardpoint; after a long time, they were the first northern artists to have made a significant contribution to the history of drawing. Among 19th-century artists, the emphasis on delineation was characteristic also of Moritz von Schwind in Germany and John Millais in England. (In the Neoclassical phase of the 20th century it was renewed, in a more open and "handwriting" fashion, by Thomas Eakins in the United States as well as by Picasso, Matisse, and Amedeo Modigliani in France.) The drawings of Delacroix, while preserving plastic qualities, show a broader stroke and are thus more pictorial. Daumier, active in all mediums primarily as a draftsman, utilized pictorial chiaroscuro effects in forcible statements of social criticism.

19th-century France

France continued to be a leading centre of the art of drawing, a form that was given a very personal note in each case in the works of Degas, Toulouse-Lautrec, van Gogh, and Cézanne. The line—the common point of departure for all of the above-mentioned artists—did not disappear until Seurat's plane shading, done in the pointillist manner.

Modern. Except for a few stylistic currents such as Tachisme (paintings consisting of irregular blobs of colour), drawing is represented in the work of practically all 20th-century artists; it is as international as modern art itself. As the other arts have become nonrepresentational, thus attaining autonomy and formal independence in relation to external reality, drawing is more than ever considered an autonomous work of art, independent of the other arts; and the sketch, study, and project—that is, the drawing as a stage in the genesis of works of sculpture, painting, and architecture—have greatly diminished in importance. Some schools and individual artists as well have concentrated on drawing and in very individualistic ways. The German Expressionists, for instance, developed especially emphatic forms of drawing with powerful delineation and forcible and hyperbolic formal description; notable examples are the works of Ernst Barlach, Käthe Kollwitz, Alfred Kubin, Ernest Ludwig Kirchner, Karl Schmidt-Rottluff, Max Beckmann, and George Grosz. In the artists' group *Der Blaue Reiter* (The Blue Rider), Wassily Kandinsky was foremost in laying the groundwork for a new evaluation of the nonrepresentational line. Paul Klee's lyrically sensitive drawings, carried out in a pen technique of unheard-of sublimity, represent a high point of modern drawing. In France, drawing plays a major role, especially in the work of the painters of the École de Paris (School of Paris), such as Pierre Soulages and Hans Hartung, who consider the line, the framework of lines, and the network of lines, as primary manifestations of form. Wols (Alfred Otto Wolfgang Schulze) and also the English artist Graham Sutherland may actually be called spiritual draftsmen who put their faith in the magic of the line. Finally, drawing occupies a considerable place in the work (including all its variants of style and form) of Picasso, once again a man who knew how to make use of its manifold technical possibilities. One is surely justified in calling him the greatest draftsman of the 20th century and one of the greatest in the history of drawing.

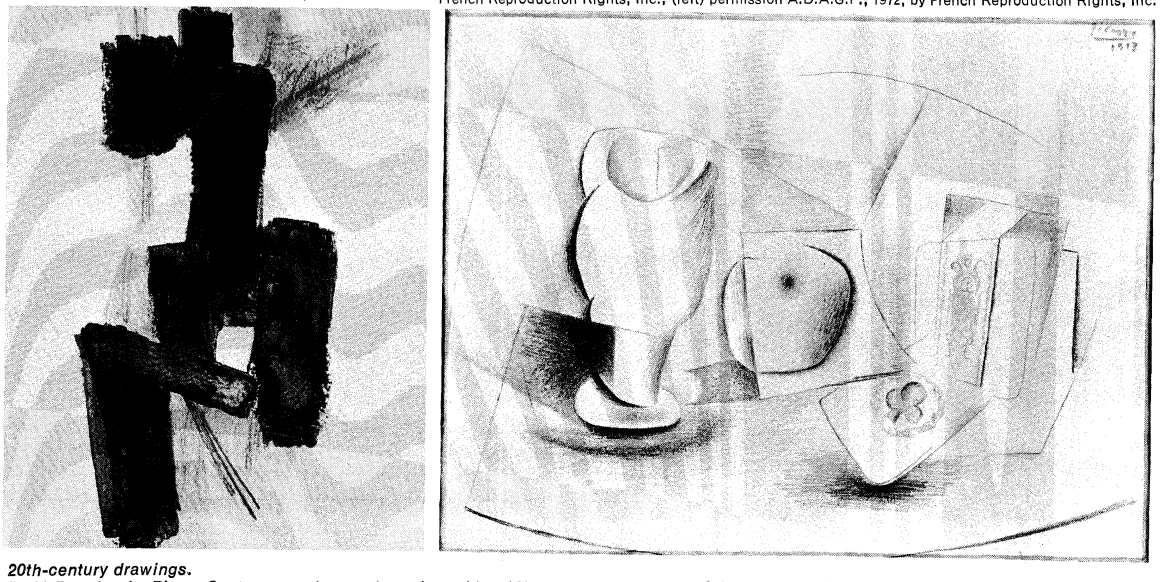
(H.R.H.)

EASTERN

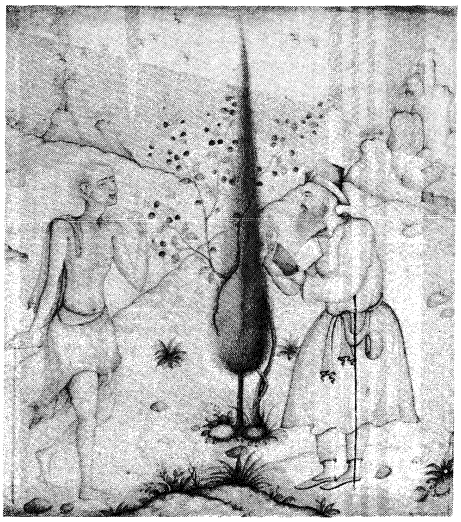
Drawings of China

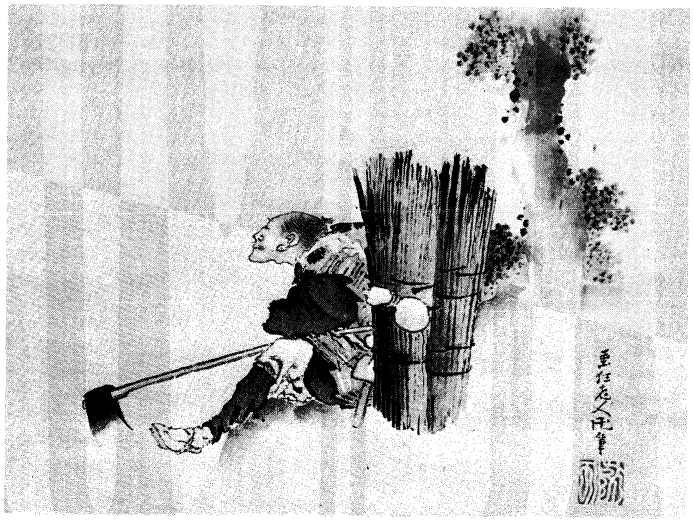
Some form of monochromatic brush drawing with ink may have been practiced in China as early as the 2nd

By courtesy of (right) Mrs. Barnett Malbin, Birmingham, Michigan (The Lydia and Harry Lewis Winston Collection); (left) Pierre Soulages; photographs, (right) Joseph Klima, Jr., permission S.P.A.D.E.M., 1972, by French Reproduction Rights, Inc., (left) permission A.D.A.G.P., 1972, by French Reproduction Rights, Inc.



20th-century drawings.
(Left) Drawing by Pierre Soulages, walnut stain and graphite, 1950. In the collection of the artist. 65 × 50 cm. (Right) "Still Life with Glass, Apple, Playing Card, and Package of Tobacco," pencil drawing by Pablo Picasso, 1913. In the Lydia and Harry Lewis Winston Collection, Birmingham, Michigan. 23.8 × 31.0 cm.





Eastern drawing.
(Left) "The Merchant and the Ascetic," Mughal drawing with colours, early 17th century. In the Smithsonian Institution, Freer Gallery of Art, Washington, D.C. 10.3 × 9.3 cm. (Right) "Woodcutter Gazing at Waterfall" (detail), by Hokusai, ink and colour on paper scroll, 1798. In the Stanford University Museum of Art, California. 29.7 × 39.0 cm.
By courtesy of (left) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C., (right) the Stanford University Museum of Art, California, Ikeda Collection

millennium BC; but the earliest pictorial work is in lacquer or on bronze vessels, contemporaneous with Alexander the Great (ruled 336–323 BC). It relies on contour and silhouette, with men and animals depicted in horizontal registers (levels, one above the other) reminiscent of Egyptian and Mediterranean work. The extent of any mutual influence between East and West cannot yet be determined. Under the Later Han dynasty (AD 23–220) wall paintings, linear in character, were produced in fresco (wet plaster) and secco (dry). Only in the Northern Wei (386–534) and T'ang (618–907) dynasties did the true character of Chinese drawing on silk or paper emerge. In the 7th century, the characteristic albums (*ts'e-yeh*) of drawings appear.

No distinction was made between drawing and painting because all Chinese pictorial art was fundamentally graphic. The artist worked with the fine point of the brush on paper or silk laid horizontally on a table. Work in pure outline was called *pai-miao*; ink applied in splashes, *p'o-mo*. Colour was used sparingly or not at all. The final work was not made direct from nature.

Hindu and Buddhist paintings at Ajantā in India and also in Ceylon reveal the essential quality in all Indian art: emphasis on a flowing, rhythmic contour to express movement and gesture. Drawings on palm leaf of the 11th century are similarly based on the use of line to depict mythological scenes.

The 14th century saw the manufacture of paper, introduced from China, permitting the production of the vertical book. Despite the Muslim prohibition of human representation, books illustrated with drawings, sometimes with flat decorative colour, were produced at the Persian and Mughal courts but not for public display. The use of a precise and expressive line constituted the basis for Persian and Indian (both Mughal and Rājput) miniature paintings, which show people in landscape or in relation to buildings.

Japanese art tended to follow that of China until the early 19th century, when the popular colour print was introduced. In the graceful feminine gestures of Utamaro's work, the Oriental love of the flowing contour is manifest, allowing his line to vary in width and density. Hokusai's drawings of social life in a humorous, almost grotesque vein, invariably reveal his complete command of the expressive line. (Ed.)

BIBLIOGRAPHY. JOSEPH MEDER, *Die Handzeichnung*, 2nd ed. (1923), a voluminous book with instructively selected illustrations, remains the basic work on the history and techniques of drawing. Another treatment of the subject, more concise in every respect, is HEINRICH LEPORINI, *Die Künstlerzeichnung*, 2nd ed. (1955). ARTHUR E. POPHAM published an introduction to drawing in *A Handbook to the Drawings and Watercolours in the Department of Prints and Drawings of the British Museum* (1939), based on the ample materials held by the British Museum. CHARLES DE TOLNAY in *History and Technique of Old Master Drawings* (1943); and JAMES WATROUS in *The Craft of Old Master Drawings* (1957), deal, from different points of view, with the history and techniques of the old masters; while HERIBERT HUTTER in *Die Handzeichnung* (1966; Eng. trans., *Drawing: History and Technique*, 1968), stresses the artistic function of drawing and includes modern works. DANIEL M. MENDELOWITZ, *Drawing* (1966), provides a historical résumé, with reference to the artistic elements and technical means of drawing; in the supplement *A Study Guide* (1967), he offers practical instructions for drawing techniques and their application, as does R. BEVERLY HALE in *Drawing Lessons from the Great Masters* (1964). JAKOB ROSENBERG illustrates the possibilities of drawing in *Great Draughtsmen from Pisanello to Picasso* (1959), with samples from the works of eight great artists. *Great Drawings of All Time*, ed. by IRA MOSKOWITZ, 4 vol. (1962), contains a summary with comments by leading authorities. M.W. EVANS, *Medieval Drawings* (1969), is useful for the early history of the art of drawing; PAUL J. SACHS, *Modern Prints and Drawings: A Guide to a Better Understanding of Modern Draughtsmanship* (1954), for the most recent developments. H. BOECKHOFF and F. WINTER, *Das grosse Buch der Graphik* (1968), gives the history of the 24 best-known collections, with comments by the various curators and the basic catalog of each collection. The number of detailed investigations in regard to individual countries, periods, and artists is too large to be listed in this bibliography. LUIGI GRASSI, *Storia del disegno* (1947), is very valuable for the role of drawing with the historical theories of art, including the elucidation of the original sources for further study. Unsurpassed in method and fundamental for an intensive study of this subtle theme is BERNHARD DEGENHART'S essay "Zur Graphologie der Handzeichnung," in *Jahrbuch der Hertziana*, vol. 1 (1937).

(H.R.H.)

Dreams

The mysterious quality of dreams may be appreciated in part from the attitudes they have evoked during human history. While any effort toward classification must be subject to inadequacies, beliefs about dreams fall into various classifications depending on whether dreams are believed to be reflections of reality, sources of divination, curative experiences, waking states, or evidence of unconscious activity (the so-called Freudian dream).

DIVERSE VIEWS ON THE NATURE OF DREAMS

Dreams as reflecting reality. Contemporary philosophers continue to argue about the differences between reality and dreams. English philosopher Bertrand Russel (1872–1970) wrote "it is obviously possible that what

we call waking life may be only an unusual and persistent nightmare" and further stated that "I do not believe that I am now dreaming but I cannot prove I am not." Philosophers generally try to resolve the question by saying that so-called waking experience seems vivid and coherent. As French philosopher René Descartes (1596–1650) put it: ". . . memory can never connect our dreams one with the other or with the whole course of our lives as it unites events which happen to us while we are awake"; or, as Russell stated succinctly: "Certain uniformities are observed in waking life, while dreams seem quite erratic."

Dreams in various cultures

Members of many cultures have variously coped with this dilemma; for example, among the Eskimo of Hudson Bay and the Patani Malay people, it is believed that during sleep one's "soul" leaves his body to live in a special dreamworld. Believers often consider it dangerous to wake someone lest his "soul" be lost. On these grounds the Tajal people of Luzon, for example, severely punish for awakening a sleeping person. In other cultures, dream events are held to be identical with reality; thus a Macusi Indian of Guyana is reported to have become enraged at the European leader of an expedition when he dreamed that the leader had made him haul a canoe up dangerous cataracts. He awoke exhausted and could not be persuaded that the dream was not real. There is a tradition in Borneo that if a man dreams that his wife is an adulteress, her father must take her back. A Zulu man is said to have broken off a friendship after dreaming that the friend meant him harm. A Paraguayan Indian, reportedly having dreamed a missionary shot at him, attempted to kill the missionary.

In other instances, dream events are believed to demand fulfillment. Jesuit priests in the 1700s reported that among Iroquois Indians it was obligatory to carry out dreams as soon as possible; one Indian was said to have dreamed that ten friends dove into a hole in the ice of a lake and came up through another. When told of the dream, the friends duly enacted their roles in it, but unfortunately, only nine of them succeeded. After dreaming of something valuable, Kurdish people were immediately expected to take it, by force if necessary. Among some natives of Kamchatka a man need only dream of a girl's favour for her to owe him her sexual favours.

Such interpretations in which the dream is given a status of reality need not imply that the two are indistinguishable. In some instances, the dream may be differentiated from reality, but dreams are accorded a superior status to the banal activities of wakefulness.

Dreams as a source of divination. There is an ancient belief that dreams predict the future; the Chester Beatty Papyrus is a record of Egyptian dream interpretations dating from the 12th dynasty (1991–1786 BC). In the *Iliad*, a character named Agamemnon is visited in dream by a messenger of the god Zeus to prescribe his future actions. From India, a document called the Atharvaveda, attributed to the 5th century BC, contains a chapter on dream omens. A Babylonian dream guide was discovered in the ruins of the city of Nineveh among tablets from the library of the emperor Ashurbanipal (668–627 BC).

Dreams in the Bible

The Old Testament is rife with prophetic dreams, those of the pharaohs and of Joseph and Jacob being particularly striking. Among pre-Islāmic peoples dream divination so heavily influenced daily life that the practice was formally forbidden by Muḥammad (*c.* 570?–632), founder of the Muslim religion.

Ancient and religious literatures express most confidence about so-called message dreams. Characteristically, a god or some other respected figure appears to the dreamer (typically a king, a hero, or a priest) in time of crisis and states a message. Such reports are found on ancient Sumerian and Egyptian monuments; frequent examples appear in the Old and New testaments. Joseph Smith (1805–1844), the founder of the Mormon religion, said that an angel had directed him to the location of buried golden tablets that described American Indians as descendants of the tribes of Israel.

Not all dream prophecies are so readily accepted. In the epic *Odyssey*, for example, dreams are classed as false ("passing through the gate of Ivory") and as true ("passing the Gate of Horn"). Furthermore, prophetic meaning may be attributed to dream symbolism. In the Bible, Joseph interpreted sheaves of grain and the moon and stars as symbols of himself and of his brethren. In general, the social status of dream interpreters varies; in cultures for which dreams loom important, their interpretation frequently is a specialized occupation of such people as priests, elders, or medicine men.

Perhaps the most famous dream interpretation book is that of the Greek geographer, Artemidorus (*c.* 2nd century AD) and is called *Oneirocritica* (from the Greek *oneiros*, "a dream"). Dream books remain widely available today. They continue to enjoy profitable sales in Europe, Asia, the Americas, Africa, and elsewhere among people who follow them in affairs of the heart, in gambling, and in matters of health and work.

Dreams as curative. So-called prophetic dreams in the Middle-Eastern cultures of antiquity often were combined with other means of prophecy, such as animal sacrifice, and with efforts to heal the sick. In classical Greece, dreams became directly associated with healing; ailing people came to dream in oracular temples where priests and priestesses advised about the cures dreams were held to provide. Similar practices, known as dream incubation, are recorded for Babylon and Egypt. In a widespread cult, suffering petitioners came to at least 600 temples of the Greek god of medicine to perform rites or sacrifices in efforts to dream appropriately, sleeping in wait of the appearance of the god or his emissary to deliver a cure. Many stone monuments placed at the entrances of the temples survive to record dream cures.

Dreams as extensions of the waking state. Even in early human history dreams also were interpreted as reflections of waking experiences and of emotional needs. Aristotle (384–322 BC), despite his contemporaries who practiced divination and incubation, in his work *Parva naturalia* attributed dreams to sensory impressions from "external objects . . . pauses within the body . . . eddies . . . of sensory movement often remaining like they were when they first started, but often too broken into other forms by collision with obstacles." In anticipation of psychoanalyst Sigmund Freud (1859–1939), Aristotle wrote that sensory function is reduced in sleep, favouring the susceptibility of dreams to emotional subjective distortions.

In spite of Aristotle's unusually modern views and even after a devastating attack by the Roman statesman Marcus Tullius Cicero (106–43 BC) on dream divination (*De divinatione*), views that dreams have supernatural attributes persisted vigorously until the 1850s and the classical work of French physician Alfred Maury, who studied more than 3,000 reported recollections of dreams. Maury concluded that dreams arose from external stimuli, instantaneously accompanying such impressions. He wrote that part of his bed once fell on the back of his neck and woke him, leaving the memory of dreaming that he had been brought before a French revolutionary tribunal, questioned, condemned, led to the scaffold, bound by the executioner and that the guillotine blade fell.

The English poet Coleridge reported that he had written "Kubla Khan" as the result of creative thinking in a dream. Having fallen asleep while reading about that Mongol conqueror, he woke to write down a fully developed poem he seemed to have composed while dreaming. Novelist Robert Louis Stevenson said that much of his writing was developed by "little people" in his dreams, and specifically cited the story of Dr. Jekyll and Mr. Hyde in this context. A German chemist, F.A. Kekulé von Stradonitz, attributed his interpretation of the ring structure of the benzene molecule to his dream of a snake with a tail in its mouth. Otto Loewi, a German physiologist, attributed to a dream inspiration for an experiment with a frog's nerve that helped him win the Nobel Prize. In all of these cases the dreamers reported having

Creative dreaming

thought about the same topics over considerable periods while they were awake.

Psychoanalytic interpretations. Among Freud's earliest writings was *The Interpretation of Dreams* (1899). His insistence that dreams are "the royal road to the unconscious" continued from it down to his last published statement on dreams, printed about a year before he died. Freud held that dreams reflect waking experience; he offered a theoretical explanation for their bizarre nature, invented a system for their interpretation, and elaborated on their curative potential.

Freud theorized that thinking during sleep tends to be primitive and regressive and that the effects of forgetting (repression) are reduced. Repressed wishes, particularly those associated with sex and hostility, were said to be released in dreams when the inhibitory demands of wakefulness diminished. The content of the dream was said to derive from such stimuli as urinary pressure in the bladder, traces of experiences from the previous day (day residues), and from associated infantile memories. The specific dream details were called their manifest content; the presumably repressed wishes being expressed were called the latent content. Freud suggested that the dreamer kept himself from waking and avoided unpleasant awareness of repressed wishes by disguising them as bizarre manifest content in an effort called dreamwork. He held that impulses one fails to satisfy when awake are expressed in dreams as sensory images and scenes. In dreaming, Freud believed:

Dream-work

> All of the linguistic instruments . . . of subtle thought are dropped . . . and abstract terms are taken back to the concrete . . . The copious employment of symbols . . . for representing certain objects and processes is in harmony (with) the regression of the mental apparatus and the demands of censorship.

It was theorized that one aspect of manifest content could come to represent a number of latent elements (and vice versa) through a process called condensation. Further displacement of emotional attitudes toward one object or person theoretically could be displaced in dreaming to another object or person or not appear in the dream at all. Freud further observed a process called secondary elaboration which occurs when people wake and try to remember dreams. They may recall inaccurately in a process of elaboration and rationalization and provide "the dream, a smooth facade, (or by omission) display rents and cracks." This waking activity he called secondary revision.

In seeking the latent meaning of a dream, Freud advised the individual to associate freely about it. From listening to the associations, the analyst was supposed to determine what the dream represented, in part through an understanding of the personal needs of the dreamer.

EFFORTS TO STUDY DREAMING

Dream reports. Though each person seems to know his own private dreams, the manner in which people dream obviously defies direct observation. It has been said that each dream "is a personal document, a letter to oneself" and must be inferred from the observable behaviour of people. Furthermore, observational methods and purposes clearly affect conclusions to be drawn about the inferred dreams. Reports of dreams collected from people after morning awakenings at home tend to exhibit more content of an overt sexual and emotional nature than do those from laboratory subjects. Such experiences as dreaming in colour seldom are spontaneously mentioned but often emerge under careful questioning. Reports of morning dreams are typically more rich and complex than those collected early at night. Immediate recall differs from what is reported after longer periods of wakefulness; psychoanalysts seem to elicit more recollections of overt sexual dreams than do laboratory investigators. In spite of these complications, there have been substantial efforts to describe the general characteristics of what people say they have dreamed.

The reported length of dreams varies widely between and within individuals (and by inference, so does the length of the presumed dreams themselves). Spontaneously reported dreams among laboratory subjects are typically short; about 90 percent of these reports are less than 150 words long, although some may exceed 1,000. With additional probing, about a third of such reports are longer than 300 words.

Length of dream reports

Some investigators have been surprised by repeated findings that suggest dreams may be less fantastic or bizarre than generally supposed. In the language of modern art, one investigator stated that visual dreams are typically faithful to reality (representational) with little, if any, abstractionist or surrealistic dreaming. Any variations from the representational, in terms of fuzziness or simplification of imagery, were characterized as impressionistic. Except for those that are very short, dreams are reported to take place in ordinary physical settings, about half of them seeming quite familiar to the dreamer; only rarely is the setting said to be exotic or peculiar.

Apparently dreams are quite egocentric, the dreamer perceiving himself as a participant, though the presence of others is typically recalled. Seldom does the person remember an empty, unpopulated dreamworld, and individuals seem to dream roughly two-thirds of the time about people they know. Usually they are close acquaintances; family members are mentioned in about 20 percent of dream reports. Recollections of notables or weird representations of people are generally rare.

The typical report is of visual imagery; indeed, in its absence, the person may say only that he had been thinking rather than "dreaming" while asleep. Rare statements about dreams dominated by auditory experience commonly are made with claims of actually having been awake. It is unusual, however, to hear of dreams without some auditory characteristics.

One typically is told of bland dreams; when there are emotional overtones, they tend to be unpleasant about 65 percent of the time. Fear and anxiety are most commonly mentioned, followed by anger; pleasant feelings are most often those of friendliness. Reports of overtly erotic dreams, particularly among those gathered in laboratories, are infrequent.

Despite their generally representational nature, dreams seem somehow odd or strange. Perhaps this is related to discontinuities in time and purpose. One suddenly may dream of himself in a familiar auditorium viewing a fencing match rather than hearing a lecture and abruptly in the "next scene" walking beside a swimming pool. Or an individual may have the experience of lying in a hallway listening to two people standing by an elevator; he may be looking at a bleeding hand and walk across an empty room to a liquor cabinet to find a roll of adhesive tape. These sudden transitions contribute to the dreamer's feeling of strangeness, and this is enhanced by his waking statements that the bulk of his dreams cannot be clearly recalled, giving them a dim, mysterious quality.

Physiological dream research. A new era of dream research began in 1953 with the discovery that rapid eye movements during sleep seem often to signal that a person is dreaming. In that year it was observed that, about an hour or so after falling asleep, laboratory subjects are apt to experience a burst of rapid eye movement (REM) under their closed lids, a change in brain waves detected by an electroencephalograph as an electrical pattern resembling that of an alert waking person, and an increased rate of breathing. Such episodes, typically lasting five–ten minutes, are followed by a period of relative quiescence and then commonly by another appearance of REM and associated responses; the cycle usually repeats three or four times a night. In 1953 it was discovered that when subjects were awakened during REM, they reported vivid dreams 20 out of 27 times; when roused during non-REM sleep, they recalled dreams in only four of 23 instances. Subsequent systematic study confirmed this relationship among REM, activated brain waves (EEG), and dream recall. Several thousand experimental studies utilizing these observable indices of dreaming have since been conducted.

Rapid eye movement

A major finding is that the usual report of a vivid, visual dream is primarily associated with REM and activated EEG. On being aroused while exhibiting these signs, people recall dreams with visual imagery about 80 percent of the time. When awakened in the absence of them, however, people still report some kind of dream activity, though only about 30 to 50 percent of the time; and in such cases they are apt to remember their dreams as being relatively "thoughtlike," realistic, and as resembling the experiences of wakefulness.

The combined duration of the REM–EEG condition (sometimes called the D-state) takes up about 25 percent of the normal sleep period in young human adults. D-state occurs episodically in bursts throughout the night, ordinarily first appearing about 100 minutes after sleep begins and lasting for approximately ten minutes. This tends to be followed by successive episodes every 90–100 minutes through the night, each successive episode being longer than the preceding one. Young adults rarely show D-state during less than 18 percent or more than 30 percent of the sleep period. Newborn infants, however, show about 50 percent D-state. This declines until, by age 10, the D-state stabilizes at about 25 percent of total sleeping time, remaining at this level until people reach their 60s. There is tentative evidence of some slight decrease in the duration of D-state sleep among very elderly persons.

Signs of dreaming among non-humans

D-state sleep has been reported for all mammals studied; it has been observed, for example, among monkeys, dogs, cats, rats, elephants, shrews, and opossums; these signs also have been reported in some birds and reptiles.

Surgical destruction of selected brain structures among laboratory animals has clearly demonstrated that the D-state depends on an area within the brain stem known as the pontine tegmentum. Evidence indicates that D-state sleep is associated with a mechanism involving a bodily chemical called norepinephrine; other stages of sleep seem to involve another chemical (serotonin) in the brain. Among other physiological changes found intimately related to D-state sleep are increased variability in breathing and heart rate, relaxation of skeletal muscles (in lower animals), and, in humans, reduction of electrical activity in muscles near the base of the tongue and penile erections or increase in vaginal blood flow.

When people are chronically deprived of the opportunity to manifest D-state activity (by awakening them whenever there is EEG evidence of dreaming), it appears increasingly difficult to prevent them from dreaming. On recovery nights (after such deprivation) when the subject can sleep without interruption, there is a substantial increase in the number of reports of dreaming. This rebound effect continues in some degree on subsequent recovery nights, depending on how badly the person has been deprived.

During D-states in the last 6½ to 7½ hours of sleep people are likely to wake by themselves about 40 percent of the time. This figure is about the same as that for dream recall, people saying they had a dream the previous night about 35 percent of the time (roughly once every three or four nights). Evidence concerning the amount and kind of dreaming also depends on how rapidly one is roused and on the intensity of his effort to recall. Some people recall dreams more often than the average, while others rarely report them. While these two groups of people show little difference in amount of D-state sleep, evidence suggests that nonrecall reflects a general tendency on the part of the individual to repress or to deny his experiences.

The psychoanalytic literature is rich with reports indicating that what one dreams about reflects his needs and his immediate and remote past experience. Nevertheless, when someone in D-state sleep is stimulated (*e.g.*, by spoken word or by drops of water on his skin), the chances that he will say he has dreamed about the stimulus, or anything similar, are quite low. Studies in which people have watched vivid movies before falling asleep also indicate some possibilities of influencing dreams but again clearly emphasize the limitations of such influences. Highly suggestible people seem likely to dream as they are told to do while under hypnosis, but the influence of direct suggestion during ordinary wakefulness seems quite limited.

Variations within the usual range of about 18 to 30 percent of D-state sleep apparently are unrelated to differences in the amount or content of dreaming. The amount of D-state further seems generally independent of wide variations in the daily activities or personality characteristics of different people; groups of scientists, athletes, housewives, and artists, for example, cannot be reliably distinguished from each other in terms of D-state activity. Such disorders as schizophrenia and mental retardation appear to have no clearly discernible effect on the amount of time sufferers spend in REM-activated EEG sleep.

Effects of drugs

Although there are exceptions, most stimulant drugs (*e.g.*, amphetamines) and depressants (*e.g.*, barbiturates and alcohol) effectively tend to reduce D-state activity. When these drugs are taken in doses ordinarily prescribed by physicians, reduction may amount to 50 percent. Withdrawal of such drugs typically results in a dramatic increase (rebound), the person showing considerably more REM–EEG symptoms than he did before being dosed.

DREAMLIKE ACTIVITIES

Related states of awareness may be distinguished from the dream experiences typically reported; these include dreamlike states experienced as one falls asleep and as he awakens, respectively called hypnagogic and hypnopompic reveries. During sleep itself there are nightmares, observable signs of sexual activity (*e.g.*, noctural emissions of sperm), and sleepwalking. Even people who ostensibly are awake may show evidence of such related phenomena as hallucinating, trance behaviour, delirium, and reactions to drugs (see HALLUCINOGEN).

Rapid eye movement is not characteristic of sleep onset; nevertheless, as individuals drift (as inferred from EEG activity) from wakefulness through drowsiness into sleep, they report dreamlike hypnagogic experiences about 90 percent of the time on being awakened. Most of these experiences (about 80 percent) are said to be visual. If dreaming is defined as at least partly hallucinatory and somewhat dramatic, then awakening from drowsiness or at the onset of sleep yields recall of experiences that may be classified as dreams for about 75 percent of the occasions. These "dreamlets" seem to differ from dream-associated REM sleep in being less emotional (neither pleasant nor unpleasant), more transient, and less elaborate. Such hypnagogic experiences apparently tend to incorporate abstract thinking and recall of recent events (day residues) and to be quite typical of falling asleep. Systematic studies remain to be made of the hypnopompic reveries commonly reported mornings before full arousal, but it seems likely that they include recollections of the night's dreams, or represent one's drifting back into transient REM sleep.

Extreme behavioural manifestations during sleep—night terrors, nightmares, sleepwalking, enuresis (bed-wetting)—all have been found generally unrelated to ordinary dreaming. Night terrors (*pavor nocturnus*) are characterized by abrupt awakening, sometimes with a scream; a sleeping child may sit up in bed, apparently terror stricken, with wide-open eyes, and often with frozen posturing that may last several minutes. Afterward there typically is no recollection of dreamlike experience. Observed in about 2 or 3 percent of children, roughly half of the attacks of night terror occur between the ages of four and seven; about 10 percent of them are seen among youngsters as old as 12 to 14 years.

Nightmares

Nightmares typically seem to be followed by awakening with feelings of suffocation and helplessness and expressions of threateningly fearful thinking. Evidence of nightmares is observed for 5 to 10 percent of children, primarily about eight to ten years of age. As early as 1897 it was reported that night terrors could be induced by rousing soundly sleeping children and suggesting dire events (*e.g.*, that the house is on fire). Later it was found that signs of spontaneously

generated night terrors and nightmares are related to abrupt awakening from deep sleep that experimentally appears dreamless. This suggests that the vividly reported fears well may be produced by emotional disturbances that first occur on awakening.

Sleepwalking, observed in about 1 percent of children, predominantly appears between ages 11 and 14. Apparently sleeping individuals rise and walk from their beds, eyes open, usually avoiding obstacles, and expressing no recollection of the episode when they wake. Studies of EEG data indicate that sleepwalking occurs only in deep sleep when dreams seem essentially absent; the behaviour remains to be reported for REM sleep.

Enuresis, or bed-wetting, occurs in about one-fourth of children over age four. These episodes seem not to be associated with REM as much as they do with deep sleep in the absence of D-state signs.

Nocturnal emission of sperm remains to be described in terms of any distinguishing EEG pattern; such events are quite rarely observed among sleeping laboratory subjects. Among a large sample of males who were interviewed about their sexual behaviour about 85 percent reported having experienced emissions at some time in their lives, typical frequency during the teens and 20s being about once a month. Of the females interviewed 37 percent reported erotic dreams, sometimes with orgasm, averaging about three to four times a year. Most often, however, openly sexual dreams are said not to be accompanied by orgasm in either sex. Males not infrequently could recall no dreams associated with emission, although most implicated erotic dreaming.

Dreamlike experiences induced as trances, deliriums, or hallucinatory behaviour by drugs seems attributable to lowered efficiency of the central nervous system in processing sensory stimuli from the external environment. The result seems to be that one's physiological activities begin to escape environmental constraint to the point that internalized, uncritical thinking and perceiving prevail.

As noted, since antiquity dreams have been viewed as a source of divination, as a form of reality, as a curative force, and as an extension or adjunct of the waking state. Psychoanalytic theorists stress the individual meaningfulness of dreams and their relation to personal hopes and fears. Contemporary research focusses on efforts to discover and describe unique, complex biochemical and neurophysiological bases of dreaming. Among the plethora of theories ranging from those that assert dreaming to be awareness of a god's voice to those that reduce the dream to physical activity in the nervous system, no single, encompassing theory seems yet to be available (see also SLEEP).

BIBLIOGRAPHY. SIGMUND FREUD, *Die Traumdeutung* (1900; Eng. trans., *The Interpretation of Dreams*, 1953), Freud's classical theory of dream interpretation, also found in *An Outline of Psychoanalysis*, trans. by JAMES STRACHEY (1949); G.E. VON GRUNEBAUM and R. CAILLOIS (eds.), *The Dream and Human Societies* (1966), a scholarly work on the place of dreams in a wide variety of cultures, with a chapter on dream research; R.M. JONES, *The New Psychology of Dreaming* (1970), an attempt to coordinate experimental findings on dreams with classical theories; EDWIN DIAMOND, *The Science of Dreams* (1962), a book, written for the layman, that serves as a bridge between early ideas about dreams and subsequent experimental work; DAVID FOULKES, *The Psychology of Sleep* (1966), a work primarily about the experimental analysis of dreams, written at a level comprehensible to the layman; H.A. WITKIN and H.B. LEWIS (eds.), *Experimental Studies of Dreaming* (1967), a collection of papers on dream research with particular emphasis on such psychological variables as pre-sleep experiences and aspects of dream recall; ERNEST HARTMANN, *The Biology of Dreaming* (1967), a technical work emphasizing biological and physiological bases. Additional relevant material may be found in W.B. WEBB, *Sleep: An Experimental Approach* (1968).

(W.B.W.)

Dreiser, Theodore

It was as the author of the then "shocking" novel *Sister Carrie* that Theodore Dreiser, in 1900, made his entrance on the American literary stage and was received so coldly that he fled it for 11 years. Although his life was devoted also to magazine editing and writing, philosophical inquiry, poetry, and social protest, it was as novelist that he showed his real genius at the same time that he highlighted the slow relaxation of American mores. The man rejected in 1900 for violating literary proprieties was acclaimed as a master in 1925 for *An American Tragedy*. That the latter was accepted was because the public, not Dreiser, had changed.

Born on August 27, 1871, in Terre Haute, Indiana, a member of a family of 12 who often lived in appalling poverty, Dreiser was so offended by his German-born father's narrow Catholicism that he later abandoned and denounced the church. His gentle, illiterate, Ohio-born mother was the parent he always remembered with love. Educated in several Indiana towns where the harried family moved in hope of betterment, he managed to complete one year at the University of Indiana with the aid of a teacher who saw promise in him. Thereafter, he wandered into journalism in Chicago, St. Louis, and Pittsburgh before reaching New York in 1894. There, aided by his then prosperous older brother Paul Dresser, a popular composer ("On the Banks of the Wabash"), he became a writer of magazine articles. He was tall and hulking, with a peasant crudity that could also be charming—a thinker influenced by T.H. Huxley, Herbert Spencer, and Darwin, believing man a product of continuing evolution that he was helpless to alter. His marriage in 1898 to Sara White, whom he had met in St. Louis, soon encountered difficulties because of his roving affections and moody temperament.

Sister Carrie

Sister Carrie sold only 456 copies. Although it was praised by the young novelist Frank Norris, and in England by Arnold Bennett and H.G. Wells, its quick death sent Dreiser into emotional prostration during which he all but lost his sanity and came near starving in a Brooklyn slum. He was saved by the kindly Paul, who had been searching for him. The novel's impropriety lay less in the immorality of the heroine than in her "common" origin and the fact that she was not punished for her transgressions. Based largely on Dreiser's family recollections, including an adventure of one of his sisters, *Sister Carrie* depicted crises and intimacies of folk life not yet welcomed in print, revolutionary in the realistic detail and profound human sympathy that shone through Dreiser's awkward writing style.

Taking refuge in editorial work, he became, in 1907, the well-paid editor in chief of the housewife-oriented Butterick magazines, which embraced the very taboos that had smothered his novel. Here—not without some wry humour over the literary pruderies his job compelled him to observe—he dealt with contributors including H.L. Mencken, who became his friend and critic. Forced out in 1910 because of a "scandalous" infatuation for the daughter of an assistant editor, he finished his second novel, *Jennie Gerhardt* (1911), which was lauded by Mencken and fared well enough to encourage him. His fascination with the machinations of the unscrupulous and successful emerged in *The Financier* (1912) and *The Titan* (1914), both patterned after the adventures of the railway magnate Charles T. Yerkes.

The door to acceptance, opening a crack, was slammed in his face in 1916 when *The "Genius"* (published the year before), largely a literary reconstruction of his own artistic and romantic travail, was suppressed by the New York Society for the Suppression of Vice. Though Mencken joined him loyally in a long campaign against such censorship, Dreiser was forced again to resort to magazine writing for subsistence. Separated from his wife, he lived thriftily in New York's Greenwich Village, then a haven for the avant-garde in literature, art, and politics. There he was known for his superstitions and his affairs with women as well as his resentment of a society that rejected honest writing. Among his friends were other rebels, including the poet Edgar Lee Masters and the writers Floyd Dell, Sherwood Anderson, George Jean Nathan, and the Powys brothers, Llewellyn and John C. He had published the autobiographical *A Traveler at*

Forty (1913) after a trip to Europe, *A Hoosier Holiday* (1916) after a motor journey to his native state, and *Plays of the Natural and Supernatural* (1916). None of these books was very profitable. Nor did he prosper with the tragic drama *The Hand of the Potter* (published 1918), *Free*, a collection of short stories (1918), the taken-from-life sketches of *Twelve Men* (1919), the philosophically unorthodox *Hey Rub-a-Dub-Dub* (1920), the autobiographical *A Book About Myself* (1922), a portion of which was to appear as *Newspaper Days* in 1931, or the descriptive *Color of a Great City* (1923). His prospects seemed bleak in 1925 when he published *An American Tragedy*, his first novel in ten years, based on a famous murder case in New York state.

An American Tragedy

This story of a young man trying to escape poverty who is driven to murder by social forces he cannot resist was an immediate critical and popular success. It was dramatized on Broadway and sold to Hollywood for a film, bringing Dreiser the acclaim and wealth that had for so long been denied him. Joseph Wood Krutch, a well-known critic, called it "the greatest American novel of our generation." Critics formerly hostile to Dreiser even agreed that it rose above his well-known difficulties with syntax and diction. As Sherwood Anderson wrote, urging Americans to read it,

> Find out, once and for all, the difference between a human flesh and blood, male man, full of real tenderness for life, and the smarties, the word slingers, the clever fellows . . . of the writing world.

Dreiser quit the Village, took an elegant apartment and a country place, bought gaudy clothing, and became the darling of literary intellectuals of that bull-market time. His visit to Russia, where he saw little hope for social improvement because his mechanistic philosophy ruled out such hope, resulted in the skeptical *Dreiser Looks at Russia* (1928). His evident borrowings from Dorothy Thompson's book about Russia caused him to quarrel with her husband, Sinclair Lewis. His *Chains* (1927) and *A Gallery of Women* (1929) comprised short stories and sketches written earlier. Indeed, his greatest literary days were over. He communed with scientists, hoping to find explanations for eternal mysteries he felt religion evaded. Never divorced, he had lived since 1920 with the beautiful and admiring Helen Richardson, a second cousin, who endured his other affairs as a necessity of his genius—emotional stimulants he felt he needed to bring out his best writing. His income was reduced by the 1929 market crash and the depression that followed it, forcing him to scale down his luxurious mode of life. His book of poems, *Moods, Cadenced and Declaimed* (1926), had a sale of 922 copies, and the autobiographical *Dawn* of the same year sold not much better. No slave to consistency or academic logic, he began to see possibilities for good in the Communism he had recently rebuffed as well as in other plans for social amelioration. He was angered by American joblessness and privation alongside opulence. Unreasonable though he could be at times toward individuals, he had endless sympathy for society's underdogs. He threw himself into social protest despite its obvious conflict with his determinist philosophy, giving so much time and effort to insurgent causes that two long-planned novels, *The Bulwark* and *The Stoic*, were put aside and his writing in general suffered. The literary hero of the '20s became the tough-talking radical reformer of the '30s. He was so independent of the "party line" that the American Communist Party utilized his prestige for years before it would admit him to membership under the careful auspices of more conformist members. Soviet Russia later rewarded him with $34,600 in royalties for Russian sales of his works.

Decline as a writer

In 1939 he and Helen moved to Hollywood, where he squeezed a living out of the capitalist system he had come to despise, in part by sales of his earlier works to the films. He published occasional pamphlets criticizing government inadequacies and mailed them out at his own expense. His *America Is Worth Saving* (1941) was little more than an expanded pamphlet against capitalism. Earlier friendships—even the one with his longtime crony, Mencken—dwindled, and although he established new ones, his California life was a break with the past. He married the adoring and neglected Helen secretly in 1944 after the death of his first wife—a triumph for her although his stormy infidelities continued. His last novels, the long-delayed *The Bulwark* (1946) and *The Stoic* (1947) suffered from his haste and fatigue.

Assessment

Although his brief popular vogue had passed when he died on December 28, 1945, at his home in Hollywood, he had won a unique place in literature through a sincerity, power, and compassion so striking as to render less important his failings in technique and polish. *Sister Carrie* and *An American Tragedy* are enduringly great folk novels as well as literary landmarks. His autobiographical works combine with his letters and other writings to reveal a rebellious life composing a drama of its own in its intense activities and involvement, his efforts to solve cosmic mysteries along with earthly political inequities, his belief in tea-leaf readers and Ouija boards existing without conflict with his excursions into pure science, his frank sensuality, his violence and unreason sometimes balanced by geniality and tenderness. His works have been widely translated, with large editions in Russia; critical interest in him still persists.

MAJOR WORKS

FICTION: *Sister Carrie* (1900); *Jennie Gerhardt* (1911); *The Financier* (1912); *The Titan* (1914); *The "Genius"* (1915); *Free, and Other Stories* (1918), including "The Lost Phoebe" and "The Second Choice"; *An American Tragedy* (1925); *Chains* (1927); *The Bulwark* (1946); *The Stoic* (1947).

OTHER WORKS: *A Traveler at Forty* (1913); *A Hoosier Holiday* (1916); *Plays of the Natural and Supernatural* (1916); *The Hand of the Potter* (published 1918), a play; *Twelve Men* (published 1919), biographical sketches; *Hey Rub-a-Dub-Dub: A Book of Essays and Philosophy* (1920); *A Book About Myself* (1922); *The Color of a Great City* (1923); *Moods, Cadenced and Declaimed* (1926; as *Moods, Philosophical and Emotional, Cadenced and Declaimed*, 1935), verse; *Dreiser Looks at Russia* (1928); *A Gallery of Women* (1929); *My City* (1929); *Epitaph* (1930); *Dawn* (1931); *Tragic America* (1932); *America Is Worth Saving* (1941).

BIBLIOGRAPHY. The bulk of Dreiser's colourfully voluminous letters and papers are at the University of Pennsylvania Library, Philadelphia. W.A. SWANBERG, *Dreiser* (1965), an exhaustive biography, avoids literary criticism. ROBERT H. ELIAS, *Theodore Dreiser: Apostle of Nature* (1949), illuminates Dreiser's philosophy and in the emended reissue (1970), contains a valuable survey of research and criticism. F.O. MATTHIESSEN, *Theodore Dreiser* (1951), is primarily criticism; while *The Stature of Theodore Dreiser*, ed. by ALFRED KAZIN and CHARLES SHAPIRO (1955), is an anthology ranging from reminiscence to criticism and contains the best general bibliography yet available. HELEN RICHARDSON DREISER, *My Life with Dreiser* (1951), is her candid account of their difficult relationship; the *Letters of Theodore Dreiser*, 3 vol., ed. by R.H. ELIAS (1959), is a rich selection. A balanced critical biography is PHILIP L. GERBER, *Theodore Dreiser* (1964); RICHARD LEHAN, *Theodore Dreiser: His World and His Novels* (1969), stresses the writer against his background. *Two Dreisers* by ELLEN MOERS (1969), combines biography and criticism in an intensive study concentrating on *Sister Carrie* and *An American Tragedy*.

(W.A.S.)

Dress

The varieties of dress (from the French *dresser*, "to set out, to arrange") worn by human beings, now and in the past, are so numerous that it is not possible to deal with all of them. Thus, this article does not consider national or regional costumes of the various peasant peoples of Europe or other parts of the world or the garb of modern primitive peoples of Africa, the Americas, and Asia. It deals chiefly with three main aspects of the subject of dress: (1) the ancient Mediterranean world; (2) western Europe (chiefly England and France), considered as standard for Europe and America, from the Middles Ages to contemporary times; and (3) East and South Asia.

Paleolithic origins

Clothes and fire presumably were discovered in Paleolithic times, and both were important in extending the range of human habitation. It is possible only to guess at

the nature of clothes worn at a time so remote. Some authorities believe that Heidelberg man may have used his powerful jaws for chewing leather (as the Eskimo does today) and so have made it pliable enough for use as clothing. Neanderthal man certainly had tools, some of them strongly suggesting the instruments used by tanners; Neanderthal woman stayed at home while the man went hunting, and it is thought that it was at this time that a differentiation arose between male and female dress. In a cave of Cro-Magnon time, at Cavillon in France, a hairpin was found in such a position near a body as to leave no doubt as to its use. Other objects resembled toggles, or double buttons, and there were also bracelets and leg ornaments. Caverns in the Dordogne have yielded bone pendants and collars of ivory, shells, stags' teeth, and fish vertebrae. Mural paintings in a rock shelter at Cogull in the Spanish province of Lérida (Figure 1), show people clothed in what seem to be skin garments somewhat resembling those of the modern Eskimo. Skin garments cannot be made without needles, and specimens of these have been found in large numbers in Magdalenian burials.

From *Historia del Arte;* photograph, E.D.I. Studio, Barcelona, Spain



Figure 1: Paleolithic women wearing fur skirts. Detail of a reproduction of a mural painting in the rock shelter at Cogull, Lérida, Spain, possibly 8000 BC. In the Museo Arqueológico, Barcelona, Spain.

The theory of one scholar is that there are strong reasons for believing that the use of fundamental garment forms of a rudimentary perfection dates from the Magdalenian. A simple tunic, the ancestor of a long line of similar pieces, the skirt and kilt (or divided coverings for the legs), the mantle, or cape (precursor of the coat), as well as moccasins or boots, are among the probabilities. If this is the case, the fundamental forms of European clothing were established at this remote period and hardly changed in essentials until the middle of the 14th century AD.

Neolithic garments

Neolithic man made use of wool and flax fibres for clothing manufacture. Decorated spindle whorls found, for example, in the Grotte du Pontal, Hérault, indicate that the vital discovery of thread had been made, and fragments of both plaited and woven cloth have been found at Robenhausen in Switzerland and elsewhere. Dyes were used, the colours being the primaries—red (obtained from hematite), yellow (from the yellow weed *Reseda luteola*), and blue (from danewort).

Primitive man not only dyed his few clothes but also painted his body; and body markings, by painting, tattooing, or scarification, served to promote vanity, denote prestige, attract the opposite sex, indicate clan membership, and frighten the enemy, and they were probably not differentiated in his mind from garments. (Ja.L.)

This article is organized as follows:

I. Western dress
Ancient Egypt
Mesopotamia
Persian, Arab, and Ottoman empires (to modern times)
The Aegean
Greece and Rome
Europe through the 18th century
Colonial America
Europe and America, 19th and 20th centuries
II. Non-Western dress
East Asia
South Asia

I. Western dress

ANCIENT EGYPT

Modern knowledge of ancient Egyptian dress derives for the most part from ancient paintings and sculpture, since very few garments have been preserved. It is necessary to remember, therefore, that artists were very much bound by tradition and, further, that modes of representation lagged behind actual changes of fashion.

Practically the only fabric that has been preserved is linen, found in graves of the Neolithic period long before the Dynastic period. The Egyptians themselves realized that flax culture was very old and believed that the gods were clothed in linen before they made their first appearance on Earth. Wool seems scarcely ever to have been used, and silk and cotton were unknown. The Egyptians had not learned to use the mordants required to dye linen successfully and, as their clothing was frequently washed, only white was really practical. Nonetheless, yellow, blue, green, and red garments were worn on occasion. Costumes made of materials with coloured patterns are represented; but whether the material was linen, and if so how these patterns were applied and how often such garments actually appeared in real life, is uncertain. Tapestry weaving was known but was extremely rare, tapestries and embroideries apparently occurring only in garments belonging to the king and to members of his family and, indeed, being treated as heirlooms. To make up for the lack of colour, the material of the dress might be elaborately pleated or arranged in a mass of fine folds; bright ribbons might be tied around the waist and hair, and those who could afford to do so wore brilliantly coloured jewelry, often combined with garlands and wreaths of flowers.

As is usually the case, new fashions originated among the rich; the general tendency was toward greater and greater elaboration in both costume and accessories.

Men's clothing

The first clothing worn by men was a narrow band around the waist to which pendants were attached, the whole arrangement being both amuletic and decorative. The loincloth and kilt were later developments of this. The characteristic masculine garment throughout the Dynastic period was a white linen kilt, a rectangular piece of cloth wrapped around the body and tied in front. The length, fullness, and method of adjustment varied with the date and the social position of the wearer. In the earlier Old Kingdom the kilt was usually short and draped smoothly around the hips. Often one or both of its lower corners were rounded; sometimes it was pleated or partially pleated and later it might be stiffened to project in front (Figure 2). In the Middle Kingdom there was a fashion for long kilts, sometimes reaching from waist to ankles, sometimes hanging from the armpits, with no belt or other interruption at the waist. In the late 18th dynasty, a double kilt appeared, long and full, with the upper one doubled and gathered in front. Fullness was always concentrated in front; and all garments—both men's and women's—were adjusted so as to fit the figure smoothly behind (Figure 3). A triangular loincloth sometimes, perhaps usually, was worn under the kilt.

Although the upper part of the body was often left bare or covered with a shawl, shirts were worn during all periods. They are first represented in art, however, in the 18th dynasty. The shirt, like the kilt, was made of a rectangular piece of linen, the material being folded and sewn up the sides, with openings left for the arms and a hole cut at the fold for the head. Several shirts have been preserved with long, tight sleeves sewn into the armholes. The richly furnished intact tomb of the architect Kha of the 18th dynasty (the contents of which are in the Museo Egizio, Turin) contained a pile of about 50 loincloths,

many kilts 20 to 23 inches (50 to 58 centimetres) wide to about 67 inches (170 centimetres) long, seven sets of loincloths and kilts tied up together, and a pile of shirts of both light summer and heavy winter weight.

Women's clothing

The earliest representations of women show them either nude or clad in tightly fitting white linen skirts down to the ankles. The characteristic woman's garment in the Old Kingdom was a long, tight slip held up by suspenders over the shoulders (Figure 2). During the Middle Kingdom and early 18th dynasty, artists copied Old Kingdom representations closely; and it is not possible to ascertain when the draped robe, which first appeared in pictures after the first half of the 18th dynasty, actually came into fashion (Figure 3). These robes were unshaped, each consisting of a length of material wound around the body, brought up over one or both shoulders, and tied at the breast. There were numerous ways of arranging the drapery; most commonly, the right arm was left free. Some pictorial representations seem to indicate a tight garment under the robe, but probably the artist was only trying to represent the first twist of the cloth, strained over the body so that folds or pleats were smoothed out. Men are also represented wearing the draped robe in the later New Kingdom.

Hirmer Fotoarchiv Munchen

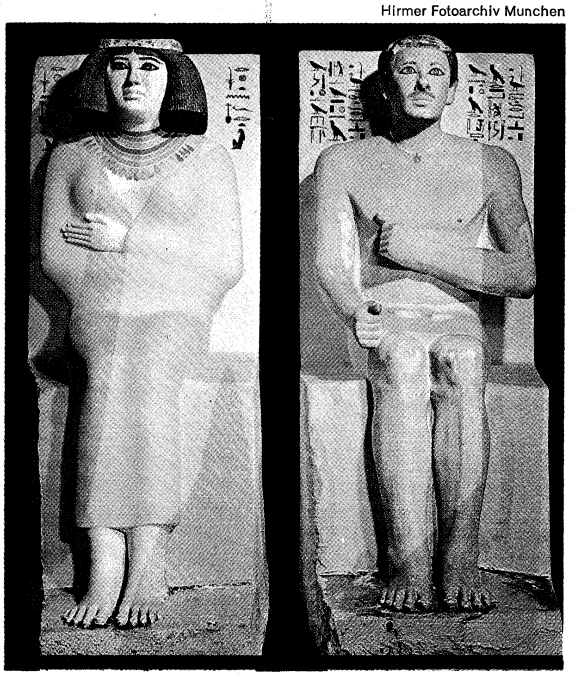


Figure 2: Egyptian dress of the Old Kingdom, 4th dynasty. Prince Ra-hotep wears a short kilt draped smoothly around his hips; Lady Nefret is clothed in a mantle. Painted limestone figures from Maydūm, c. 2600 BC. In the Cairo Museum.

Women, like men, wore heavy mantles for warmth. Examples survive of materials with a self-pattern on the right side that appears on the reverse as a thick, warm tufting. In the privacy of their homes most women may have gone about without clothing, as did their maids, who are often so represented.

Children are invariably shown naked, their heads shaved except for the "lock of youth" that hangs down one side. Still, a few children's shirts—smaller models of their parents'—have been preserved.

Sandals of woven rushes or of leather were worn on occasion by men, women, and children; but probably the Egyptians usually went barefoot, as, indeed, they still do in rural districts.

The clothing described above was for the most part that of the official class. The king himself is shown either in the severely simple short kilt of the Old Kingdom or in contemporary costume, with decorations peculiar to himself—an elaborately worked jacket, a jewelled belt and apron, and a variety of crowns or royal headdresses. At the other end of the social scale were the labourers and

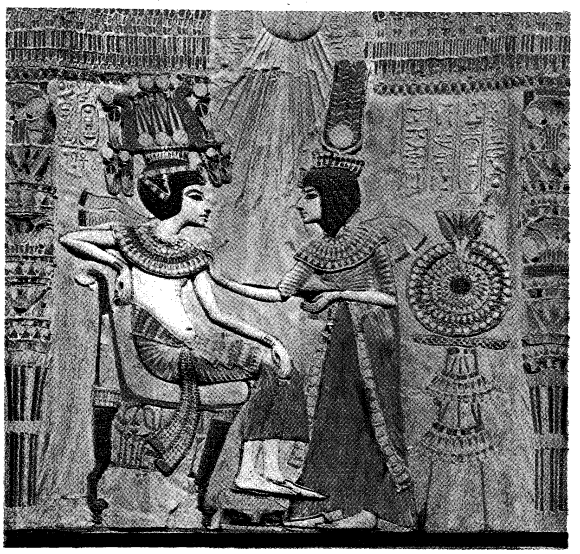


Figure 3: Egyptian dress of the New Kingdom, 18th dynasty. King Tutankhamen wearing a double kilt, long and full, with the upper one doubled and gathered in front; Queen Ankhesenamen in a draped robe tied at the breast and leaving the right arm free. Detail from the back of the throne of Tutenkhamen (reigned 1361–52 BC). In the Cairo Museum.
Hirmer Fotoarchiv Munchen

workers in the fields, who went nude or perhaps wore a belt or loincloth of linen or leather.

Wigs worn by men and women

Heavy wigs or a padding of false hair, worn by both men and women, are known from an early period. They served not only as an adornment but also to protect the wearer's head from the burning rays of the sun, thus in a way acting as hats. Semicircular kerchiefs, tied by the corners at the nape of the neck under the hair, were sometimes worn to protect the wig on a dusty day. Wigs were dressed in many different ways, each characteristic of a given period; generally speaking, the hair became longer and the arrangement of curls and braids—set with beeswax—more complicated as time went on.

The earliest records indicate that the Egyptians grew hair on their chins. They frizzed, dyed, or hennaed this beard and sometimes plaited it with interwoven gold thread. Later, a metal false beard, or postiche, which was a sign of sovereignty, was worn by queens as well as kings. This was held in place by a ribbon tied over the head and attached to a gold chin strap, a fashion existing from about 3000 BC to 1580 BC.

Cosmetics, like wigs, served a practical as well as an ornamental purpose. During the Predynastic period, men as well as women applied a line of green paint around the eyes, which helped to absorb some of the sun's glare. In dynastic times the colour of paint used was a dark gray. Red ochre seems to have been used to colour the cheeks and henna the palms, nails, and, in the late period at least, the hair. (Ed.)

MESOPOTAMIA

Typical Mesopotamian garments

Simplicity characterized Mesopotamian clothing in the periods of that civilization's greatest flowering. Garments typically consisted of a rectangular piece of loosely woven wool, bleached or dyed red or blue. Linen was employed to a lesser extent, its use increasing in the 1st millennium BC. The warp formed a straight fringe along the short sides, sometimes knotted into tassels. The weft could be made to stick out in loops, forming a wavy fringe on the long sides. Red and blue garments had white fringes, and white garments often had coloured fringes. As Mesopotamian art suppresses most folds, the rendering of the borders is the best clue to the draping of the garment. Men normally draped their clothes counterclockwise, women clockwise. Since manual labour and warfare require freedom of movement, the long garments denote leisure and peaceful pursuits. In the earliest periods, clothes were discarded during the performance of feats of strength or service in the temple.

Protoliterate and Early Dynastic periods. During the Protoliterate period (*c.* 3400–*c.* 2900 BC) men of rank wore an ankle-length skirt, the rectangular cloth wrapped so that the visible short side was in front.

During the Early Dynastic period (*c.* 2900–2334 BC) the cloth was wrapped as before; but the short side was behind the left hip, where the ends of the belt or the cloth itself were knotted. A fringe of long weft loops encircled the skirt below. In the later part of the period, a luxury version had rows of loops all over (Figure 4).

By courtesy of (left) the Metropolitan Museum of Art, New York, Fletcher Fund, 1940, (right) Iraq Museum, Baghdad; photograph (right), Fratelli Fabbri Editori, Milan, Italy



Figure 4: *Sumerian votive figures from the Square Temple of Abu at Tall al-Asmar, wearing Mesopotamian dress.* (Left) Male wearing skirt with long weft loops encircling the bottom, c. 2600 BC. In the Metropolitan Museum of Art, New York. (Right) Female draped in cloth placed over the left shoulder and then passed under the right arm, c. 2900 BC. In the Iraq Museum, Baghdad.

Women wore a similar but larger cloth. With one short side it was placed over the left shoulder, then it was passed under the right armpit and again over the left shoulder, from where the other side hung down to the ankles, covering the left arm except for the hand. A hidden pin or stitching may have held the two layers of material together on the left breast (Figure 4).

Akkadian, Neo-Sumerian, and Old Babylonian periods. Men of rank now also wore a full-sized cloth (at least 100 by 50 inches [250 by 125 centimetres]), placed with the centre of one short side on the left lower arm (which had to be kept horizontal), wrapped around the back, under the right armpit and over the left arm and shoulder, thus giving freedom of movement to the legs and right arm. The other short side curved up the back. One corner was either tucked under the right armpit or brought forward over the right shoulder. When the cloth grew even longer, it was wrapped around twice, ending over the left arm. Exceptionally, the cloth might be wound clockwise, the first turn forming a skirt and the second brought over the left shoulder with the same result of disengaging the right arm while enveloping the left.

Garments of men of rank

The beret, worn by rulers in the north in the Early Dynastic period, was current in the Akkadian period. Thereafter a turban, sometimes covered with curls as if made of fur, was the royal headgear.

For women, an arrangement covering both shoulders, not unknown before, now became common. The centre of one long side was placed across the bosom. Both short sides were passed under the armpits, crossed in the back and brought forward over the shoulders and upper arms, so that two points hung down in front. The hair might be enveloped in a cloth held in place by a headband.

Kassite period. Mountaineers from the border of Iran who ruled Babylonia during this time (*c.* 1595–*c.* 1157 BC) introduced heavy, richly decorated clothes and soft shoes.

Assyrian period. The basic garment of the Assyrians was the rectangular tunic. The slits for neck and arms often had embroidered borders, and tassels sometimes appeared around the bottom. Depending on the freedom of movement desired, tunics were ankle length or knee length, girt with a cummerbund over which a narrow belt might be worn. Sandals were the common footgear (Figure 5).

Over the tunic, the king threw a long shawl, draped like the traditional garment but, being narrower than the latter, covering the right side only from the waist down. For the lesser members of the court the width of the shawl decreased, becoming a mere sash. The king also wore a headband with rosettes, either separately or over a pointed cap similar to a fez.

By courtesy of the Museum of Fine Arts, Boston



Figure 5: Assyrian king (possibly Ashurnasirpal II), with an elaborately dressed beard, wearing sandals and a full-length tunic decorated with embroidery and tassels, amber figurine, 9th century BC. In the Museum of Fine Arts, Boston.

Assyrian women wore a sleeved tunic and a large shawl, wrapped so as to cover both shoulders.

The Babylonians also wore sleeved and belted tunics, drawn smooth in front so that the folds of the material were concentrated in the back of the garment. Diagonal shoulder bands may denote rank. The royal headdress (generalized in Persian times) took the shape of a pointed cap with a streamer. (M.N.v.L.)

Importance of beards

The Assyrians (like the ancient Persians) devoted great care to oiling and dressing their beards, using tongs or curling irons to create elaborate ringlets and frizzles, in a tiered effect. Assyrians resorted to a black dye for eyebrows, hair, and beard, whereas the Persians used henna, which produced an orange-red colour, a style that existed from 1900 BC. Gold dust, gold thread, and scented yellow starch were sometimes used in the hair and beard for festive occasions. (W.J.Tu.)

PERSIAN, ARAB, AND OTTOMAN EMPIRES (TO MODERN TIMES)

The primitive costume of the Persians was made of animal skins, for wool was little known before the time of

Cyrus and cotton and silk were not known at all. The costume was in the form of a coat and breeches; and this, with minor modifications, has lasted until modern times. With growing prosperity, the Persians began to make both coat and trousers of coloured cloth. They wore at first a conical leather hat, which was exchanged for a time for the Phrygian cornet (Figure 6). The conical



Figure 6: Achaemenid Persian man wearing a coat, trousers, and a Phrygian cornet. Silver statue, *c.* 559–330 BC. In the Vorderasiatisches Museum, East Berlin.
By courtesy of Vorderasiatisches Museum, Berlin

cap was sometimes circled by bands of cloth, in which perhaps may be seen one of the origins of the turban. A relief found in the ruins of the palace of Persepolis shows various figures bringing presents, clad in a mixture of Persian and Hebrew modes (the Hebrew dress was longer than the Persian and, in the Semitic tradition, without trousers). There is hardly any trace in ancient Persian monuments of the dress of women.

Scythians and Parthians wore trousers with coats that were open at the front and crossed with a belt to keep them in place. The coats of the women were longer but otherwise similar, except that the belt was sometimes replaced by a long piece of cloth wound around the hips and knotted in front.

Modifications brought about by the Arab conquest

7th century. Arab conquest in the 7th century AD brought modifications. Arab costume itself has been singularly unchanging, as can be seen from occasional representations in monuments of ancient Assyria and Egypt. The Arabs' earliest garments seem to have been a piece of cloth wrapped around the body below the armpits and a long, full shirt with or without sleeves. The descendant of this latter garment is the garb of many modern Egyptians. The Arabs wore over it a mantle of coarse wool or of camel's hair in the shape of a sack open at the bottom, with holes for arms and head. This is the *aba* mentioned in the Bible as the costume of the Hebrew prophets. It was usually cream coloured with wide vertical stripes of black, white, brown, and blue. The headdress was a small piece of cloth wound round the head or folded in a triangle and kept in place by a ribbon tied in a knot at the front or a cord, plain or ornamented, encircling the crown. This is the *ḥā'ik*, the characteristic Arab headdress even of the 20th century. Under it is worn a felt cap (or several felt caps one over the other) coloured red, white, or brown.

No representation is known of the early dress of Arab women, but it probably resembled closely the dress of men and has continued to do so. The shirt, or chemise, reaching to the ground and very wide, was open over the chest but buttoned at the neck. A scarf was worn around the waist. The mantle consisted of a large rectangle of woolen cloth, black or blue with yellow or red stripes. It was sometimes worn over the head on top of a small fichu (triangular scarf) wound around the crown. The usual veil is a band of black crepe or white muslin, sometimes long enough to reach the feet, kept in place by two ribbons attached to a circlet around the head and often ornamented with pearls or coins.

Persian dress

11th century. The great Turkish migration of the 11th century brought some changes to both Arab and Persian costume. The Persian costume at this time consisted of a shirt, a caftan (a Turkish word describing a kind of long coat open at the front, with short sleeves, and sometimes edged with fur), and a cloak. The cap of lambskin in the form of a truncated cone, introduced by the Turks, is still worn in Persia. A cap of similar form but red in colour and made of felt, the fez, became the characteristic headgear of Turks until abolished by Kemal Atatürk in 1925.

Characteristically, the Persian trousers were baggy and gathered in at the ankles. Slippers were made of leather and had turned-up points. Kurds, Afghans, Georgians, and Armenians long preserved these modes, which are by no means extinct.

16th century. Documentation increases during the 16th century, when Europe became conscious of the Turkish threat, and there was a growing interest in exotic dress. The costume worn by Persian women—and still surviving—consisted of a white cotton chemise with long sleeves; long woolen stockings; woolen embroidered slippers; wide trousers of dyed cotton, tied around the ankles and hanging in folds; an underwaistcoat with long sleeves open to the elbow and a short-sleeved overwaistcoat; a small cap; and a shawl tied around the waist. A veil of white muslin was either wound around the head or suspended just below the eyes. Out of doors, women threw over their shoulders a piece of cloth that covered them completely.

The caftan and *'abā*

Men's dress at this period, at least among the upper classes, consisted of trousers at first extremely ample (still worn among the Kurds) but gradually growing tighter. The camise, or shirt, often striped and patterned, worn outside the trousers and falling sometimes to the knees and sometimes to half-calf, was provided with long sleeves. The caftan was buttoned across the chest and closed by means of a scarf (Figure 7). Its long skirts were sometimes turned back and attached to the belt. An outer garment as long as the caftan and always more ample was of a different colour. In the 20th century the overmantle was provided with long wide sleeves, sometimes slit. A very ancient garment, long retained, was the *'abā*, a kind of long cloak, originally made of camel hide, thrown over the shoulders. Sometimes it had a single wide sleeve, a peculiarity preserved by the modern Kurds.

The waistcoat, dating from the end of the 16th century, was generally of red or blue cloth with braid of a contrasting colour and adorned with small buttons sewed very close together. Over this was sometimes worn a short coat, rather like an Eton jacket, similarly ornamented. Red leather slippers were worn by all except clergy and magistrates, who wore blue.

Turkish headgear

The original headgear of the Turks was probably a round Tatar cap edged with fur, but after the capture of Constantinople the sultan imitated the Prophet Muḥammad by surrounding his cap with a large amount of white muslin wound around and around. It was at this period that the turban became usual among the Turks. Functionaries were compelled to wear it in a form determined by their rank. The costume of the lower classes consisted simply of trousers and a cloak of hide or cloth.

Women's costume was hardly to be distinguished from



Figure 7: Persian prince wearing a caftan buttoned across the chest and closed with a scarf. "Prince Muhammad Beg of Georgia," watercolour on paper by Reẓā 'Abbāsī, Isfahan, early 17th century. In the Islamisches Museum Berlin, East Berlin.
By courtesy of the Islamisches Museum Berlin

that of men except that the trousers continued to be more ample. The chemise, made of white muslin and falling to mid-thigh, was décolleté and had no sleeves or very short ones. The dress was open in front and reached the ground. Veils of black horsehair or white muslin were attached to the headdress, a small cap that was often ornamented with precious stones. The hair was worn either loose and flowing or in plaits encased in a kind of cloth tube.

20th century. In the second half of the 20th century, the visitor to Near Eastern countries finds little enough of the traditional costumes described. After the foundation of the Turkish republic there was a progressive invasion of European influences, particularly in Turkey itself. Men largely adopted some form of European dress, the women less so, except in Egypt. Wealthy Egyptian women obtained their clothes from the best couture houses of Paris; and even peasant women, if they had not abandoned the veil, wore one of net that hardly concealed the features. On the other hand, the black horsehair veil still worn by the women of Damascus and Baghdad makes it difficult for an observer to decide whether the wearer is coming or going.

Beginning of European influence

The typical Turkish fez is still worn by men in Egypt, the rest of the costume being altogether European. This was the rule among the Turks in the days of the last sultans, who themselves wore frock coats with Turkish headgear. Indeed, the process of Europeanizing had begun in the middle of the 19th century. In 1873 the Imperial Ottoman Commission for the Universal Exhibition of Vienna sponsored an elaborate work containing photographs of men and women in all parts of the Turkish empire, and the pages of this volume reveal clearly not only the variety of costume in the different regions but also the layers of influence out of which they had been built up. In what was then the *vilâyet* of Baghdad, men are shown dressed in the pure Arab mode, while the women wear a recognizably Persian costume. The peasant of the neighbourhood of Jerusalem wears a long white cotton shirt, with a broad leather belt, and over it a garment resembling a Western academic gown, with broad black and white stripes; his fez is wound around with white muslim. A Christian artisan in the region east of the Jordan River wears over his long open shirt, the jubbah, a long, sleeved garment resembling a Western mackintosh. A peasant woman of Damascus is seen covered with jewelry "like an Assyrian idol of which her costume conserves the archaic form." A peasant woman of Lebanon, on the other hand, wears garments of biblical simplicity. The clothes of a woman of Beirut are plainly *à la française*. A Kurdish soldier has a turban like that of the Afghans. In Istanbul the same bewildering variety is apparent. The old-fashioned bourgeois is still clad in the costume of his ancestors; the young government employee wears black trousers and a frock coat—there was nothing Oriental about his garb except the fez. Between these two extremes could be found every combination and permutation possible. Especially after World War II the process of Europeanization was greatly speeded up. (Ja.L.)

THE AEGEAN

Cretan and Mycenaean clothing

Clothing worn by the Cretans in the 3rd and 2nd millennia BC and by the Mycenaeans in the second half of the 2nd millennium was very different from Oriental as well as from later Greek costume. Early in the 3rd millennium in Crete, men wore a short loincloth attached to a tight, broad belt (Figure 8, right). A relief vase from the early 2nd millennium shows a prince wearing long hair, uncovered, and an officer wearing a helmet. From Late Minoan times, a life-size stucco relief showing a king and a wall painting showing young men carrying precious gifts indicate that elaborate loincloths, with woven patterns and embroidered borders and adorned in front with tassels, were worn. Similar are the costumes of envoys from Crete pictured in Egyptian tombs.

On the mainland, in Mycenae, Tiryns, and Sparta, where the Cretan civilization was accepted and imitated in the second half of the 2nd millennium, men wore short tunics with long sleeves for war and the hunt, a helmet with boars' teeth, high sandals, and leggings. The figure of a musician carved on the sarcophagus of Ayía Triádha in Crete (c. 1400 BC) shows that this tunic was adopted there from the Mycenaeans; the lyre player wears a long, sleeveless tunic with a stripe running around the shoulders and down to a decorative border. The priests who bring offerings wear fur skirts.

Costumes of Cretan noblewomen

The costume of Cretan noblewomen is much more complicated, reflecting the luxurious living and gay spirit of court life. A lady of the 3rd millennium wore a bell-skirted dress, the sleeveless bodice of which exposed the breasts and stood away from the nape of the neck. Her cap was high and turbanlike. In the Middle and Late Minoan periods this dress became a tailored, well-fitted, elegant robe. The temple repository in Knossos shows priestesses with bell-shaped skirts made of small strips sewn horizontally together or of different-coloured flounces falling one over another (Figure 8, left). An apron, similar to the loincloth of the men, hangs from a broad, tight belt. The jacket is open in front, but the breasts are covered by a diaphanous vest. Narrow sleeves cover the upper arms. The headdress is a high turban or a crown.

When Cretan civilization was taken over by the mainland, women accepted this rich dress and made it even richer. The best example is provided by the figures of women from a fresco at Tiryns; their dress exhibits flounces (strips of material attached to the garment at their upper edges), some smooth, others with scale and half-moon patterns in different colours; the jacket has borders of flower petals. The hair is elaborately bound up with ribbons and rings, with rows of small curls before and behind the headband, a chignon on the crown of the head from which a ponytail falls, and four long strands falling down the back (Figure 8, left). When women went to the hunt, as shown in another fresco from Tiryns, they wore a simple tunic like that of the men, except that it had short sleeves and no belt. In bull games women wore the apronlike loincloth and high sandals of the men. On the Late Minoan sarcophagus of Ayía Triádha, the priestess, like the priest, wears a skirt made of fur.

In the shaft tombs of Mycenae many golden diadems,

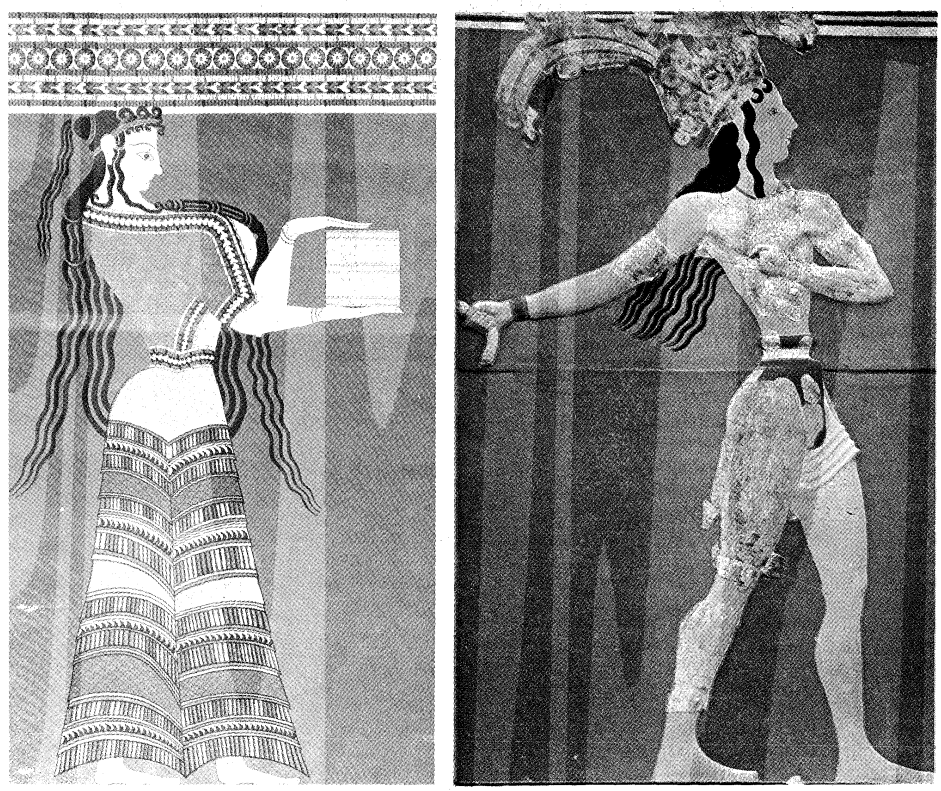


Figure 8: *Middle Minoan period dress.*
(Left) Woman wearing a bell-shaped skirt with flounces, an open jacket, and a diaphanous vest covering the breasts; her hair is intricately dressed with rows of small curls on the forehead and a chignon from which a ponytail falls. Fresco painting, *c.* 1600 BC. In the palace at Tiryns, Greece. (Right) Priest-king wearing elaborate loincloth attached to a tight, broad belt. Fresco from the palace at Knossos, Crete, destroyed *c.* 1400 BC. In the Archaeological Museum, Iráklion, Crete.
(Left) From *Historia del Arte;* photographs, (left) E.D.I. Studio, Barcelona, Spain, (right) Andre Held, Switzerland

breastplates, and buttons have been found, the last used perhaps as jewelry or sewn on dresses as ornaments.

GREECE AND ROME

Greece. The clothing of the Greeks, like their civilization, changed with the three periods of their history: the Archaic (*c.* 750–*c.* 500 BC), the Classical (*c.* 500–323 BC), and the Hellenistic (323–30 BC). In the Archaic period the dress was influenced by the Near East; in the Classical period it was original and harmonious; in the Hellenistic period it was luxurious.

In the Archaic period, overall patterns, as well as figured representations, ornamented fabrics. In the Classical period, patterns were restricted to borders and central stripes. Priestesses wore only white, but other women used many different colours. In the Hellenistic period these became more brilliant, and many mixed colours, particularly violet, were used. Servants and artisans wore dark dresses.

The main dress of Greek men and women in all periods was a linen (later sometimes woollen) shirt, the *chitōn.* Homer mentions the garment often, terming it radiant white and long. Ionian men wore it with a train, and a long *chitōn* was worn by all men in the Archaic period, later only by elderly men, priests, and charioteers (Figure 9, left). After the Persian war, young men wore a short, narrow version. In its simplest form, the *chitōn* was made of two panels of linen cloth seamed up the sides and across the top, with openings left for the head and arms. Actors and priests (for example, the priest of Athena on the central slab of the east frieze of the Parthenon) wore sleeved *chitōn*s.

The *chitōn* worn by women was much wider than that of men (sometimes as wide as the woman's measurement from fingertip to fingertip, with the arms outstretched). Hence the openings for the arms were left at the top, rather than on the sides; in adjusting the garment the wearer brought the two edges together at the shoulder and fastened them with a brooch or brooches along the upper arms. The waist was gathered in with a girdle, the extra length (the garment was about a foot longer than the wearer was tall) making a *kolpos,* or pouch, well illustrated by the caryatids (columns in the form of female figures) of the Erechtheum.

Artisans, warriors, and slaves wore a short *chitōn,* often with the right shoulder left free.

The *peplos*

A garment worn exclusively by women, in the early Archaic and later again in the Classical period, was the *peplos,* a large rectangular piece of wool folded vertically and worn with an overfold at the top in several different ways (Figure 9, right). Every four years the young girls of the best families of Athens, under the guidance of the priestess, wove a new *peplos* for the cult statue of Athena Polias, the city goddess.

Athletic Spartan girls and other young girls wore the *peplos* without a belt and with one side left open so that the leg showed. In most cases, however, a belt was worn, either around the waist or, in the Hellenistic period, much higher, just below the bust. If the overfold of material at the top was long, the belt was laid over it; and it is seen thus in many statues of Athena. If the overfold was short, the belt was placed below it and a pouch made of the excess length. For young girls both overfold and pouch were long, so that the dress could easily be lengthened for the growing girl by shortening overfold, or pouch, or both. When *peplos* and *chitōn* were worn together, the *peplos* served as the overdress.

The *peplos* and the *chitōn* became more and more similar in form with the passage of time. Sometimes the *peplos* was worn without an overfold; sometimes the *chitōn* was worn with one—an example of the latter is that worn by the caryatids of the Erechtheum. In the Hellenistic period very costly materials—cotton and silk, in addition to thin linen and wool—were used. A heavy woolen *chitōn* (*peronātris*) was fixed on the shoulders with clasps (*peronē*).

The *peplos,* being merely a rectangular piece of material, could also be used as a mantle (*himation*). As such,



Figure 9: *The Greek chitōn and peplos.*
(Left) Grecian youth wearing long tunic, or *chitōn*. Charioteer from Delphi, bronze statue, *c.* 470 BC. In the Delphi Museum, Greece. (Right) Woman wearng a *peplos*, a large rectangular piece of wool folded vertically and worn with an overfold at the top, the main dress of Greek women in early Archaic and Classical periods. The "Peplos Kore," *c.* 540–530 BC. In the Acropolis Museum, Athens.
(Left) Toni Schneiders, (right) Hirmer Fotoarchiv Munchen

The *himation* and *chlamys*

in the Archaic period, it was laid symmetrically over both shoulders and the ends were allowed to hang down in front. In the Classical period it was commonly draped in such a way that a long end hung down from the left shoulder in front; the *himation* then was carried back over the left shoulder, across the back, across the right shoulder or hip to the front, across the body, and again thrown back over the left shoulder, arm, or hip (Figure 10, left). Women draped the *himation* over their heads when walking outside the house, as did both men and women in mourning. The many possibilities for style of draping were used by sculptors to reflect the character of the wearer of the *himation*.

A smaller mantle, the *chlamys*, worn by men only, was draped around the upper body and fastened with a brooch on the right shoulder, leaving the right arm free (Figure 10, right). The four corners hung freely on the right side. The two lower ones were rounded during the Hellenistic period.

Men and women generally wore sandals, the soles of which were fastened by leather thongs crossed in various ways and bound above the ankle. Men also on occasion wore high leather boots, and women sometimes wore soft closed shoes. Women's shoes became very luxurious in the Hellenistic period. White and red were preferred colours.

Hair styles and beards

In the Archaic period men wore beards, which were often curled into ringlets with tongs, and long hair. After the Persian Wars they cut their hair short; and in the Hellenistic period they shaved their beards on the order of Alexander the Great, who feared the beards might serve as handles to the enemy. Women often wore their hair parted in the centre and bound up in various ways with ribbons, diadems, and scarves; chignons and loose back hair were common. Little girls wore braids and twisted locks. In the Hellenistic period the so-called melon coiffure, with many partings, like the rind of a melon, was fashionable for women and children. **(M.Bi.)**

Etruria. Whatever the origin of the Etruscans, the people of the region now called Tuscany developed their civilization in the so-called orientalizing period in close contact with Greek and Near Eastern culture. Some differences were due to the great wealth their mineral deposits brought them, although their clothes conformed in general to the Greek fashion of the period. Athletes and young men wore short pants (*perizoma*); older men wore a long *chitōn* covered by the heavy woollen *chlaina*, which was pinned at the shoulder by a brooch, or fibula. There was a short *chitōn*, although this was worn less frequently than in Greece or in the East, and a three-quarter length *chitōn*, more Oriental than Greek. Men wore their hair bobbed, as in Greece. Women also wore a long *chitōn*, with a peculiarly Etruscan mantle hanging straight down the back. Long back braids, known elsewhere in the Mediterranean in Late Mycenaean and

(Left) SCALA, New York, (right) Hirmer Fotoarchiv Munchen



Figure 10: *The Greek himation and chlamys.*
(Left) "Sophocles" wearing large *himation* draped over both shoulders, Roman copy of a Greek statue of *c.* 340 BC. In the Vatican Museum, Rome. (Right) Boy in a *chlamys*, a short men's cloak consisting of a piece of material kept in place by a brooch on the right shoulder. Roman copy of a Greek original of the 1st century BC. In the Istanbul Museum.

orientalizing times, were typically Etruscan. Hats included the simple *pileus*, as in Greece, the *petasos*, and pointed Oriental models. Judging from the rich gold jewelry found in tombs, their taste ran to luxurious necklaces, earrings, bracelets, fibulae, and pectorals.

Around the middle of the 6th century many elements were adopted from the Greeks of Asia Minor (Ionia), probably often by way of Greek colonies in southern Italy and Sicily. The thin linen Ionian tunic (possibly an Etruscan word) or *chitōn* replaced the earlier woollen, straight *chitōn*, although the *peplos* with overfold was

never adopted in Etruria. Pointed shoes became the rage, for men as well as women, from *c.* 550 to 475 BC. The typical hair style for women was the high chignon, or *tutulus*, often covered by a veil, or the Greek kerchief, or *mitra*. Men wore their hair long. Mantles for women remained rectangular, though they were not worn in the contemporary Greek fashion. Men adopted a rounded mantle, the *tebenna*, the forerunner of the later Roman toga. The short version, worn by the Etruscans in Rome at the time of the last Tarquin, was what the Romans later called the *trabea* and adopted as ritual religious dress. Roman ritual costume preserved many fashions of this period (*tutulus*, or *sex crines*, *calcei*, *toga praetexta*, *bulla*), much as modern church vestments preserve the dress of the Middle Ages. From *c.* 450 to 300 BC Etruscan dress conformed more closely to Greek dress, although some typically Etruscan elements remained, such as the rounded mantle for the men and a special *chitōn* for women. By the Hellenistic period, Etruria had been conquered by the Romans. Etruscans wore the standard Hellenistic fashions, except for the *tebenna*. The longer model was worn by the Romans as the toga, symbol of Roman or Italian citizenship. (L.B.W.)

Rome. After the Tarquins were driven out and the Etruscans defeated and brought under the supremacy of Rome in the 4th century BC, the Romans kept the Etruscan dress, at least for their priests (Figure 11, left). After the conquest of Magna Graecia, however, in the 3rd century, Roman women adopted the *chitōn* (Latin *tunica*) with a high belt and *himation* (*palla*). Men also began to wear the *himation* (*pallium*). Nevertheless, the Romans had a number of original forms of dress. The most characteristic feature of Roman dress is its sharp differentiation by social and professional classes.

The tunic, a shirt corresponding to the Greek *chitōn*, was introduced to Italy by the Etruscans (Figure 11, left). It consisted of two pieces of linen sewn up the side and across the top, with holes left for the head and arms. Plebeians, herdsmen, and slaves wore narrow tunics of coarse linen in dark colours; patricians wore fine white wool. For children, the tunic was made wide, with large sleeves. The *tunica angusticlavia*, or tunic with narrow stripes, was worn by children of patricians until 16, when they received a pure white tunic (*tunica pura*). Magistrates and knights wore the *tunica angusticlavia*, senators and other high officers, the *tunica laticlavia* (with broad stripes). Women's tunics had a broad central stripe down the front. At funerals and other rites and by brides, the tunic was worn ungirt (*tunica recta*).

In later times the sleeves were woven in one piece with the body of the tunic, and the decoration became variegated, particularly for children (Figure 11, right). In the 2nd century AD the sleeves became very wide, the stripes often being repeated on them; this garment was the *dalmatica*, fabricated of white Dalmatian wool. It was later taken over as a Christian liturgical vestment.

Development of the toga

The distinctive Roman mantle, the toga, originally Etruscan (Figure 12, top), was worn during the early period by both men and women of all classes. Gradually, it was abandoned by women, then by labouring people, then by the main body of patricians. Throughout the history of the empire, however, it remained the state dress, the garment of the emperor and of high officials of the state. Colour and purple border were rigidly prescribed for most wearers, and the style of draping became extremely complicated—so complicated that rich men had special slaves whose chief duty was handling the toga (Figure 12, bottom). For triumphs and, later, for consuls, the toga was embroidered (*toga picta*). Candidates for office wore an all-white *toga candida*.

Footwear was differentiated according to classes. Women wore closed shoes in white, red, green, or yellow. Men wore sandals, those of the patricians being red with a moon-shaped ornament on the back. Senators wore brown shoes with four black leather straps wound around the leg up to the middle of the calf and tied in two knots. Consuls wore white shoes. Soldiers wore heavy boots with toes free. (M.Bi.)

Roman wigs and beards

In Rome, wigs came into use in the early days of the empire. They were also known to the Carthaginians; indeed, the Greek historian Polybius said that Hannibal used wigs as a means of disguise. Juvenal relates that the empress Messalina assumed a yellow wig, a badge of prostitution, for her visits to brothels. Later, the fashion-

By courtesy of (left) Bibliotheque Nationale, Paris; photograph, (right) Alinari



Figure 11: *The Etruscan and Roman tunic.*
(Left) Etruscan priest in long tunic and cloak, bronze figurine, 5th century BC. In the Bibliothèque Nationale, Cabinet des Médailles, Paris. (Right) Imperial Roman long-sleeved tunic. Statue of Commodus, reigned AD 180–192. In the Vatican Museum, Rome.



Figure 12: *The Etruscan and Roman toga.*
(Top) Etruscan musicians in mantles, precursor of the Roman toga. Detail from Tomb of the Leopards, fresco painting, 5th century BC. In the necropolis at Tarquinia, province of Viterbo, Italy. (Bottom) Roman dress showing complicated toga drapings. Detail from processional frieze on the south side of the Ara Pacis Augustae precinct walls, Campus Martius, Rome, marble, 13 BC.
(Bottom) From *Historia del Arte;* photographs, (top) SCALA, New York, (bottom) E.D.I. Studio, Barcelona, Spain

able ladies of Rome almost universally used hairpieces, and references in the writings of Ovid and Martial indicate that the golden hair imported from Germany was most favoured. Men also occasionally wore wigs for disguise or, like the emperor Otho, to conceal baldness. Women continued to have wigs of different colours as part of their ordinary wardrobe; Faustina, wife of Marcus Aurelius, is said to have had several hundred. Some portrait busts—for example, that of Plautilla in the Louvre—had removable hair, so that by changing the wig a statue would never be unfashionable.

The Romans frowned on the curled beards of the Greeks, considering them effeminate, and preferred a trim, well-groomed shape. The use of the razor was encouraged in 6th-century BC Rome in an effort toward hygienic reform, but shaving did not become general until about 454 BC, when a group of Greek Sicilian barbers came to the mainland from Sicily. Barbershops were situated on the main streets but were patronized only by those who could not afford slaves. The Roman general Scipio Africanus, according to the scholar Pliny the Elder, was the first Roman to shave daily. Philosophers, however, retained their beards. A beard at this time also signified mourning. (M.Bi.)

EUROPE THROUGH THE 18TH CENTURY

Middle Ages. In the early Middle Ages the clothes of both men and women in western Europe showed very little change from those of five centuries before, when Gaulish and Germanic costume had been assimilated by that of Rome, toga abandoned, and trousers—usually loose and cross-gartered—added. Short tunics, sometimes two at once, often made of undyed woollen cloth, were worn by men almost universally. Shoes were simple moccasins, often made of untanned leather. An ample cloak completed the outfit. Such a costume remained that of men of the lower classes for many centuries. Women wore a kirtle, a rather long, shapeless dress concealing the lines of the figure. The sleeves were long and wide, and beneath them could be seen the tight-fitting sleeves of an underbodice. There was very little decoration, and colours were simple—mainly earth colours and vegetable dyes. The hair was entirely concealed by a veil.

Typical dress of the early Middle Ages

Feudal lords returning from the Crusades introduced Oriental luxuries such as silks and damasks (see below *Non-Western dress: China*). Trousers were replaced by close-fitting hose, revealed by a tunic that at first reached to the knees and later to the ankles. The long, full gown had tight, embroidered or bejewelled sleeves; and around the hips was a decorated belt from which hung a pouch, a sword, and a dagger. There was also a straight overmantle fastened around the neck by a cord, chain, or jewelled clasp. Improvements in weaving and dyeing made possible a wider range of colours, and the garments of both sexes began to be trimmed with fur. The typical male headgear consisted of a hood framing the face and forming part of a short cape over the shoulders. The hood was sometimes lengthened at the back to form a point or even a long tail known as a liripipe.

The shaping of garments to the body, for both men and

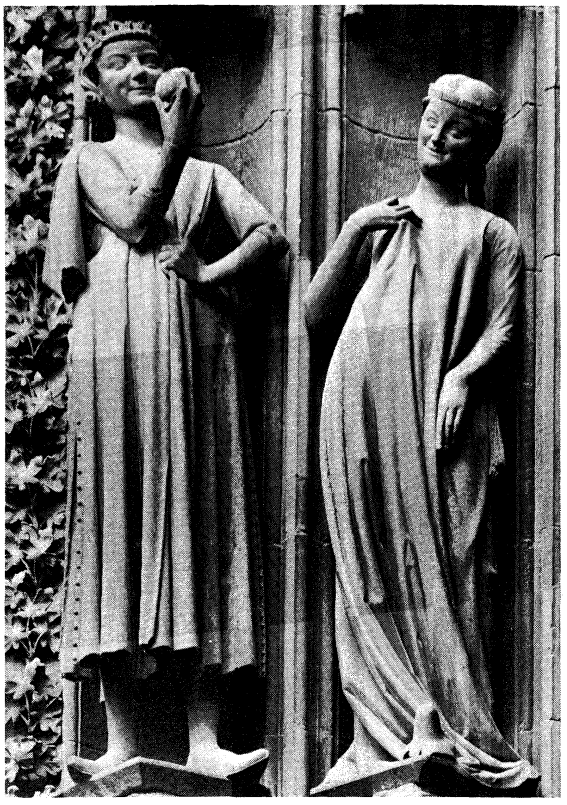


Figure 13: Medieval costume of the High Gothic period exemplified by man (left) wearing full tunic over close-fitting hose and *crackowes*, slim, pointed shoes; woman in a long, shapeless dress over a tightly sleeved underbodice. Statues on the west side of Strasbourg cathedral, France, *c.* 1280–1300.
Archives Photographiques, Paris

Appearance of "fashion" in the mid-14th century

women, began at the end of the 13th century; and by the middle of the 14th something that can be recognized as "fashion" appeared (Figure 13). Men wore a gipon, a kind of short tunic, tight-fitting and buttoned down the front; the sleeves were very tight and long enough to cover the knuckles, the neck was round and low. When the gipon was worn without an overgarment called a cotehardie (a close-fitting, long-sleeved tunic long enough to cover the buttocks), it was belted around the hips. Tight hose were revealed, the two legs often of different colours. The upper garments were sometimes similarly of divided colour and adorned with heraldic emblems; for at this period both men and women often wore their coats of arms on their own persons. Shoes, long and pointed, were known as *crackowes* after the Polish city of Cracow (Kraków); it is thought that the marriage of Richard II of England to Anne of Bohemia (Poland at that time being part of the kingdom of Bohemia) was responsible for this fashion. Gloves, at first worn by kings and bishops as a status symbol, were worn universally among the upper classes in the 14th century.

A characteristic male overgarment from 1380 to 1450 was the houppelande, later called the gown. It fitted the upper part of the body closely but fell below the waist in folds; it was belted around the hips. The neck had a high, upright collar, the edge of which was often "dagged," or cut in fancy shapes. The hood also was dagged; and between 1390 and 1410 a peculiar mode arose of inserting the head into the face opening, winding the folds around like a turban, and binding them in place with the liripipe. The dagged edge thus formed a kind of cockscomb. Hats were also worn with a plume attached by means of a jewelled brooch.

Women continued to wear the kirtle but, like men's clothes, more closely fitted and now décolleté. Sleeves were extremely tight, with many buttons. The cotehardie worn by women had long sleeves sometimes reaching to the ground. Over it, women wore a sideless surcoat, consisting of a front and back descending to below the waist, which remained fashionable for more than a century. In headdresses there was increasing elaboration. Older women continued to wear the veil and the wimple, a piece of cloth covering the chin; but court ladies adopted a kind of circular arch of ridged or pleated linen. Another fashion was to frame the face by means of two pillar-shaped structures of net enclosing the side hair.

After 1400, men were clean-shaven; they were, indeed, almost compelled to be so because of the introduction of a helmet with a chinpiece instead of the cowl of mail usual in the preceding century. They also cut their hair short; and after 1410 the "bowl crop," which looked as if the hair had been cut around a bowl, was usual.

About 1420 the gipon, now called a doublet, began to develop a collar. The cotehardie worn over it was very short at the beginning of the century, but knee length from about 1410 to 1450. The houppelande, or gown, showed considerable variation in the shape of the sleeves. The belt was now usually worn at the waist instead of around the hips. Materials were richer, and there was an increasing use of fur trim. Hose now began to be joined together like modern theatrical tights and were attached to the lower edge of the doublet by "points," or strings. The hood became rare except among the peasantry, the usual form of headgear being the chaperon, a kind of ready-made turban. There was an increasing use of hats, some of them resembling the top hats of a later age. Women's headgear assumed fantastic shapes; the hennin, or steeple headdress (a cone with a veil attached at the top), for example, was fashionable in France and Flanders between 1460 and 1480 (though it never became popular in England). There was also a turban for women and a so-called butterfly headdress of transparent gauze raised above the head by means of long pins.

Increasing use of hats

In the second half of the 15th century men wore, under a very short doublet, a kind of padded waistcoat, known

SCALA, New York

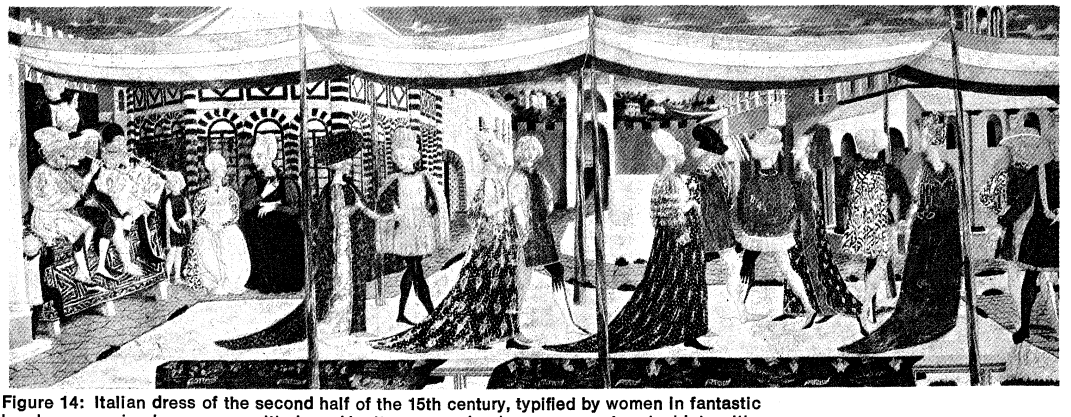


Figure 14: Italian dress of the second half of the 15th century, typified by women in fantastic headgear wearing long gowns with deep décolletages and trains; men wearing doublets with hose attached to the lower edge. Detail of the so called "Adimari Wedding" cassone, Florentine, *c.* 1470. In the Accademia, Florence.



Figure 15: Northern European middle class dress of the first half of the 16th century. Men in small hats wearing furred gowns; man (left) wearing gold chain, a sign of prosperity, and a doublet reaching almost to the knees. "The Ambassadors," oil painting by Hans Holbein the Younger, 1533. In the National Gallery, London.
By courtesy of the National Gallery, London; photograph, A.C. Cooper Ltd.

confusingly as a petticoat. The shoulders of the doublet were puffed and padded. A tall brimless hat, like a Turkish fez, was worn. For women, the sideless surcoat was replaced by the gown, often with a very deep décolletage. Skirts were very long, often with a train (Figure 14).

16th century. The characteristic English fashions of the 16th century were beginning to appear at the time of the accession of Henry VII in 1485. For women, the butterfly headdress was replaced by a low hood lined with fabric of a contrasting colour and turned back in a broad fold. When this was stiffened in the shape of a flat arch it developed into the "Tudor" headdress.

Men were still clean-shaven but wore their hair long. Hats were sometimes adorned with large feathers. Long, pointed shoes gave place to extremely broad shoes. Sleeves, sometimes detachable, also became noticeably wider. They were nearly always puffed and slashed, a curious fashion supposed to have been derived from the costume of German mercenaries. The doublet was longer, reaching almost to the knees, but parted in the front to reveal the codpiece, a pouch at the crotch of the breeches. Under the doublet, men wore a waistcoat, over it, a jerkin or jacket, and over that, a sleeveless and heavily furred gown. All prosperous men wore a gold chain; and this, together with the sleeveless furred gown, has survived as the traditional dress of English mayors. The hat was generally small and worn over one ear. After 1530 it was the fashion to cut the hair short and wear a full beard (Figure 15). Puffed and slashed sleeves

Women wore the kirtle and over it the gown, tight-fitting to the waist and falling in ample folds to the ground. Toward the end of Henry VIII's reign the gown was given a curious bell-shaped sleeve, narrow at the shoulder and opening out to an enormous fur cuff. Trains became shorter after 1530 and 10 years later disappeared entirely. The neck of the gown was cut low and square, revealing the neck of the kirtle and, below that, the top of the chemise.

The English hood, or Tudor arch headdress, was gradually replaced by the French hood, which was smaller and worn farther back on the head, exposing the hair above the forehead.

The German influence that had been dominant in the first half of the century gave place about 1550 to Spanish. The new effect was narrower, tighter, and more formal. The doublet had a narrower sleeve, and its skirts shrank almost to the waist. Puffed trunk hose, frequently slashed in contrasting colours, covered the hips and thighs. The characteristic cloak was short, reaching only to the hips, and about 1560 it was the fashion to wear it attached to one shoulder only.

The most striking characteristic of both men's and women's costume in the second half of the 16th century was the ruff, a frill of folded linen, worn around the neck, that gradually grew larger and larger. It was at first supported by wires, but the invention of starching made these unnecessary. Women's ruffs were worn above or divided in front to show the décolletage. Sometimes worn with the ruff was the rebato, a wired collar edged The ruff

By courtesy of Mr. Simon Wingfield Digby, Sherborne Castle, England

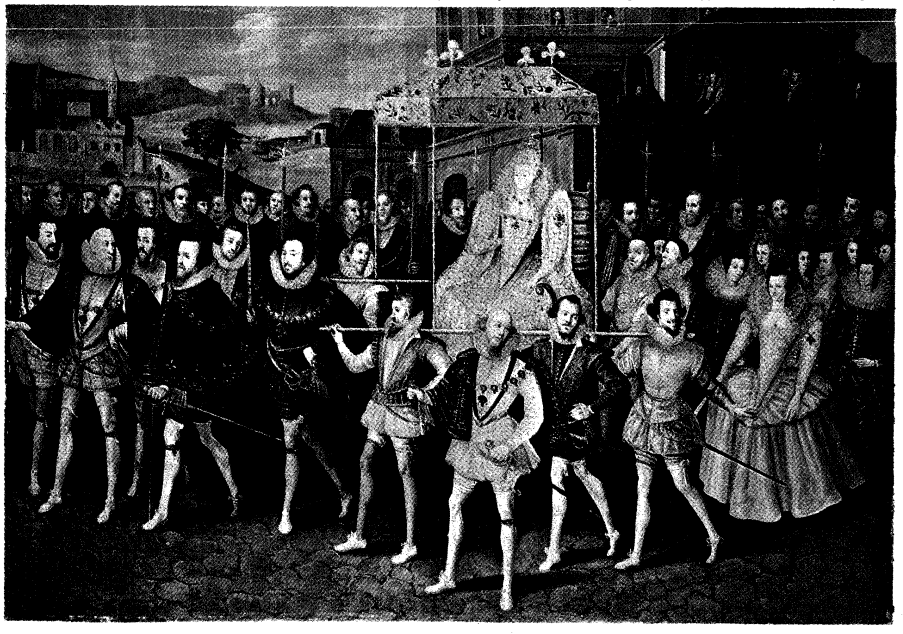


Figure 16: English dress of the second half of the 16th century. Queen Elizabeth wearing a rebato, a wired collar edged with lace; jewels and pearls adorn her neck and stomacher. Men wearing doublets, puffed trunk hose, and ruffs. "The Procession of Queen Elizabeth I," oil painting attributed to Robert Peake the Elder, *c.* 1600. In the collection of Simon Wingfield Digby, Sherborne Castle, Sherborne, Dorset.

with lace and standing up behind the head in the shape of wings. This is the style familiar from portraits of Queen Elizabeth I (Figure 16). Much jewelry was worn by men, chiefly in the form of gems sewn onto the doublet. Women wore stomachers (a jewelled garment worn over chest and stomach) as well as necklaces and ropes of pearls.

Women's skirts grew wider and wider and were kept distended by a roll farthingale (a padded roll worn around the hips) or by a wheel farthingale, which made its wearer look as if she were standing in a large drum. The dresses, however, were sometimes sufficiently short to show the clocked (with the sides ornamented with woven or embroidered decoration) and molded stockings made possible by the invention of the stocking frame. In the last quarter of the century, women wore their hair turned back from the forehead over a pad or a wire known as a palisado. Men generally wore their hair short, and beards were almost universal: pointed, forked, or spade-shaped, sometimes even dyed crimson. Many men adopted the curious fashion of wearing an earring in one ear. The flat caps that had been worn in the first half of the century were replaced among the fashionable by hats of various kinds. Apprentices and others of low degree were ordered to wear woollen caps, but the law was ineffective and was repealed in 1597. Women wore hats only for riding or travelling.

17th century. The modes of the 16th century continued almost unchanged until after the death of Queen Elizabeth I and even during the early years of the reign of James I. The change came about in 1610 with the disappearance of trunk hose and the vogue for longer and tighter breeches. Tall hats ornamented with feathers gave place to low-crowned, wide-brimmed hats. Instead of shoes, there were wide boots reaching to the knee and turned down or half-length boots with wide tops. Ruffs were replaced by wide collars falling over the shoulders and edged with lace. This became the characteristic Van Dyck collar of the Cavaliers; in a plainer form it was worn by the Roundheads also. Women's neckwear showed the same transition, although sometimes the ruff and falling collar were worn together. Women began to wear a square piece of material pinned around the back of the head known as a head roll. The bodice was now short-waisted, with a round instead of a square décolletage, often laced up the front with a ribbon. The sleeves were wide and sometimes puffed and slashed. The skirt, instead of being distended over a farthingale, was now bunched up to reveal a petticoat (Figure 17). Among the men, long hair showed Royalist and short hair Puritan sympathies.

Change from 16th-century modes

The Puritans' costume was similar to that worn by their Dutch contemporaries; the Cavalier costume was essentially French. The Restoration of Charles II was a real victory for French over English fashion. The doublet had become so short that the shirt was revealed between it and the breeches, and its sleeves also were very short, reaching just to the elbow. The breeches, loose at the knee like very wide shorts, were called Rhinegraves. Every item of clothing was lavishly decorated with ribbons. The falling collar had become a kind of bib edged with lace (Figure 17).

Revolution in masculine attire

Then, in 1666, there was a complete revolution in masculine attire. The diarist Samuel Pepys records: "This day the King begins to put on his vest . . . being a long cassock close to the body of black cloth . . . and a coat over it." The other great diarist of the period, John Evelyn, calls it "a comely dress after the Persian mode." The new costume, which by 1670 was firmly established in both England and France, consisted of a long coat with wide, turned-back sleeves and a row of buttons down the front. Some of these were left unbuttoned to reveal a garment almost identical but without sleeves. Neither coat nor waistcoat (as the undergarment was to be called) was provided with collar or rever (portions of a garment turned back to reveal the reverse side), but a cravat of lace or muslin, sometimes with a bow, was worn. Breeches were tight-fitting; and the stockings, gartered below the knee, were pulled up over them. The outfit was completed by a hat that had already begun to be cocked but had not yet become a tricorne, or three-cornered hat (Figure 18, top).

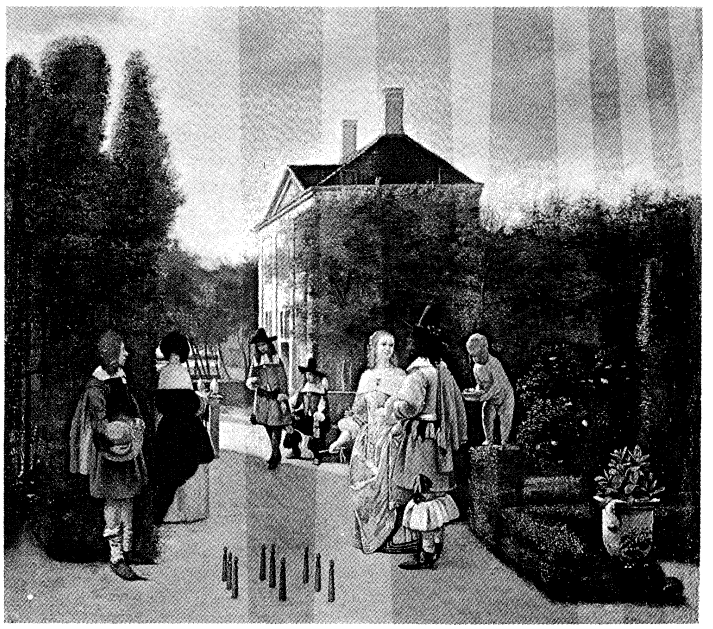


Figure 17: Mid-17th-century Dutch dress. Men and woman wearing wide collars falling over the shoulder. Man (right) wearing Rhinegraves, the woman (centre) a skirt bunched up to reveal a petticoat. "A Game of Skittles," oil on canvas ascribed to Pieter de Hooch, 1660–68. In the St. Louis Art Museum.

By courtesy of the St. Louis Art Museum, Missouri

From Versailles, where Louis XIV began to wear a wig in the late 17th century, the fashion spread throughout Europe (Figure 18, top). After the Restoration in England, under Charles II, the wearing of the peruke became general. Samuel Pepys records that he parted with his own hair and "paid £3 for a periwigg of human hair," and on going to church found that "it did not prove so strange as I was afraid it would."

There was not much change in women's dress during the first 20 years of the reign of Charles II, though women's garments gradually became stiffer and more formal. They aimed at an effect of height; and this tendency culminated in the last decade of the 17th century in the fontange, or topknot, a high lace cap supported on a wire frame.

18th century. Men's dress remained static for almost the whole of the 18th century, the typical costume consisting of a long square-tailed coat, embroidered waistcoat, tight knee breeches, and clump-heeled, square-toed shoes. The three-cornered hat was universal (Figure 18, bottom and 20). Under Queen Anne the wig attained its maximum development, covering the back and shoulders and floating down over the chest. Many varieties were perfected, the cheaper versions being of horsehair. (Smaller, less pretentious wigs, custom-ordered from London, were also worn in the American colonies.) Women, in general, did not wear wigs but had their own hair, sometimes with the addition of false hair, elaborately dressed and powdered.

As the century progressed, the wig and hat grew smaller. The fashion of powdering the hair gave a touch of artificiality and served to distinguish the upper from the lower classes. The waistcoat also grew gradually smaller until it assumed something like its 20th-century form. But there was no real change in essentials until the eve of the French Revolution.

Women's dress was not so constant. Shortly after the death of Queen Anne in 1714, the farthingale returned under another name and in a rather different form, known at first as the hoop skirt and later under the French name of *panier*, or pannier. By 1730 it had become a general European fashion. Panniers were structures of osier reeds or whalebone, rather like a country-

woman's basket, worn on the hips so that the skirt was extended sideways and not all the way around. The extreme width even had an effect on the architecture of the period—for example, the curved balusters of staircases, especially constructed to allow for the passage of women's voluminous skirts. The hair, however, was now dressed close to the head, the towering fontange having been completely abandoned.

(Top) Reproduced by permission of the Trustees of the Wallace Collection, London, photograph, J.R. Freeman & Co. Ltd., (bottom) by courtesy of the Staatliche Schlosser und Garten, Berlin

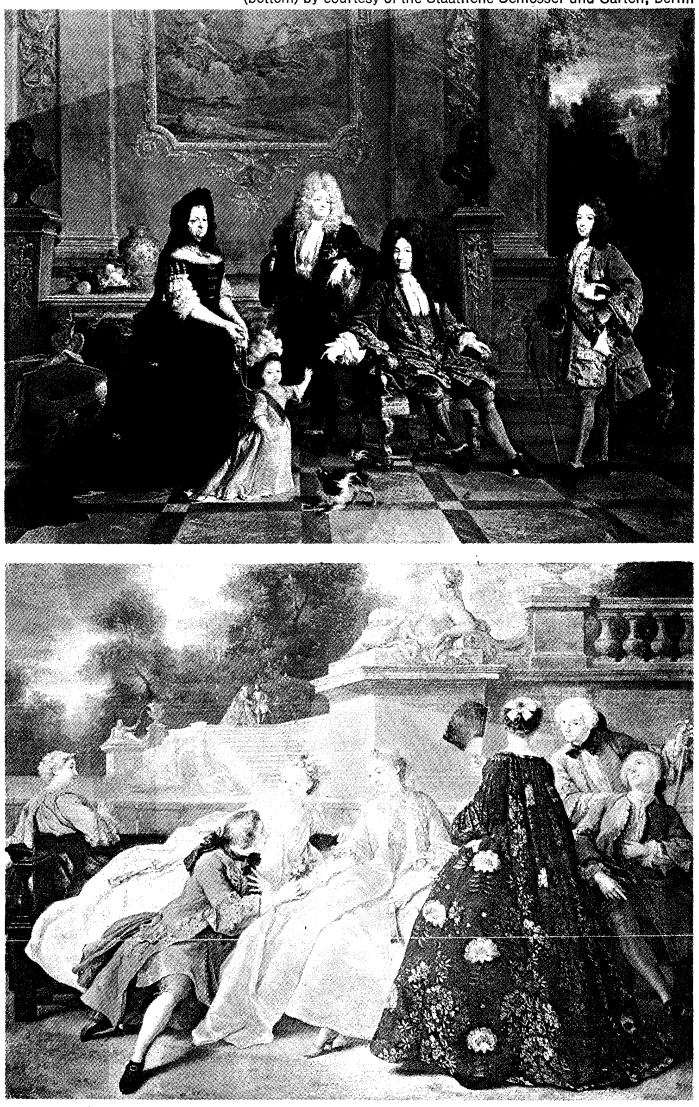


Figure 18: *French dress of the Louis XIV and Louis XV periods.*
(Top) Male attire of long coat with wide, turned-back sleeves, waistcoat, lace cravat, and tight-fitting breeches; the wigs are perukes. "Louis XIV and His Family," oil painting by Nicolas de Largillière, 1711. In the Wallace Collection, London. (Bottom) Women wearing "Watteau gowns," loose dresses worn over tight bodices, with long, vertical pleats falling from the shoulders to the ground. Men wearing powdered wigs, long, square-tailed coats, embroidered waistcoats, tight knee breeches, and stockings. "The Declaration of Love," oil painting by J.F. de Troy, 1731. In the Staatliche Schlösser und Gärten, West Berlin.

"Watteau gown"

A charming fashion of the first quarter of the century was the sacque, or "Watteau gown," a loose gown, worn over a tight bodice, with long, vertical pleats falling from the shoulders to the ground (Figure 18, bottom). At first it was a *négligé*, but by the 1730s it had become the accepted *robe à la française*, or French dress. As dresses grew less formal, the materials of which they were made became lighter; heavy brocades gave place to flowered taffetas and damasks, and even to delicate lawns and dimities. Men's coats, plain and dun coloured at the beginning of the century, began to be embroidered all over or made of patterned velvet. The shirt ended in a frill of lace at the wrist and throat. The silk stockings, now usually white, were no longer rolled over the knees. The three-cornered hat became much smaller, and it was sometimes carried under the arm so as not to disturb the position of the wig.

In women's dress, in 1750, there was a striking decrease in the size of hoops. The triangular-shaped bodice, long and with a tight waist, was cut very low in a square décolletage. The bodice was sometimes laced across the front with ribbons, the corset thus not being a separate garment but forming part of the dress (Figure 19). In cold weather small capes and even large cloaks were worn. Muffs were a popular accessory with both men and women, some of them large enough to hold a small lap dog.

In the 1770s women's hair began to rise, and the general effect was again one of height rather than width. The shape of dresses began to be modified, the side panniers replaced by a pad worn at the back, giving the appearance of a bustle.

Change from court to country modes

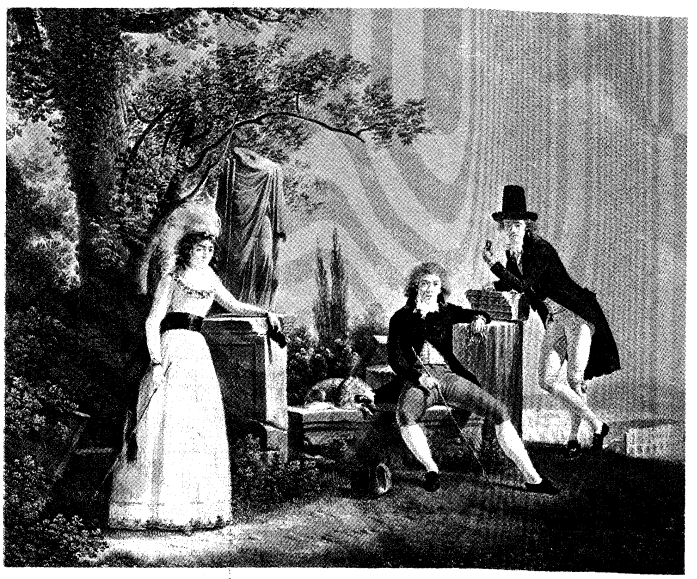
The 1790s saw a fundamental change in both men's and women's costume, essentially a change from court to country modes. Instead of the tricorne, men began to wear a hat which, with its narrow brim and high crown, is the obvious ancestor of the top hat. Embroidery disappeared from men's coats, and the skirts were cut away in front. What resulted was a hunting or riding costume. Boots were substituted for pumps, and swords ceased to be carried. Men began to wear their own hair instead of wigs, and toward the end of the century they gave up the use of powder (Figure 20, top).

Women's clothes showed a new simplicity based upon what were supposed to be classical modes. About 1795 the waist became extremely high and remained so for a quarter of a century (Figure 20, bottom). The materials used were very plain and thin; in France transparent fabrics were occasionally worn. High-heeled shoes were replaced by heelless slippers; and caps were discarded in favour of bandeaux, or narrow bands for the hair. From 1794 to 1797 there was a fashion for wearing two or three large ostrich plumes in the hair; and this fashion, slightly modified, persisted in court dress into the 20th century. Dresses were so scanty that cashmere shawls and wraps of various kinds came into fashion as well

By courtesy of the Brooklyn Museum, New York, Dick S. Ramsay Fund



Figure 19: Eighteenth-century American woman wearing full-skirted dress with pointed bodice, rounded neck, and close-fitting sleeves ruffled at the elbow. "Deborah Hall," oil painting by William Williams, 1766. In the Brooklyn Museum, New York.



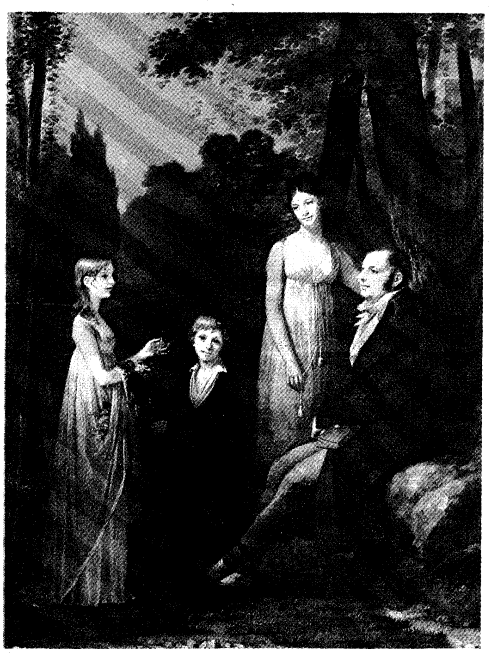


Figure 20: *Late 18th- and early 19th-century dress.* (Top) Woman wearing dress showing new simplicity based upon classical modes; the man (right) wears a hat, which, with its narrow brim and high crown, was the ancestor of the top hat. "Family Portrait," oil painting by François Sablet, *c.* 1794. In the Musée Cantonal des Beaux-Arts, Lausanne, Switzerland. (Bottom) Women wearing high-waisted "Empire" dresses reflecting the influence of ancient Greek and Roman costume; man in long coat, waistcoat, and tight breeches. "The Schimmelpenninck Family," oil painting by Pierre Paul Prud'hon, 1801. In the Rijksmuseum, Amsterdam.
(Bottom) By courtesy of the Rijksmuseum, Amsterdam; photograph, (top) Giraudon

as the spencer, a short tailless coat with tight-fitting sleeves.

Special clothes for children

These modes, both masculine and feminine, had been anticipated by children's clothes a generation before. It was not until the middle of the 18th century that there were any special clothes for children, who were dressed as replicas of their elders except that boys wore their own hair instead of wigs. Like adults, they had full-skirted coats, tricorne hats, and lace ruffles; little girls wore long dresses and were tightly laced, even from infancy. Jean-Jacques Rousseau's theories of education induced many upper class parents to clothe their children more sensibly. By the 1770s, boys were wearing soft shirts and comfortable trousers, and girls loose white dresses with a hint of the high waist. (Ja.L.)

COLONIAL AMERICA

Early period. The settlers of the New World brought with them habits and ideas concerning dress that were characteristic of their places of origin. Basically, however, the dress of the American colonists was completely under English influence. Historical events of 17th-century England had their influence on costume; and plain dress and rich dress became, in effect, the respective symbols of Puritan and Cavalier. Many Virginia colonists leaned toward the Cavaliers; Puritan ideas prevailed in Massachusetts. Virginia dress, though it differed little in design from that of New England, was in general more costly. The Puritans omitted such extravagances as fine brocades, rich laces, ribbons, and feathers; the change to simpler dress that had begun before their departure for America continued.

Probably the greatest change in clothing in America, as opposed to Europe, took place in everyday working costume, the Americans wearing heavier and warmer clothing made of stronger and stouter materials. The distinguishing characteristic of all clothes listed in the inventories of the colonization companies is their wearing quality, and the terms "heavy cloth" and "strong durable stuff" are often encountered. Men and boys wore comfortable, durable jackets and breeches, for example, made from deerskin and buckskin tanned to the consistency of fine chamois with the use of animal brains, a process the colonists had learned from the Indians.

Difference between American and European clothing

Men wore breeches full at the waist and knee; the mandilion, a loose cloak, often lined with fabric of a different colour; doublet and jerkin, somewhat similar garments resembling a jacket, often sleeveless when designed to be worn over another garment; the cassock, similar to doublet and jerkin but longer; and the buffcoat, a strong coat of buff leather, with or without sleeves, tied by a wide sash around the waist. The doublet was especially important, worn by both men and women and by children, its high, tight collar often surmounted by a stiff linen collar or band. It was made of two thicknesses of cloth and welted at the armhole, the welt, or wing, being a piece of cloth set over the armhole where body and sleeves met. The everyday dress of women was a short gown of durable material, with a full skirt over a homespun petticoat, covered by a long apron of white linen. The more stylish dress was longer and made of finer material. It often had the virago sleeve—full at elbow and shoulder and drawn in at intervals by strings of narrow ribbon—that appears in most 17th-century portraits of American women and children.

Women's everyday dress

Slashed clothing was fashionable, as in England; in the openings made by the slashes could be seen rich materials. In 1634, however, the general court of Massachusetts forbade men and women to make or buy clothes with more than one slash in each sleeve and another in the back. The starched ruff of the early 17th century gave way to a falling band, the common form of which was a broad, plain, linen collar. Both men and women wore this collar and plain linen turnback cuffs.

Stockings were either knitted or cut from woven cloth and sewn to fit the leg. They were attached to men's breeches by points, or strings, which were also used to secure other garments; later, sashlike garters replaced points. Both men and women wore stout leather shoes with medium heels. Men also wore French falls, a buff leather boot with a high top wide enough to be crushed down. After 1660 the jackboot, a shiny black leather boot large enough to pull over shoe or slipper, replaced the French falls; oxfords of black leather were worn by school children.

Headgear

Both men and women wore a steeple hat of felt or the more expensive beaver. Men also wore the montero cap, which had a flap that could be turned down, and the Monmouth cap, a kind of stocking cap. Women of all ages wore a French hood, especially in winter, when it was made of heavy cloth or fur-lined; this hood, tied loosely under the chin, is seen in many portraits of the time. Sometimes the steeple hat was worn on top of the hood.

As colonial wealth increased, so did refinements in

clothing. Many portraits of American women in the last half of the 17th century show costume as stylish as that worn in England. The skirt of the woman's gown was gathered to the bodice and hung full to the floor or was open in front to show the petticoat; sometimes the sides were caught together in the back. This increasing back fullness developed into the bustle, which lasted until about 1711, when the hoop came in. Small, elegant aprons became fashionable.

In the 1660s, as in France and England, the doublet, jerkin, and cassock were gradually transformed into the coat, which became the general style of man's attire. In the last quarter of the 17th century, the characteristic men's clothing of the 18th took shape: the periwig, the tricornered hat, the ruffle of lace or sheer linen at throat and wrists, the long waistcoat, the skirted coat with wide, turnback cuffs, tight breeches, sash garters, and buckled shoes.

The rise of "fashion"

18th century. Fashion was a noticeable element in the first half of the 18th century in America. No laws restricted fine dress in this prosperous period, and merchant ships brought to all the seaport towns elegant fabrics as well as other luxuries. Dresses and suits of homespun, stout shoes, and durable cloaks were worn as before; but those who could afford to do so had stylish dress. There was no waste of fabric, however. The cherished brocade dress, the fine suit, and embroidered waistcoat were often bequeathed by will to the second and even the third generation. There were no fashion plates (illustrations of clothing styles), but jointed dolls dressed in the latest styles were sent at regular intervals from London to the American colonies.

The most fashionable woman's dress during most of the 18th century was the sacque, or "Watteau gown" (see above *Europe: 18th century*), whose ample folds displayed to full advantage the costly material of which it was often made. The commonest woman's dress, however, one worn with variations both before and after the sacque, was a dress with a long-waisted, pointed bodice, a full skirt, a rather low, rounded or square neck, and a close-fitting sleeve with a ruffle at the elbow (Figure 19). Stays, or corsets, were a necessary article of dress at the time.

The woman's dress shoe was a brocade slipper with a French heel and usually a buckle. Often the pattern, or overshoe, worn to shield the slipper, was made of the same material. As in the 17th century, cloaks were worn extensively, called by such names as pompadours, roquelaures, capuchins, and cardinals. The cardinal, as the name suggests, was a hooded scarlet cloak and was worn during the entire century.

Women's headgear

Women's hoods, which previously had been made ordinarily of black silk or heavy cloth, now came in fine fabrics and gay colours. The calash, a great bonnet that resembled the extension top of a *calèche*, or French carriage, was worn with a high, pompadour style of hairdress. This head covering, usually of thin silk, with whalebone or rattan run at two- to three-inch (five- to eight-centimetre) intervals through shirrings, could be pulled over the face or pushed back. The mobcap, a huge white lace or lawn cap with deep-hanging frills, was also worn. Gypsy hats and the skimming-dish hat, or skimmer, both of leghorn straw, were worn after 1750.

From 1770 to 1780 the polonaise gown was popular. The bodice was fitted as before; but the full skirt was curved away from the front and caught up on either side at the back so that it fell in three large loops, showing an elaborately quilted or embroidered petticoat. A long apron of sheer, embroidered linen was worn with the open-front gowns of this period. Gold beads and fans for women and snuffboxes for men were important accessories.

Little had happened to the costume worn by men at the close of the 17th century except for a shortening of the waistcoat, a decrease in the size of the turnback cuffs and the wig, and the replacement of the cravat by the stock, which fastened at the back so that the frilled edge of the shirt could be seen at the open front of the waistcoat (Figure 21).

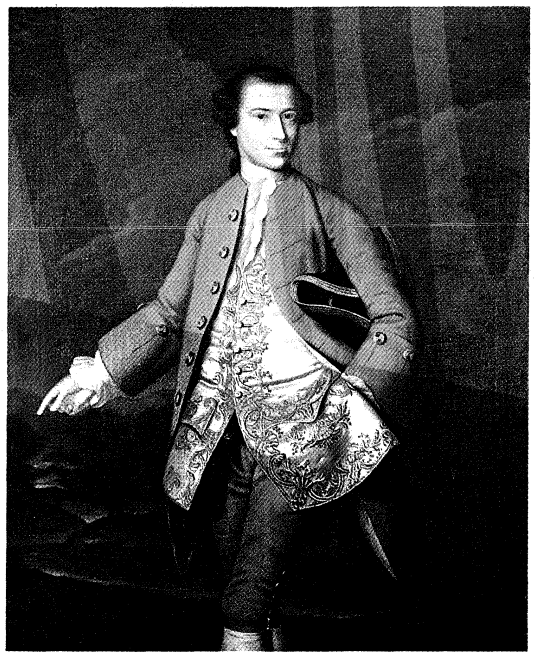


Figure 21: Eighteenth-century American man wearing a wig, ruffle of sheer linen at throat and wrists, long waistcoat, skirted coat with turn-back cuffs, tight breeches, and stockings. "Portrait of Theodore Atkinson, Jr.," oil painting by John Singleton Copley, 1757. In the Museum of Art, Rhode Island School of Design, Providence.
Museum of Art, Rhode Island School of Design, Providence

After 1760, breeches were fastened at the knee with buttons and a small buckle; and the front of the coat, which was developing a collar, was cut away at the sides. Unpowdered, natural hair became popular for both men and women. (A.W.M.)

EUROPE AND AMERICA, 19TH AND 20TH CENTURIES

The Regency dandy

Early 19th century. The dress of the Regency dandy was a smartened version of European late 18th-century country clothes. Beau Brummell, whose clothes were copied by the Prince Regent himself, was so concerned with fit that he had his coat made by one tailor, his waistcoat by another, and his breeches by a third. The neckcloth was so elaborate and voluminous that Brummell's valet sometimes spent a whole morning getting it to sit properly. It was at this period (*c.* 1811–20) that English modes for men became everywhere accepted as correct, even in Napoleonic France (the top hat, for example, became almost universal).

France continued, however, to dictate women's fashions, except for an interlude between 1802 and 1814, when communication was difficult and English dresses diverged noticeably from those worn in Paris. After the abdication of Napoleon in 1814, however, English women's clothing immediately reverted to Parisian styles. The French waist was still high, but the skirt had become wider at the hem and was rather heavily decorated at the lower edge to make it stand out. The invention of the fashion plate meant that fashions now filtered down the social scale more quickly. The only real change in men's dress after 1814 was that trousers began increasingly to replace breeches; in summer they were usually made of white duck.

Women's dress

In 1820 women's waistlines resumed their normal position, becoming steadily tighter and tighter. Skirts continued to expand, sometimes weighted at the lower edge with a band of fur. Colour came back again after a long eclipse, and there was a vogue for tartans. A little puffed and sometimes slashed sleeve began about 1825 to be worn with another transparent sleeve over it. When this was made opaque, the sleeve assumed the leg-of-mutton shape so characteristic of the period. After 1830, while skirts grew shorter, sleeves became enormous. Hats were also extremely large and ornamented with flowers and ribbons.

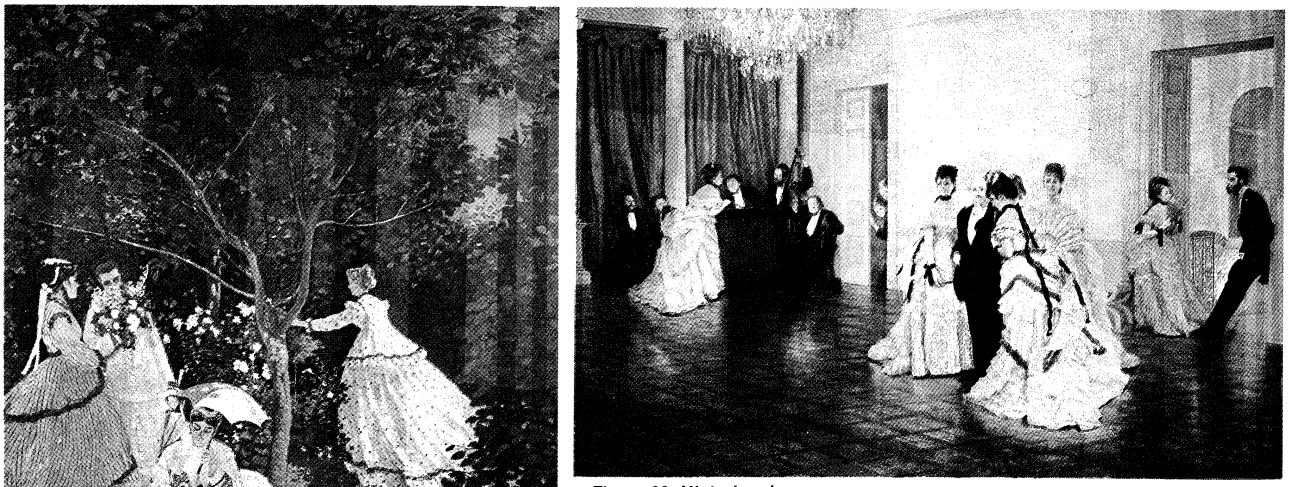


Figure 22: *Victorian dress.*
(Left) Women wearing light summer dresses with crinolines. "Women in the Garden," oil painting by Claude Monet, 1866–67. In the Louvre, Paris. (Right) Women wearing bustles and men in cutaway coats. "Too Early," oil painting by James Tissot, 1873. In the Guildhall Art Gallery, London.
(Right) By courtesy of the Guildhall Art Gallery, London, photograph, A.C. Cooper Ltd., (left) Giraudon

Victorian age. With the accession of Queen Victoria in 1837 these romantic modes began to be modified. The wide sleeves disappeared or, rather, the bulge was placed farther down the arm. Large hats, indeed hats of any kind except turbans, were abandoned in favour of the poke bonnet. Shawls were universally worn. The general effect was subdued and modest. On the other hand, evening dresses developed a wide décolletage that exposed the shoulders; this early Victorian style, which would have been thought extremely improper in the 18th century, was accepted without question.

The increasing amplitude of skirts in the early 1850s caused a revival of the 18th-century hoop in an improved form—the crinoline, a stiff fabric made of horsehair and linen thread (Figure 22, left). This item was soon replaced by a series of flexible hoops, sometimes forming a separate structure and sometimes sewn into the petticoat.

Men's clothes at this period had a somewhat sombre look. The top hat had become the shiny silk hat, usually black. The clothes, except for a surviving fancy waistcoat, were also black, although shepherd's plaid trousers were occasionally worn. The cutaway coat was now worn as evening dress only (Figure 22, right). High neckcloths had been abandoned for collars and ties not very different from those worn in the 20th century.

Crinolines and bustles

The crinoline reached its largest dimension about 1860. After that it began to slip to the back and was indeed no more than a half crinoline. Then, about 1869, it was replaced by the bustle, although hoops were sometimes still employed (Figure 22, right). Hats, worn tilted over the forehead, replaced bonnets. In men's clothes there was a steady evolution toward less formal modes, a short black coat sometimes replacing the frock coat or morning coat. The bowler hat, invented in the 1850s for country wear, began to be worn in town. The suit, with coat and trousers made of the same material, was worn by younger men.

As the 1870s progressed, the bustle was gradually reduced until it was no more than a fullness at the bottom of the skirt. Almost all dresses had long trains, even the street dresses of fashionable women. In 1880 the fullness of the skirt disappeared for a brief period; dresses were not only smooth over the hips but fairly narrow around the hem. The waist was very tight and the general effect was of a corset worn outside the dress. The bodice was high-necked and very closely fitted. By 1884, however, the bustle was back again in a slightly different form, stretching straight out at the back like a shelf. Hats were small, perched squarely on top of the head.

Not all of the women of the 1880s, however, wore these fashionable clothes. Followers of the Aesthetic Movement in England wore looser garments—though the waists were still tight—with enormous sleeves supposed to resemble those worn by women in early Florentine paintings. The humorous journals of the period made great play with the contrast between fashionable and Aesthetic modes.

Invention of men's sports clothes

For men, the 1880s provided a certain relief from formality by the invention of various kinds of sports clothes. Men took to knickerbockers and tweed jackets, straw hats and "deerstalkers" like that immortalized by Sherlock Holmes. Women also wore straw hats and bodices of naval cut to match the double-breasted "reefers" of the men. In the 1890s women began to wear knickerbockers for cycling, a real revolution in female attire. They also thought it necessary for this purpose to adopt men's stiff collars and trilby hats (Figure 23, top).

In 1890 sleeves rose to a point on the shoulders, and this feature expanded until it produced the characteristic balloon sleeves of the middle of the decade, familiar from the drawings of Aubrey Beardsley and the posters of Toulouse-Lautrec. Muffs, almost universally worn, were so small that they could be carried on one forearm, the other hand being perpetually occupied in lifting the dress. In evening dress, décolletage, which had been modest in the 1870s and 1880s, became more daring.

Children's clothes were less sensible and comfortable than they had been 50 to 60 years before. What had started in the 1820s as rational dress for boys had been formalized into the rigid discomfort of the Eton suit with its stiff white collar. Fortunate boys were dressed in sailor suits and unfortunate ones as "Little Lord Fauntleroy," in velvet suits with lace collars and cuffs and with the hair dressed long in curls. Little girls were dressed in elaborate and easily soiled garments with much lace. Their skirts were shorter than those of adult women, but the waists were nearly as tight.

As the century drew to a close, the wide sleeve slipped and became a bulge over the forearm, facilitating the wearing of the short capes then fashionable. A prodigious quantity of lace was worn, at the wrists or in the form of a frilled shirt front attached to the bodice. For informal wear, blouses had become popular, these too adorned with frills of lace.

Men's daytime clothes had divided into formal and informal wear. Formal wear consisted of a long frock coat with silk lapels, white or gray waistcoat, striped trousers, an all-the-way-round stiff white collar and a silk (top)

hat. Informal wear consisted of a checked suit or knickerbockers and a Norfolk jacket, and a straw hat. Cloth caps also were worn. Even with this costume, however, it was usual to wear a stiff collar. Men's evening dress had also divided into a formal tail coat and an informal dinner jacket, a stiff shirt being worn with both.

Beards and moustaches

From about 1840 to 1870, side whiskers attained a longer, more luxuriant growth and were sometimes referred to as "mutton chops" or "Piccadilly weepers." In the United States they were called "dundrearys" after Lord Dundreary, a character in the play *Our American Cousin*. The name "burnsides," or "sideburns," came from the name of a U.S. Civil War general, Ambrose Burnside. The "imperial," a pointed tuft of whiskers on the chin, was named in honour of Napoleon III. Long side whiskers merging into a moustache became known as the "Franz Josef" in honour of the Emperor of Austria. In the 1880s the trend was toward the clean-shaven, although beards and moustaches were retained by older and professional men. In the British Army the "walrus" moustache became popular; it was forbidden to shave the upper lip. In the U.S., a moustache that curled upward at the ends, called a "handle bar," found great favour. Professional men favoured the pointed Vandyke beard. Eventually, whiskers became a distinguishing mark of butlers and footmen, who wore them in the 20th century.

Early 20th century. The first decade of the 20th century showed comparatively little change in the essential lines of feminine costume. The mature type of women was admired: tall, small-waisted, heavy-bosomed, and with a peculiar stance—the S-shaped look, resulting from what was strangely known as the "health" corset, a corset boned in such a way as to throw the hips back and the bosom forward. The effect was accentuated by the Russian blouse and cascades of lace descending from the bust. Never since the days of William III had so much lace been worn. Lace collars and collarettes, lace sleeves, lace plastrons (trimming on the bodice), lace overbodices, and lace petticoats, only to be glimpsed occasionally, but requiring the finest workmanship—there was hardly any part of women's dress that was not adorned with this expensive form of decoration. Real lace in such quantities being unobtainable and machine-made lace somewhat despised, a compromise was discovered in Irish crochet, for which there was a considerable vogue, especially in 1907. Skirts were long and sweeping, with a curious resemblance to the dominant line of *art nouveau* furniture. The hair was piled high on the head, and a flat pancake hat projected forward as if to balance the trailing skirts (Figure 23, bottom).

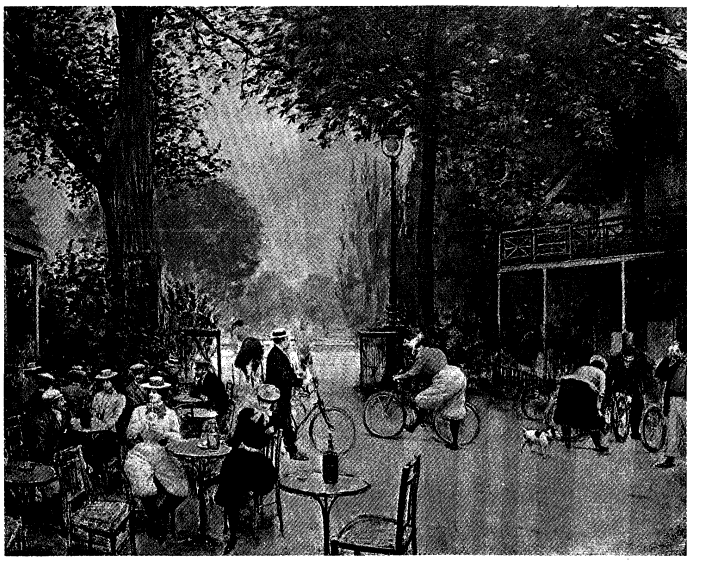
Introduction of pajamas

A revolution in nightwear took place in the early years of the century. For 300 years or more, women had worn in bed a long smock and men a longer version of the day shirt. Pajamas now began to take their place. They are thought to have originated in India (*pājāma* in Hindi means "drawers") and, once introduced into the West, became steadily more popular not only for men but also for women.

In 1908 a slight modification began to be discernible in women's clothes. The bust was no longer thrust quite so far forward as before; the exaggerated overlap of the blouse was abandoned, and skirts became a little narrower at the hem, though they still trailed on the ground and had to be gathered up in the hand when crossing a wet or muddy street. The change resulted, in part, from the coming of what was called the empire gown, though it did not resemble very closely the high-waisted dress of a century before. Its effect was to make the figure appear tubular; by 1910 the process was complete. At the same time hats became extremely wide and the wide skirt was replaced by a very narrow one.

The hobble skirt

The hobble skirt of 1911, one of the strangest garments ever worn by women, shackled the legs so completely that walking was almost impossible. The feminine silhouette, as so often, resembled a triangle, but a triangle standing on its apex. With the enormous hats went very large muffs; and the handbag reappeared, since it was impossible for such narrow skirts to contain pockets. In the draped effect of evening dresses there is a curious Orientalism, due in part to Paul Poiret, the most fashionable dressmaker of the period, and in part to the immense success of the Russian ballet. For day wear the tailor-made, or suit, was very popular. Fashionable trimming consisted of buttons sewn in the most unlikely places all over the costume.



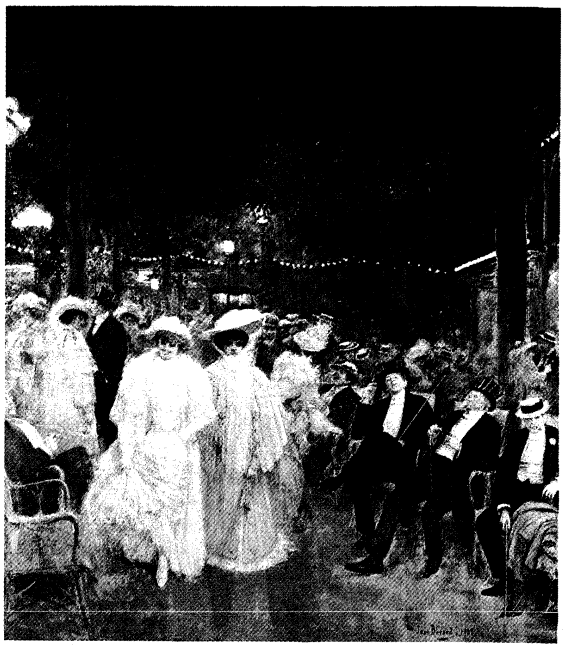


Figure 23: *Turn-of-the-century fashions.*
(Top) Women wearing knickerbockers, men's stiff collars, and trilby hats for cycling. "The Cycle Hut in the Bois de Boulogne," oil painting by Jean Beraud, *c.* 1901–10. In the Musée de l'Ile de France, Château de Sceaux. (Bottom) Women dressed in cascades of lace, with corsets that threw the hips back and the bosom forward in the S-shaped look; men in long frock coats, white waistcoats, all-the-way-round stiff white collars, and silk top hats. "Jardin de Paris. The Night Beauties," oil painting by Jean Beraud, 1905. In the Musée Carnavalet, Paris.
(Bottom) By courtesy of the Musee Carnavalet, Paris; photograph, (top) Cliche Flammarion/Musee Chateau de Sceaux ,France

A kind of tunic overskirt made its appearance in 1912, and by early 1914 had become the feature of dress upon which the attention of designers was concentrated (Figure 24). It was to provide the jumping-off stage for the development of the next mode, which involved a fundamental change in the neckline. The new fashion was known as the V-neck and was considered by many to be dangerous to both health and morals. Men's clothes remained extremely formal.

The outbreak of World War I in 1914 did not bring



Figure 24: "Afternoon Dress with Hoop Tunic," by Bernard Boutet de Monvel, 1914. In the collection of François Boucher, Paris.
In the collection of F. Boucher, Paris, permission A.D.A.G.P. 1972 by French Reproduction Rights, Inc.; photograph, Flammarion

about any immediate change in women's dress; but about the middle of 1915 the narrow underskirt was abandoned and a lampshade tunic became the whole dress. This skirt reached to mid-calf, revealing high laced boots. Hats were now small and were usually trimmed with a vertical feather. The shortish skirts and high boots remained in fashion with only slight modification until the end of 1918.

Post-World War I. By 1921, the skirt fullness of the war period had disappeared, and waists had disappeared with it. The general effect was tubular, the corset having been completely abandoned. The extremely low waist was indicated only by the lower edge of the now inevitable jumper. Hair was bobbed, but the close-fitting, bell-shaped cloche hat had not yet appeared. Hoop skirts, or panniers, made a brief appearance in 1923 but did not long affect the prevailing trend. It was not until 1925 that the extremely short skirt made its appearance. With

The tubular look

it were worn silk stockings in various shades of flesh pink, a startling innovation.

Skirts became even shorter in 1927, when it was impossible for a woman to sit down without showing her knees (Figure 25, left). The short skirt lasted for day wear until 1930, but evening dresses, by careful degrees—tails, trains, sidepieces, and transparent hems—became long again. In 1930 day skirts suddenly lengthened to mid-calf or lower and the waistline moved back to its normal position. There was a temporary vogue for large sleeves, not, as in 1830 and 1895, to decrease the apparent size of the waist, but to make the hips look smaller. The emphasis throughout the 1930s was on slim hips and backless dresses, and sometimes the whole outfit looked as if it had been designed to be seen from the rear. The shoulders were heavily padded to make them look square. The long reign of the cloche hat came to an end, to be replaced by a variety of very small hats that were worn perched forward over one eye (Figure 25, centre).

Movement toward more informality in men's clothes

Men's clothes showed a steady movement toward more informality. The frock coat disappeared altogether, and the morning coat and silk hat were worn only on ceremonial occasions. Wide trousers known as "Oxford bags" came in 1924, and trousers remained wide until the end of the 1930s. Soft collars replaced the stiff white linen variety; and for golf and other sports men wore baggy knickerbockers known as knickers or plus fours, often with a gaily coloured sweater.

Women's sports clothes became more and more scanty, and it became usual to play tennis in shorts or very short skirts, without stockings. Swimsuits were extremely scanty.

Two marked general tendencies may be noted: the disappearance of class distinction in women's day clothes and a marked divergence between day clothes and evening clothes. Many women were now engaged in some kind of work away from home during the day, and there was evolved for this purpose a kind of working uniform, consisting of the tailor-made, or suit. Trousers, in the form of slacks, were increasingly worn for sports but not yet for shopping. On the other hand, even girls in the lower income groups assumed for evening some kind of "glamour" dress, a dress based ultimately on what had been seen in motion pictures. This influence tended to inhibit change over the years, since as much as two years might elapse between the shooting of a film and its gen-

VOGUE covers Copyright © (left) 1928 (renewed 1956), (centre) 1931 (renewed 1959), (right) 1947 by The Conde Nast Publications Inc.



Figure 25: *Women's fashions from the 1920s to the 1940s.*
(Left) Tubular, long-waisted, short-skirted dress. *Vogue*, 1928. (Centre) Slim-hipped daytime suit showing skirt at mid-calf and the waistline in its normal position. *Vogue*, 1931. (Right) Christian Dior's "New Look," with narrow shoulders and long skirt. *Vogue*, 1947.

eral release. The "glamour" toilettes of the 1930s are therefore curiously static and tend to have the appearance of fancy dress.

World War II. Just before World War II there was an attempt on the part of dress designers to bring back the boned corset. Fashion commentators spoke of "the figure of eight which the new clothes demand; so get it, by hook, by crook, or by corset." Apart from this trend, there was a wild variety of styles, especially in evening dress. There were hooped skirts and hobble skirts (a kind of throwback to the fashions immediately preceding World War I) and even harem skirts. But these developments were cut short by the war.

By the summer of 1939, day skirts were almost as short as they had been in the late 1920s. Shoes were very square-toed with low heels. When war came, the little hats that had served for a decade were often replaced by head scarves and, in wet weather, by plastic hoods. Housewives as well as factory workers took to slacks, thereby making stockings unnecessary. Even with skirts, many women began to go about with bare legs, on which stockings were sometimes painted with a line drawn down the back to simulate the still obligatory seam. Clothes rationing was introduced in Britain in June 1941, and for the rest of the war fashion may be said to have ceased to exist in that country. In the United States, rationing of clothing worked little hardship, but restrictions on uses of fabrics (leading to the virtual disappearance of silk and nylon) affected clothing. The daytime silhouette for women continued to be broad-shouldered and short-skirted. The short dress for evening reappeared. The hair was worn upswept in the back, with a high pompadour in front, or else hanging loosely to the shoulders or below.

Post-World War II. When communication with France was restored in 1945, there was a cry that "fashion is going feminine," but in fact there was at first very little change. Day dresses had more rounded shoulders and a slightly smaller waist; evening dresses sought to revive the modes of 1939. Then in the spring of 1947 Christian Dior launched a "New Look" from Paris (Figure 25, right). The shoulders were narrow, there was a new emphasis on the bust, and the skirt was much lengthened and had a wide, billowing hem. Hats, after their wartime eclipse, returned to popularity, with much experiment in form and colour. Brims, in general, were large and flaring or small and turned back against the crown.

The "New Look"

A popular development in postwar fashions was the late-afternoon cocktail dress, which by means of a minor adjustment, such as discarding a bolero, could be made to serve as an evening dress also. There was a certain revival of such Victorian modes as the crinoline for formal occasions. It was as if women were unconsciously striving to return to a more settled age. There was a marked emphasis on the bosom, the desired effect sometimes obtained by means of a padded or inflated brassiere. Soon, however, the "New Look" gave place to what seemed like a revival of the modes of the 1920s, with the straight lines and displaced waists of typical postcrisis periods. Dior brought out an H-line and then an A-line, and it seemed as if the designers were unable to decide whether to echo the fashions of 1800 or those of 1925. By the early 1960s, they appeared to have made up their minds —fashionable dress resembling very closely that of the '20s, with a lowered waist and very high hemline. Shoes were excessively pointed, with stiletto heels. In the servantless postwar world, women often wore slacks for housework and shopping; in some communities long shorts (Bermudas or Jamaicas) replaced slacks; and culottes, or divided skirts, reappeared.

Men's clothes after the war showed a curious reversion, not to Victorian but to Edwardian modes: tighter trousers, coats buttoned higher, and the revived bowler hat. There was a passing craze for fancy waistcoats. Slacks, sport shirt (worn without a tie), and a jacket resembling battle dress, together with a variety of sweaters (jerseys in Great Britain), became ordinary wear for many young men. In the evening the tailcoat almost entirely disappeared, but the dinner jacket maintained its popularity.

Revival of the Edwardian look for men

(Ja.L.)

The fashion revolution that began on London's Carnaby Street in 1957 with the introduction of the "Mod" look led to the permissive, youth-oriented, and anti-establishment fashions of the 1960s, when styles never moved so swiftly, had so many contradictions, or displayed such variety. If the result was easily natural, individual, and unrestrained, it was acceptable. A preference for making your own fashion became current and a taste for ethnic, nostalgic, erotic, or "far out" effects were "in." This was the era of the erratic hemline (mini, midi, and maxi); pants suits, hot pants, and short shorts; boots and platform shoes; and flared or bell bottom trousers. Borrowing freely from the opposite sex, unisex fashions came into vogue by the end of the 1960s. Women often

Pictorial Parade Inc.

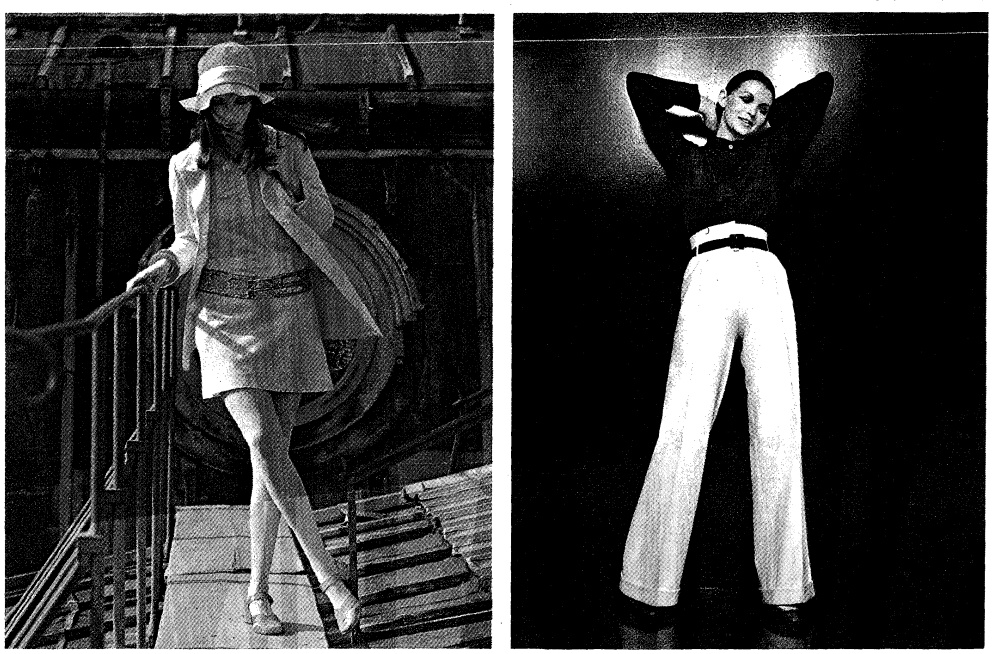


Figure 26: *Dress in the 1960s and '70s.*
(Left) Mini-skirted outfit with clump-heeled shoes, 1969. (Right) Trousers by Yves Saint-Laurent, Paris, 1970.



Figure 27: *Men's dress, early 1970s.*
Men wearing "mod" suits, wide-legged trousers, broad ties, and a trench coat with casual turtleneck sweater, 1971.
Reprinted, courtesy of the *Chicago Tribune*

adopted such masculine attire as military surplus and work clothes and the stylish male had long hair, carried a pocketbook, wore jewelry, and used a wide variety of cosmetics. By the early 1970s, however, there was a return to traditional elegance and classical styling. The layered look (adding garment to garment) was popular and fashion design was often inspired by the styles of the 1930s and the early 1940s. (Ed.)

II. Non-Western dress

Western-style clothes, which many persons prefer to wear during business hours, are now a common sight in many of the large cities of eastern and southern Asia; but they do not dominate the scene and are usually exchanged for traditional dress when the day's work is done. In Korea and especially in Japan, traditional styles of dress reflect marked Chinese influence, though both countries developed characteristic styles of their own. In like manner, modes of dress in the Indian subcontinent have been a source of inspiration to some of the countries of Southeast Asia and of the East Indian archipelago.

EAST ASIA

China. More than 2,000 years before the beginning of the Christian era, the Chinese discovered the marvellous properties of silk and shortly thereafter invented looms equipped with devices that enabled them to weave patterned silks rapidly enough to satisfy the demand for them by luxury-loving Chinese society. Thus, centuries before Chinese silks began to be shipped westward and still more centuries before the West learned the secret of sericulture, the people of China had already established ultrarefined standards of elegance in matters of dress.

The earliest period of Chinese history for which reliable visual evidence of clothing styles is obtainable is the Han dynasty (206 BC–AD 220). Han bas-reliefs and scenes painted in colour on tiles and lacquers show men and women dressed in wide-sleeved kimono-style garments which, girdled at the waist, fall in voluminous folds around their feet. This *p'ao*-style robe continued to be worn in China until the end of the Ming dynasty (1644). The graceful dignity of post-Han *p'ao* is revealed in Chinese figural paintings of the 8th to 13th centuries AD (Figure 28) and also in the modern kimono of Japanese women; for this Chinese style, introduced into Japan in the 8th century, has been little changed in the country of its adoption during the intervening centuries. Even today the kimono continues to be worn and is commercially very successful.

The *p'ao*

The Chinese records indicate that at least as early as the T'ang period (618–907) certain designs, colours, and accessories were used to distinguish the ranks of Imperial, noble, and official families; but the earliest visual evidence of these emblematic distinctions in dress is to be found in Ming portraits. In some of these, emperors are portrayed in voluminous dark-coloured *p'ao* on which the 12 Imperial symbols, which from time immemorial had been designated as Imperial insignia, are displayed. Other Ming portraits show officials clothed in red *p'ao* that have large bird or animal squares (called "mandarin squares," or *p'u-fang*) on the breast, specific bird and animal emblems to designate each of the nine ranks of civil and military officials having been adopted by the Ming in 1391.

When the Manchus overthrew the Ming in 1644 and established the Ch'ing dynasty, it was decreed that new styles of dress should replace the voluminous *p'ao* costume. The most formal of the robes introduced by the Manchus was the *ch'ao-fu*, designed to be worn only at great state sacrifices and at the most important court functions. Men's *ch'ao-fu* had a kimono-style upper body with long, close-fitting sleeves that terminated in the "horse-hoof" cuff introduced by the Manchus, and a closely fitted neckband over which was worn a detached collar distinguished by winglike tips that extended over the shoulders. Below, attached to a set-in waistband, was a full, pleated or gathered skirt. Precisely stipulated colours and pattern arrangements of five-clawed dragons and clouds, waves and mountains were specified for the *ch'ao-fu* of emperors, princes, nobles, and officials; the bright yellow of the emperor's robe and the 12 Imperial

Manchu styles

By courtesy of the National Palace Museum, Taipei, Taiwan, Republic of China



Figure 28: The legendary Emperor Yao wearing a *p'ao*-style robe, a wide-sleeved kimono-style garment, which, girdled at the waist, falls in voluminous folds at the feet. Hanging silk scroll by Ma Lin, Southern Sung dynasty (1126–1279). In the National Palace Museum, Taipei, Taiwan.

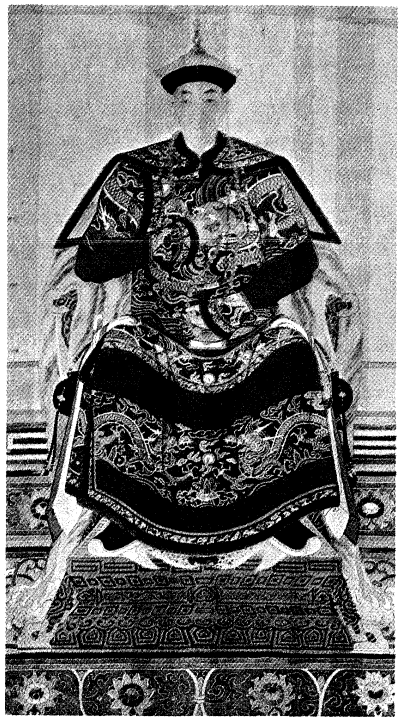


Figure 29: *Chi-fu*, or "dragon robe," worn by Ho-shen, minister to the Ch'ien-lung emperor (reigned 1736–96). Painted silk wall hanging. In the Metropolitan Museum of Art, New York.
By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1942

symbols emblazoned on it clearly established his lofty rank. All other ranks wore "stone blue" *ch'ao-fu* decorated in accordance with prescribed rules about the number, type, and arrangement of dragon motifs.

Among women, only those of very high rank were permitted to wear *ch'ao-fu*. Women's robes were less commodious than the men's and were cut in long, straight lines with no break at the waist. The narrow sleeves with horse-hoof cuffs of these *ch'ao-fu* robes and the arrangement of their dragon, cloud, mountain, and wave patterns were essentially the same as those of the so-called dragon robes discussed below. They were clearly differentiated from the dragon robes, however, by their cape-like collars and by flaring set-on epaulets which, gradually narrowed, were carried down under the arms. Stole-like vests, always worn over women's *ch'ao-fu*, were also a distinguishing feature of this costume. The colour of the empress's *ch'ao-fu* was bright yellow, related shades being worn by princesses and wives of princes.

"Dragon robes"

Chi-fu, or "dragon robes" (*lung-p'ao*) as they were usually called, were designed for regular court wear by men and women of Imperial, noble, and official rank (Figure 29). The *chi-fu* was a straight, kimono-sleeved robe with a closely fitted neckband that continued across the breast and down to the underarm closing on the right side, the long tubular sleeves terminating in horse-hoof cuffs. The skirt of the *chi-fu* cleared the ground to permit easy walking and in men's garments was slit front and back as well as at the sides to facilitate riding; the extra slits were the only feature that distinguished the *chi-fu* of men below the rank of emperor from those of their wives. All *chi-fu* were elaborately patterned with specified arrangements of dragons, clouds, mountains, and waves, to which were added auspicious and Buddhist or Taoist motifs. Distinctions in rank were indicated by the colours of the robes and by slight variations in the basic patterns; but, because of the large number of personages who wore *chi-fu*, these distinctions were not always easily recognizable. Emperors' *chi-fu*, either yellow or blue, were always distinguished by the 12 Imperial symbols; but they have also been found on a number of empresses' robes (which have identifying dragon sleeve bands), even though the Ch'ing regulations did not prescribe them for wear by the empress. A large number of *chi-fu*, including those worn by emperors and empresses, were acquired by Western museums after 1912, when the overthrow of the Manchu regime and the establishment of the Chinese republic did away with the official hierarchies of Imperial China; these provide a rich but complicated field of study.

The *p'u-fu*

The *p'u-fu*, a three-quarter-length coat worn by men and women over their *ch'ao-fu* and *chi-fu*, became an important adjunct of Ch'ing dynasty dress because it proclaimed the wearer's exact rank at a glance, for it was made of plain purplish-black silk on which his insigne was emblazoned in bright colours and gold. The specially designated *kun-fu* of the emperor and the *p'u-fu* of his family were distinguished by specific numbers and arrangement of five-clawed dragon medallions; the *p'u-fu* of nobles had squares enclosing four-clawed dragons or other mythical beasts; and civil and military officials each had their appropriate *p'u-fang*.

The informal Manchu *ch'ang-fu*, a plain long robe, was worn by all classes from the emperor down, though Chinese women also continued to wear their Ming-style costumes, which consisted of a three-quarter-length jacket and pleated skirt. Men's *ch'ang-fu*, cut in the style of the *chi-fu*, usually were made of monochrome patterned damask or gauze; women's *ch'ang-fu* had wide, loose sleeves finished off with especially designed sleevebands and gay woven or embroidered patterns. Modern traditional Chinese dress developed directly out of these *ch'ang-fu* styles. Men's robes underwent very little change. Women first narrowed the *ch'ang-fu*, then cut the sleeves off at the shoulders, and finally shortened the skirt and slit it up the sides to form the sheath style.

Japan. The earliest representations of dress styles in Japan are in 3rd- to 5th-century-AD clay grave figures, a few of which show men and women wearing meticulously detailed two-piece costumes consisting of crossed-front jackets that flare out over the hips, the men's worn over full trousers, which, banded above the knees, hang straight and loose beneath; women's jackets are worn over pleated skirts (Figure 30).

By courtesy of the National Museum, Tokyo



Figure 30: Japanese grave figure, wearing meticulously detailed two-piece costume of flared, crossed-front jacket and pleated skirt. Clay, 3rd to 5th century. In the National Museum, Tokyo.

Earliest representations of Japanese dress styles

Two-piece costumes appear to have been worn regularly during the 7th and 8th centuries, the jackets of this period being called *kinu*, the men's trousers, *hakama*, and the women's skirts, *mo*. It is known, however, that during



Figure 31: Two Japanese men and a woman (foreground) wearing the *kamishimo*, a jumperlike garment with extended shoulders; woman (far right) and men (background) in the *yukata*, a cotton kimono; wealthy merchant (seated right) wearing a short, black *haori* coat and *hakama*, men's skirt-trousers. Detail of "Shuin-Boeki-sen," a folding screen, late 17th century. In Ōsaka Castle, Japan.
By courtesy of Kumata Shrine, Osaka, Japan

the Nara period (710–784) Japanese court circles adopted Chinese court dress, the most characteristic feature of which was the long kimono-style *p'ao* garment; thus, it must be supposed that the *kinu*, *hakama*, and *mo* were the accoutrements of middle and lower class society, though these garments may also have been adapted for wear under the *p'ao*. It is clear that emblematic colours and patterns as well as the *p'ao* style were borrowed from China because modern court dress in Japan, which has been little changed since the 12th century, has many purely Chinese characteristics.

The most important court costumes

The most important court costumes of Japan are the *sokutai* of the emperor and the *jūni-hitoe* of the empress, which are worn only at coronations and at very important ceremonial functions. (Similar costumes are worn by the crown prince, by princes and princesses of the blood, by high officials, and by ladies-in-waiting.) The voluminous outer robe (*ho*) of the emperor's *sokutai* is cut in the style of the Chinese *p'ao* but is given a distinctively Japanese look by being tucked up at the waist so that the skirt ends midway between the knees and the floor. This *ho* robe is yellow (the colour worn only by emperors and their families in China), and it is patterned with *hōō* birds and *kilin* (japanized versions of the mythical Chinese *feng-huang* and *ch'i-lin*). The outer and most important of three kimonos worn under the *ho* is the *shitagasane*, which has an elongated back panel that forms a 12-foot (four-metre) train. The *shitagasane* is made of white damask, as the the baggy white trousers (*ueno-hakama*) that are a characteristic feature of the *sokutai* costume. Both of these garments and a cap-shaped headdress (*kammuri*) of black lacquered silk, with an upright pennon, decorated with the Imperial chrysanthemum crest, are purely Japanese in style; but the ivory tablet (*shaku*) carried by the emperor when wearing the *sokutai* was undoubtedly inspired by tablets of jade that Chinese emperors carried as symbols of their Imperial power.

The outermost garment of the empress' *jūni-hitoe* costume is a wide-sleeved jacket (*karaginu*) that reaches only to the waist and has a pattern of *hōō* bird medallions brocaded in colours of the empress' choice. Attached to the waist at the back of the *karaginu* is a long, pleated train (*mo*) of sheer, white silk decorated with a painted design. The outer kimono (*uwagi*) is very large to accommodate the many layers of kimono worn under it, the abnormally long skirt swirling out fanwise around the wearer's feet. This, too, is made of rich brocade, its design and colours being a matter of personal taste. Under the *uwagi* is a plain purple kimono and under that a robe known as the *itsutsu-ginu*, which has multiple bands of coloured silks (usually five) attached at the edges of the sleeves, at the neckline, and at the hem, giving the appearance of several robes worn one over another. No special interest attaches to the *hitoe* kimono worn under the *itsutsu-ginu* or to the *kosode* worn next to the body, but the divided skirt (*naga-bakama*) that completes the costume is an extremely picturesque garment. Made of stiff, red cloth and fastened high up under the breasts, the *naga-bakama* covers the feet in front and is carried out in a train in back. Worn with the *jūni-hitoe* is an elaborate coiffure known as *suberakashi;* and, affixed directly over the forehead, are special hair ornaments consisting of a lacquered, gold-sprinkled comb surmounted by a gold lacquered chrysanthemum crest.

Other types of 12th-century formalized dress

Other types of dress formalized in the 12th century were the *noshi* (courtiers' everyday costumes) and the *kariginu*, worn for hunting. Both of these garments were voluminous hip-length jackets worn with baggy trousers tied at the ankles. At this time also it became necessary to devise special costumes for the newly formed samurai caste. The *hitatare*, the formal court robe of samurai, and the *suo*, a crested linen robe designed for everyday wear, though more closely related in style to the Chinese *p'ao* than the *noshi* and *kariginu*, were characterized by V-shaped necklines accentuated by inner-robe neckbands of white. Several centuries later the samurai adopted the *kamishimo*, a striking jumperlike garment, with extended shoulders and pleated skirt-trousers, which was worn over the *hitatare*. This costume probably inspired a later fashion of wearing skirt-trousers (*hakama*) over a full-length black kimono which, together with the short black *haori* coat, is today the approved formal attire for Japanese men (Figure 31).

The basic kimono style adopted by Japanese women during the Nara period has remained amazingly close to that of the *p'ao* robes worn by the women of T'ang China. The kimono is usually thought of as a Japanese invention, and it was in fact the master designers and dyers of Japan who, in the 17th and 18th centuries, evolved styles of decoration that have made it the most beautiful garment in the world (Figure 32). The practice of wearing short-sleeved kimono (*kosode*) as outer garments and belting them in with narrow sashes (*obi*) originated during the Muromachi period (Ashikaga shogunate; 1338–1573), when samurai women began to wear a voluminous outer kimono (*uchikake*) as a kind of mantle. Eventually, the *kosode* came to be worn only by married women, the long-sleeved *furisode* being reserved



Figure 32: Japanese woman wearing a kimono. "The Courtesan Itsutomi Holding a Plectrum," print by Chobunsai Yeishi, *c.* 1794.
By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.

for young unmarried girls. The wide *obi*, which is today the most elegant feature of the Japanese costume, was not adopted until the early 18th century; and it was at this time also that women first began to wear the short *haori* coat, which has come to be an important feature of Japanese women's dress.

The *yukata*

The *yukata*, which is worn by both men and women, is a cotton kimono with stencil-dyed patterns (usually done in shades of indigo) that were originally designed for wear in the home after a bath. Because it has become accepted practice to wear *yukata* on the street on warm summer evenings, the cottons designed for them have become increasingly handsome.

Korea. Some of the basic elements of modern traditional dress in Korea, the *chŏgori* (jacket), *paji* (trousers), and *turumagi* (overcoat), were probably worn at a very early date, but the characteristic two-piece costume of today did not begin to evolve until the period of the Three Kingdoms (*c.* 57 BC–AD 668). During the early part of this period both men and women wore tight, waist-length jackets and short, tight trousers; and it is believed that the Koreans' traditional fondness for white clothing dates from this period (Figure 33).

Everyday costumes of the nobility

Korean records state that special costumes for court wear modelled after those of T'ang China were adopted during the reign of Kim Ch'unch'u in the 7th century; but Chinese influence on Korean dress at this period is verifiable only in changes that occurred in the everyday costumes of the nobility. Noblewomen formerly had worn tight trousers and jackets (which continued to be worn by the poorer classes); now they began to appear in wide-sleeved, hip-length jackets, belted at the waist, and in full-length skirt-trousers. The corresponding dress for noblemen was a narrower, tunic-style jacket, cuffed at the wrists, belted, and worn with roomy trousers bound in at the ankles. The most striking evidence of Chinese influence at this time is to be seen in the style of the *turumagi* overcoat worn by noblemen, pictured in fresco paintings as a voluminous full-length garment made almost exactly like the *p'ao* robe of T'ang China. One-piece robes were never worn in Korea until the late 13th century, when the court was forced to adopt Mongol dress; after Mongol domination ended in 1364, Koreans wore the one-piece robe only at wedding ceremonies.

In the 15th century, Korean women began to wear pleated skirts (*ch'ima*) and longer *chŏgori* (jackets), a style undoubtedly introduced from China. Noblewomen wore full-length *ch'ima* to indicate their social standing and began gradually to shorten the *chŏgori* until eventually it attained its present length, just covering the breast. This style made it necessary to reduce the fullness of the skirt somewhat in order to make it possible to extend it almost up to the armpits, which remains the fashion (Figure 33).

The adoption of Chinese-style *p'u-fang* (mandarin squares) as emblems of rank for civil and military officials (who wore them on their *turumagi*, or overcoats) appears to have been the only notable example of Chinese influence on men's dress at this period. Otherwise, few changes were made until 1894, when class distinctions were relaxed by government decree. It was at this time that the *turumagi* was shortened and narrowed to its present form.

The most picturesque costume of modern Korea

The most picturesque costume of modern Korea is that of men of leisure, *yangban*, who are past 60 years of age. The *yangban* wear white almost exclusively, their costumes consisting of full trousers tied at the ankles with ribbons, over which is worn a short *chŏgori* and a fitted vest and, over all, a loose *turumagi*, which falls just below the knees and is tied at the breast. The patriarchal appearance of the *yangban* (who is usually bearded) is accentuated by a black horsehair hat, its flat brim and high crown giving him somewhat the appearance of an American colonial Pilgrim Father (Figure 33). Younger men wear a similar costume (though not the hat) in gray, light blue, or light brown.

Women's costumes feature a bolero-style white *chŏgori*, finished off at the neck by a figured band or ribbon that ties from left to right, and high-waisted *ch'ima*, which, in formal costumes, is a full, billowing garment made of beautifully patterned silk.

By courtesy of Chun Sung-woo, Korea



Figure 33: Korean women wearing short *chogori*, a jacket, with raised up skirts. Man in foreground wears loose *turumagi*, an overcoat, which ties at the chest. Inked and coloured leaf from an album of 30 leaves by Sin Yun-bok (1758–?). In the collection of Chun Sung-woo, Korea.

SOUTH ASIA

The Hindu population of South Asia comprises about 2,000 castes whose members wear clothes and ornaments that clearly indicate their caste. The subject of dress, therefore, cannot be dealt with satisfactorily in a few paragraphs. Some of the principal features of upper class Hindu and Muslim dress and the history of their development can, however, be sketched briefly.

The dhoti and sari

The ancient origin of two of the most characteristic garments of modern India, the dhoti worn by men and the sari worn by women, is verifiable in sculptured re-



Figure 34: *South Asian dress.*
(Left) Indian woman wearing a *ghāghrā*, open-front, pleated skirt, with a long apronlike panel over the front opening, and a *colī*, short-sleeved, breast-length jacket. Detail from a miniature of the Rajasthani style, late 18th century. In a private collection. (Right) Indian wearing the *jāmah*, a long-sleeved coat that reaches to the knees or below and is belted in with a sash. Detail from "The Emperor Shah Jahan," oil painting by Bichitra, 1631. In the Victoria and Albert Museum, London.
(Right) By courtesy of the Victoria and Albert Museum, London; photographs, (left) P. Chandra, (right) EB Inc.

liefs dating back as far as the 2nd century BC. In these and in slightly later reliefs, both men and women are pictured wearing a long piece of cloth wrapped around the hips and drawn between the legs in such a fashion that it forms a series of folds down the front. The upper bodies of both men and women were unclothed, though women wore a narrow cloth girdle around the waist. Men are pictured wearing large turbans, women with head scarves that fall to the hips. Women also wore a great amount of jewelry—bracelets, anklets, and girdles—men's ornaments consisting solely of bracelets.

No major change in costume appears to have been made until the 12th century, when the Muslims conquered north and central India. In this part of the subcontinent, radical new dress styles were adopted to conform with Muslim practice, which required that the body be covered as completely as possible. Men's costumes thereafter consisted of the *jāmah*, a long-sleeved coat that reached to the knees or below and was belted in with a sash, and wide trousers known as *isar* (Figure 34, right). These garments and the *farjī*, a long, gownlike coat with short sleeves, which was worn by priests, scholars, and high officials, were made of cotton or wool, silk being forbidden to men by the Qur'ān. Somewhat modified, these traditional styles continue to be worn by upper class men of Pakistan and Bangladesh.

Muslim women's garments

Women's garments, dictated by the Muslim conquerors, consisted of wide-topped trousers snugly fitted around the calves of the legs, a long shirtlike garment, and a short, fitted outer jacket. Silk was not forbidden to women; and highborn women, forced to spend their lives in seclusion, devoted a great deal of time and money to their costumes. The Mughal emperor Akbar's Rājput wives, inspired by the profusion of luxurious fabrics available in India, designed a graceful new style of dress, which Muslim women adopted forthwith. This costume consisted of an open-front pleated skirt, or *ghāghrā*, worn with a long apronlike panel over the front opening, and a short-sleeved, breast-length jacket called a *colī* (Figure 34, left). The *ghāghrā* and *colī* continue to be basic elements of Muslim women's dress, the loose front panel having been replaced by the traditional sari, which is worn as an overgarment, one end draped around the hips, the other carried up over the shoulder or head.

Dress in southern India was little affected by Muslim rule in the north. The dhoti continued to be worn by most Hindu men (it is forbidden to some castes) and the sari by women. Some additions to these traditional costumes have been adopted. On formal and semiformal occasions many Hindu men wear a long, full-skirted, white cotton coat, which reaches to the knees and buttons down the front from top to bottom, over jodhpur-style white trousers; and most Hindu women wear a short *colī*-style jacket under a sari or a loose waist-length bodice. (Pa.S.)

BIBLIOGRAPHY

Western: R. TURNER WILCOX, *Mode in Costume*, rev. ed. (1948, reprinted 1969), a good, readable account, profusely illustrated by line drawings with comments; HILAIRE HILER, *From Nudity to Raiment* (1929); HILAIRE and MEYER HILER, *Bibliography of Costume: A Dictionary Catalog of About Eight Thousand Books and Periodicals* (1939); KATHERINE MORRIS LESTER and ROSE N. KERR, *Historic Costume: A Resumé of Style and Fashion from Remote Times to the Nineteen-Sixties*, 6th ed. (1967); MARY BROOKS PICKEN, *The Fashion Dictionary* (1957); FRANCOIS BOUCHER, *Histoire du costume en occident* (1965; Eng. trans., *20,000 Years of Fashion*, 1967); IRENE PENNINGTON HUENEFELD, *International Directory of Historical Clothing* (1967); JOYCE ASSER, *Historic Hairdressing* (1966); MARGOT HAMILTON HILL and PETER A. BUCKNELL, *The Evolution of Fashion* (1967); BLANCHE PAYNE, *History of Costume, from the Ancient Egyptians to the Twentieth Century* (1965); DOUGLAS GORSLINE, *What People Wore* (1952); MARGOT LISTER, *Costume* (1967); RICHARD CORSON, *Fashions in Hair: The First Five Thousand Years* (1965); PHILLIS CUNNINGTON, *Costume* (1966); ALISON GERNSHEIM, *Fashion and Reality, 1840–1914* (1963); J. STEVENS COX, *An Illustrated Dictionary of Hairdressing and Wigmaking* (1966); REGINALD REYNOLDS, *Beards* (1950); MILLIA DAVENPORT, *The Book of Costume*, 2 vol. (1948, reprinted 1964). (*Antiquity*): JAMES LAVER, *Costume in Antiquity* (1964); LEON and JACQUES HEUZEY, *Histoire du costume dans l'antiquité classique: L'Orient* (1935), most helpful in interpreting the evidence from the monuments; HENRY F. LUTZ, *Textiles and Costumes Among the Peoples of the Ancient Near East* (1923), mainly discusses the evidence from the texts. (*Middle Ages to Modern*): DOREEN YARWOOD, *English Costume from the Second Century B.C. to 1952* (1953); PERCY MACQUOID, *Four Hundred Years of Children's Costume from the Great Masters, 1400–1800* (1923); WOMEN'S WEAR DAILY, *Sixty Years of Fashion: 1900–1960* (1963); MARGARETE BRAUN-RONSDORF, *Modische Eleganz* (1963; Eng. trans., *Mirror of Fashion*, 1964); ALFRED RUBENS, *A History of Jewish Costume* (1967); C. WILLETT CUNNINGTON, PHILLIS CUNNINGTON, and CHARLES BEARD, *A Dictionary of English Costume, 900–1900* (1960); ANNE BUCK, *Victorian*

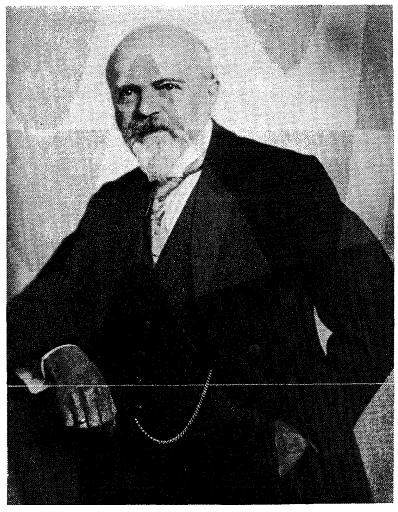
Costume and Costume Accessories (1961); A. HYATT MAYOR, "Change and Permanence in Men's Clothes," *Metropolitan Museum Bulletin*, New Series, 8:262–269 (1950).

Non-Western: ALAN PRIEST, *Costumes from the Forbidden City* (1945) and *Japanese Costume* (1935); HELEN FERNALD, *Chinese Court Costume* (1946); SCHUYLER CAMMANN, *China's Dragon Robes* (1952); KEN-ICHI KAWAKATSU, *Kimono*, 4th rev. ed. (1956); ELIZABETH KEITH, *Old Korea, the Land of Morning Calm* (1946); CORNELIUS OSGOOD, *The Koreans and Their Culture* (1951); G.S. GHURYE, *Indian Costume* (1951); JAMILA BRIJ BHUSHAN, *The Costumes and Textiles of India* (1958); A. LEIX, "Indian Costume," *Ciba Review*, no. 36 (November 1940); HELEN B. MINNICH and SHAJIRO NOMURA, *Japanese Costume and the Makers of Its Elegant Tradition* (1963); SUMIKO HASHIMOTO, *Japanese Accessories* (1962); MAX TILKE, *Orientalische Kostüme in Schnitt und Farbe* (1923; Eng. trans., *Oriental Costumes: Their Designs and Colors*, 1923); HOWARD S. LEVY, *Chinese Footbinding: The History of a Curious Erotic Custom* (1966).

Driesch, Hans Adolf Eduard

Through his many and varied pioneering experiments in embryology, Hans Driesch exerted an uncommon influence on biology. Possessed of a philosophical cast of mind, he worked in experimental embryology as a gifted amateur for over 20 years before becoming a professor. As soon as he assumed this position he became a systematic philosopher rather than a scientist, devoting himself fully to systematic philosophy. His inability to interpret his experimental results in strictly scientific terms led him into vitalism—the concept that the processes of life cannot be explained only by the laws of physics and chemistry—for which he was the last great spokesman.

By courtesy of the Ruprecht-Karl-Universitat, Heidelberg, West Germany



Driesch.

Driesch was born at Bad Kreuznach, a German spa, on October 28, 1867. For his early education, his father, a well-to-do gold merchant in Hamburg, sent him to a prominent humanistic gymnasium that had been founded by a friend of Martin Luther. Driesch's interest in zoology was aroused while he was still a child by the unusual live animals his mother kept in their home.

Following the prevailing custom in Germany, Driesch attended several universities (at Hamburg, Freiburg, and Jena); he studied zoology and also chemistry and physics, which were not usual topics of study at that time for zoologists. He did his doctoral work at Jena under Ernst Heinrich Haeckel, whose main interest was in phylogeny, a special branch of evolutionary theory. But Driesch's own work took a different direction. His doctoral dissertation in 1887 dealt with factors controlling the growth of colonial hydroids.

For the next ten years Driesch travelled extensively; he also experimented during this period with marine eggs, frequently at the international Zoological Station in Naples. In 1891 he separated the first two cells formed by a dividing sea urchin egg and discovered that each would form a whole larva. A similar experiment had been performed on the frog's egg by Wilhelm Roux in 1888, but with quite different results; each of the first two cells formed only half an embryo, and Roux concluded that the parts of an organism are determined at the two-cell stage. Driesch, however, concluded that the fate of a cell is not determined at the two-cell stage, but by its position in the whole organism. He published his first wholly theoretical monograph that year and, in 1892, speculated that vitalistic interpretations of biological data might be reasonable. His experimental results gave strong impetus to the then new science of experimental embryology.

Work in experimental embryology

Driesch made many other less well known but equally important contributions to embryology. He produced a giant larva by fusing two embryos. By compressing dividing eggs he caused an abnormal distribution of nuclei, thereby proving that the nuclei are all equivalent; this experiment was an important forerunner of modern genetics. He recognized that nuclei and cytoplasm interact and postulated that the nucleus exerts its influence on the cytoplasm by means of ferments, or enzymes. In 1896 he shook sea urchin larvae to displace their skeleton-forming cells and observed the displaced cells return to their original positions. This experiment was the first demonstration of embryonic induction—that is, the interaction between two embryonic parts resulting in differentiation that would not have occurred otherwise—the theoretical aspects of which he had speculated upon in a monograph published in 1894.

Belief in vitalism

By 1895 Driesch was a convinced vitalist. He felt himself driven to this position by his inability to interpret his cell-separation experiments in mechanistic terms; he could not envisage a machine that could divide into two identical machines. Driesch applied the Aristotelian term entelechy to denote a vital agent that could regulate organic development. Although such an agent could not be explained by physical science, he believed that its actions were related to the activity of enzymes, which he recognized as important in development.

In 1899 Driesch married Margarete Reifferscheidt, by whom he had two children, Kurt and Ingeborg, both of whom became musicians. Settling in Heidelberg, he continued to perform embryological experiments until 1909, when he was at last habilitated—the procedure then required to enter the German university hierarchy—in natural philosophy. As a member of the faculty of natural sciences, he held successive professorships of philosophy at Heidelberg beginning in 1912 and transferred to Cologne in 1919 and to Leipzig in 1921. As a philosopher, he was strongly influenced by Immanuel Kant, and metaphysics was one of his specialties; logic was another. Perhaps because of his leanings toward vitalism, he also became interested in parapsychology.

Driesch's work was of immediate importance in stimulating the progress of experimental embryology. His experimental and theoretical studies—widely disseminated in his own time though now forgotten—on embryonic induction, enzyme action, and nuclear and cytoplasmic interaction led to work that continues today, but in a less vitalistic framework. In 1935 Driesch was forced into early retirement by the Nazis, but he continued to write until his death in Leipzig on April 16, 1941.

BIBLIOGRAPHY. MARGARETE DRIESCH, "Das Leben von Hans Driesch," in ALOYS WENZL (ed.), *Hans Driesch, Persönlichkeit und Bedeutung für Biologie und Philosophie von heute* . . . (1951), includes a bibliography of articles and monographs; CURT HERBST, "Hans Driesch als experimenteller und theoretischer Biologe," *Arch. EntwMech. Org.*, 141:111–153 (1941–42), analyzes his scientific work. JANE OPPENHEIMER, "Hans Driesch," *Dictionary of Scientific Biography*, vol. 4, pp. 186–189 (1971), is the only biography available in English. The best known of Driesch's many books on embryology and on philosophy is his *Science and Philosophy of the Organism*, 2 vol. (1908; 2nd ed., 1929). Driesch's posthumous autobiography is *Lebenserinnerungen. Aufzeichnungen eines Forschers und Denkers in entscheidender Zeit* (1951).

(J.M.O.)

Drug and Drug Action

Drugs, chemical substances that affect the functions of living things, are used in treating, preventing, and diagnosing diseases. Numerous other definitions of the word drug exist; in the popular sense, the term drug refers mainly to substances that affect psychological or behavioral functions or lead to varying degrees of dependence or addiction. Pharmacology, the science of drugs, deals with all aspects of drug action and can be considered a part of biology and an important basic medical science. The experimental study of drugs and drug action dates from the mid-19th century, and a German pharmacologist, Oswald Schmiedeberg, has been called the father of pharmacology.

Drug sources

The most important source of drugs today is chemical synthesis. The increase in the manufacture of drugs has resulted in the development of 25,000 or more drugs and drug products. Numerous drugs of natural origin also are available; these are obtained from plants, animals, minerals, bacteria, and fungi. In most instances, however, either an active principle is purified from a natural substance or some chemical modification is introduced. Drugs of natural origin include the antibiotics, which are derived from the growth media of bacteria or fungi (*e.g.*, the penicillins, streptomycin), and the immunizing agents (*e.g.*, vaccines, antitoxins, antiserums), whose exact chemical nature has not yet been established. Drugs produced by numerous pharmaceutical businesses throughout the world are prescribed by physicians or purchased without prescriptions (the so-called patent medicines or over-the-counter drugs).

Many drugs are potentially dangerous chemicals that can cause serious, sometimes fatal, poisoning if used incorrectly; governments of various countries, therefore, have established certain legal requirements concerning drug use. In the United States, for example, control of drugs is based on laws enacted under the Federal Food, Drug, and Cosmetic Act, which are enforced by the U.S. Food and Drug Administration (FDA). The purposes of governmental regulations include the protection of the drug user against the potential hazards of drugs and against fraud. The *United States Pharmacopeia* (USP), the *National Formulary* (NF), and the *Pharmacopoea Internationalis* are examples of official publications that provide names, chemical and physical properties, and other information concerning drugs.

Associated with the benefits that have resulted from modern drug development has been an increase in medical and nonmedical uses and abuses of drugs. Most hospital patients, for example, receive four or five drugs during a hospital stay, and large quantities of drugs also are used by private physicians and by individuals for self-medication. Widespread abuse of drugs has been reflected by an increase in the occurrence of drug toxicity, also referred to as iatrogenic disease. Increased consumption of psychoactive drugs and narcotics (see below *Drugs that produce stupor and narcosis*) has resulted in a drug cult or drug culture.

GENERAL ASPECTS OF DRUG ACTION

Chemical structure and drug receptor interactions. The primary basis of drug action is an interaction between a drug molecule and some specific cellular component, which is called a receptor or receptive substance. Functional changes produced by a drug on various organs (*e.g.*, heart, lung, kidney) are termed effects. In contrast, the basic interactions between a drug molecule and a receptor are referred to as the action or mechanism of action of a drug. The mechanism by which a drug influences a cellular process to cause a functional change in an organ has not yet been established for all drugs. Most drugs are selective in their activity; *i.e.*, a main action may be associated with some primary effect. Such selectivity provides a convenient basis for the classification of drugs; *e.g.*, certain drugs affect the central nervous system (CNS), or the heart, or the blood vessels. This selectivity, however, is relative, since no drug exerts only a single action. Although the exact nature of the receptor is not yet known for many drugs, evidence exists that receptors may be associated with specific enzymes (*i.e.*, organic catalysts) in cell cytoplasm or membranes, or with other molecules found in the cell.

Affinity and efficacy

Drug-receptor interactions involve two distinct processes: affinity is a term applied to the specific localization of a drug at its binding sites in a receptor; intrinsic activity, or efficacy, is the specific functional change that results from the binding of a drug. The degree of receptor binding, however, is not always a criterion of the degree of efficacy because drugs also may be bound nonspecifically in body tissues (*e.g.*, plasma proteins) without inducing any pharmacological effects. The magnitude of the action of a drug also is proportional to the concentration of a drug at a receptor; *i.e.*, intensity of drug action increases with increasing drug dosage. This graded response, which has been attributed to the occupation of increasing numbers of receptors by increasing numbers of drug molecules, is referred to as receptor occupancy.

Changes in the chemical structure of a drug can modify its actions. The relationship between chemical structure and drug action, referred to as a structure–activity relationship (abbreviated SAR), has been extensively studied. Among numerous groups of drugs, chemical modification of a parent drug molecule has been widely used in studying drug receptor interactions and in attempting to develop new drugs. General similarities in actions and effects of several chemically derived classes of drugs, or cogeners (*e.g.*, barbiturates, sulfonamides), are known. On the other hand, it is not always possible to predict the results of chemical modifications; *e.g.*, one type of chemical change, acetylation, of the parent molecule of morphine leads to increased efficacy, but substitution of a methyl group in the morphine molecule leads to formation of nalorphine, a potent morphine antagonist.

Evidence also exists that drug–receptor interaction depends on the three-dimensional structure of a drug molecule. One isomer (*i.e.*, structural arrangement) of norepinephrine, for example, induces significantly more intense physiological activity than another; this is the case also with the effects of the isomers of amphetamine on the central nervous system. These structure–activity relationships may be the result of fundamental differences in intramolecular forces and the three-dimensional structure of a drug molecule.

Mechanisms of drug action. In attempts to establish mechanisms of drug–receptor interactions, the nature of the effects of drugs that have opposite actions (*i.e.*, drug antagonists) have been extensively studied. Drugs that combine with receptors and initiate action have both affinity and efficacy (or intrinsic activity); these drugs are called agonists. Drugs that react with receptors but show no efficacy are called antagonists. An agonist usually cannot combine with receptors that have been blocked by an antagonist; this is termed receptor blockade. In some types of drug antagonism, however, an increase in the amount of an agonist can overcome a blockade induced by an antagonist. Known as competitive antagonism, this phenomenon may be the result of competition between the two drugs for specific receptor sites. On the other hand, in noncompetitive antagonism the receptor blockade induced by an antagonist is not affected regardless of the amount of agonist added. It has been suggested that, in this case, agonist and antagonist may act at different binding sites on the same receptor.

It has been observed that the relationships between agonist and antagonist obey the law of chemical kinetics or law of mass action, in that the velocity of the reactions between agonist and antagonist at a receptor site is directly proportional to the quantities of the reacting substances. Most drug–receptor interactions also have been shown to behave in a manner similar to the interactions between enzymes and specific tissue chemicals usually referred to as substrates, in which reversible enzyme-substrate complexes are formed. Drug–receptor interactions, therefore, have been described as a lock and key relationship between drug molecules and receptors. The

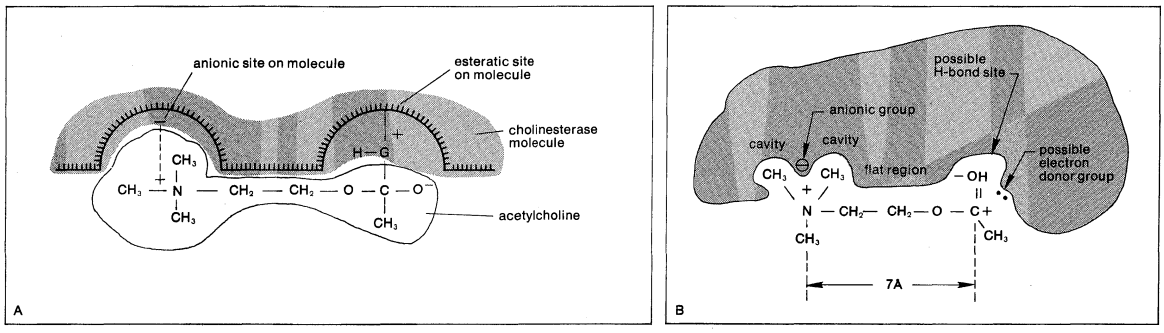


Figure 1: (A) Representation of the active centre of a cholinesterase enzyme molecule and its interaction with acetylcholine. *G* represents an electron-rich centre. (B) Hypothetical complex between acetylcholine and its receptor.

Adapted from A. Goldstein, L. Aronow, and S.M. Kalman, *Principles of Drug Action: The Basis of Pharmacology* (1968); Hoeber Medical Division, Harper & Row

specificity of drug action among chemically related drugs apparently depends on the degree of fit between a drug and its receptor molecules; an optimal fit would correspond to maximal specificity of drug action.

Figure 1 illustrates the concept of drug and receptor interactions. Figure 1A shows the molecular fit involved in interactions between a drug, acetylcholine, which acts as a substrate, and the enzyme cholinesterase (with specific anionic and esteratic binding sites within the structure of the molecule), which acts as a protein receptor for the acetylcholine molecule. Figure 1B illustrates the complex acetylcholine might possibly form with a hypothetical receptor. The fit between acetylcholine and its receptor at the two binding sites suggests ionic bonding (*i.e.*, attraction between oppositely charged particles, positive attracted to negative). On the other hand, in other types of drug–receptor interactions, more stable binding may occur between drug and receptor and involve covalent bonds, in which two atoms share a pair of negatively charged particles, or hydrogen bonds. The administration of an antagonist drug also may cause changes in the shape of a protein receptor; such conformational changes in protein structure could partially or completely prevent the actions of a specific agonist by changing the positions of the binding sites so that chemical interaction would no longer be possible.

Role of physical and chemical properties

Although the fundamental mechanisms of drug–receptor interactions are not yet known with certainty, it is known that the mechanism of action of some types of drugs is related to special physical or chemical properties. Magnesium sulfate, for example, is poorly absorbed in the large intestine, and the retention of water that results there causes the cathartic action of the salt; the antidotal action of dimercaprol (British Anti-Lewisite; BAL) in mercury poisoning is due to the affinity of mercury ions for certain chemical groups on the BAL molecule that complex with mercury, thus making it inactive. Alterations in control mechanisms involved in transmission of nerve impulses and in the permeability of the cell membrane to various substances also are involved in some types of drug actions. Nonspecific interferences in cell functions caused by the properties of some drugs (*e.g.*, the general anesthetics) also can reversibly interfere with the normal functions of brain cells.

Quantitative evaluation of drug responses and efficacy. The intensity of the effects of a drug depends on the dose, or quantity of the drug used. Although drug responses may be increased or decreased by numerous factors (see below *Factors modifying drug responses*), dose–effect relationships can be determined for any drug by plotting the dose of the drug against the intensity or magnitude of the effects observed under controlled experimental conditions. Although the relationship between dose and effect varies considerably among different types of drugs, generally, when increasing doses of a drug are plotted against the intensity of the effects produced, an S-shaped curve is obtained. Drugs with similar actions have similar types of curves.

Drug potency

If the amount of one drug required to produce a given response is smaller than the quantity of another drug required to produce an equal effect, the former is said to be more potent than the latter; conversely, if the quantity of one drug required to produce an effect is larger than the quantity of another drug required to produce the same effect, the former is less potent than the latter. The term potency, therefore, refers only to comparative drug quantities; potency alone is not a criterion of the efficacy of a drug. The relative degree of drug effectiveness can be determined experimentally from dose–effect curves. The slope of the curve, which indicates the rate of increase in response with increasing doses of a drug, also provides some indication of its margin of safety (see below *Drug interactions and drug toxicity*). In the quantitative assessment of drug action, the use of dose–effect curves is indispensable.

As a result of hereditary variability, individuals differ in their response to the same dose of a drug. For exact comparisons of the relationships between drug efficacy and drug toxicity, therefore, a dose–percent curve, in which the dose of the drug is plotted against the percent response, is used. The dose required to produce a specific response in 50 percent of individuals tested is called the median effective dose and is abbreviated ED50. In comparative toxicity tests with animals, death is the end point, and the dose required to cause death in 50 percent of the animals is termed the median lethal dose (abbreviated LD50). For an experimental evaluation of the margin of safety of a drug, therefore, the ratio between the median toxic dose and the median effective dose (LD50/ED50), referred to as the therapeutic index of a drug, is widely used to predict the safety range of a drug for therapeutic use.

In the case of drugs that contain unknown quantities of active principles or mixtures of active principles, chemical analysis is not used to determine activity; instead, quantitative evaluation of activity is performed by biological assay or bioassay methods. These methods involve comparing dose–effect curves of a preparation of the unknown with dose–effect curves of an official reference standard of the same drug; *e.g.*, pituitary extract or insulin, supplied by either the World Health Organization or the government of the country involved. In order to achieve uniformity in such drug preparations, evaluation of the preparations is carried out in laboratory tests involving animals.

Factors modifying drug responses. Both the quality and the quantity of the effects of a drug can be influenced by numerous factors. These factors include those concerned with the drug itself; *i.e.*, dose, route of administration, frequency of administration, absorption, distribution in the body, metabolism, and rate of excretion. Other factors are concerned with features of the experimental animal, or patient, to whom the drug is given; *i.e.*, age, weight, sex, hereditary factors, physiological variability, and associated disease states.

Influences and bias on the part of patient and physician can significantly modify drug responses, thus interfering with the interpretation of the therapeutic efficacy of a drug. In order to avoid such complications, test responses to new drugs require the use of a dummy preparation or

substitute drug, referred to as a placebo, which consists of an inactive constituent (*e.g.*, lactose, glucose) in place of the active drug. In addition, so-called blind or double blind procedures are employed. In blind tests a patient does not know if he is given a placebo or the real drug; in double blind tests neither the patient nor the physician knows if the real drug or a placebo is being tested.

Variation in drug responses due to excessive natural sensitivity to a drug is termed idiosyncrasy; decreased natural sensitivity is termed resistance. Repeated use of certain drugs (*e.g.*, morphine) can lead to acquired resistance, or tolerance, in which higher and higher doses are required to produce the same quantitative response. An acute form of tolerance occurs with some drugs tested by repetition at short intervals; this tolerance is referred to as tachyphylaxis. Repeated use of a drug can also lead to drug dependence (*i.e.*, if the drug is not available, an individual shows abnormal emotional symptoms) or to drug addiction, which is a more intense form of dependence associated with both physical and psychological distress.

Drug combinations

Another important factor that can modify drug responses is drug combinations. The use of fixed dose combinations of drugs may be dangerous because individuals vary in their susceptibility to different drugs. In some cases, however, drugs may be legitimately combined in order either to improve their therapeutic efficiency or to lessen any toxic or adverse effects associated with high doses of either drug used alone.

Drug interactions and drug toxicity. When two or more drugs are combined, the resultant effects of either drug may be either increased, decreased, neutralized, or antagonized. When the combined effects of two drugs equal the expected sum of the individual effects, the response is termed summation. The combined effects may, however, be additive; *i.e.*, not equal to the expected sum of the two effects. When combined responses to two drugs exceed the sum of the expected effects of the individual agents, the enhanced response is termed synergism or potentiation.

No drug is entirely nontoxic; therefore, caution is a prerequisite to drug use. Drug toxicity may involve the development of effects that range from trivial or mild to serious or fatal. Apart from their primary effect, unavoidable side effects may also develop from secondary or associated effects that result from use of a particular drug. Toxic reactions to a drug, often the result of drug allergy, or hypersensitivity, may be associated with skin rashes, high fever (called drug fever), and damage to white or red blood cells. Allergy is often an unpredictable feature of the use of new drugs; there is, as yet, no satisfactory experimental procedure for evaluating this type of potential drug toxicity. Drug allergy primarily involves immunological (*i.e.*, antigen–antibody) reactions. Drug toxicity may increase in the presence of liver or kidney disease, due to impairment in the detoxification or excretion of drugs. The use of drugs during pregnancy requires special precautions in view of possible damaging effects to the fetus; for example, development of fetal abnormalities (teratogenic effects) were observed with the drug thalidomide. Drug toxicity is also a problem in infants and children because drug detoxifying enzymes are not completely developed in the young.

THE FATE OF DRUGS IN THE BODY

General characteristics. In order to exert its action, a drug must reach its receptor, or target, site in an adequate quantity. The amount of a drug that reaches a receptor site depends on the dose given and the route of administration used (*e.g.*, oral, injection, etc.). Most drugs are carried to tissues or organs of the body in the plasma of the blood stream, ultimately crossing the walls of very small blood vessels, called capillaries, to reach their target sites. Variable amounts of different drugs, however, may be combined (*i.e.*, bound) reversibly with protein molecules in the plasma; some drugs also are stored in different body tissues in a free or bound form. Drugs may also be modified chemically in the body through the complex series of events known as metabolism. Finally, both the residue of the unaltered drug and the products, called metabolites, of its metabolism are eliminated from the body by excretion, mainly through the kidneys. The ultimate fate of a drug in the body is dependent, however, on its inherent physical and chemical properties and on various factors related to its administration, distribution, storage, metabolism, and elimination.

Figure 2 is a schematic illustration of the general fate (*i.e.*, absorption, distribution, and excretion) of drugs. Plasma, body fluids, and cells are separated by cellular components known as membranes (vertical broken lines in Figure 2), the living barriers through which a drug must pass during its absorption, distribution, and excretion. Although the exact nature of biological membranes has not yet been clearly established, there is evidence that all cell membranes are similar in structure and consist of double layers of lipid (fat) molecules with layers of protein molecules on the inner and outer surfaces of the membrane. Openings, referred to as "pores," may exist at intervals in cell membranes; in addition, normal cell membranes are selective with respect to the passage of certain charged atoms through them. The outer layers of cell membranes contain enzymes that are important in transferring substances through the membranes; these enzymes may be affected by certain drugs.

Importance of membranes

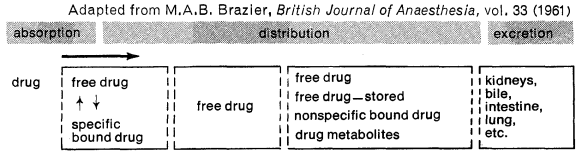


Figure 2: Absorption, distribution, and excretion of drugs.

The passage of a drug through cell membranes is achieved by two general mechanisms, namely, passive transfer and active transport. Passive transfer involves simple, or passive, diffusion of a drug through a membrane by physical processes; that is, the rate of passage is directly proportional to the amount of a drug on each side of the membrane. Passive diffusion can occur in either direction across a membrane (inward or outward), and small molecules cross membranes at a faster rate than do large ones. The rate of diffusion also depends on the relative ease with which the drug dissolves in oil or water and the ratio of ionized, or charged, to un-ionized molecules of the drug in solution. Since most drugs are either weak organic acids or bases, their degree of ionization depends on the hydrogen ion concentration (or pH) of fluids on either side of the cell membrane. Because an un-ionized drug dissolves more readily in lipids than does an ionized one, and because membranes contain lipids, most drugs diffuse through membranes in the un-ionized form. The rate of diffusion of a drug is influenced by changing the hydrogen ion concentration (pH) at the cell membranes. On the other hand, a water-soluble ionized drug generally is poorly absorbed; *i.e.*, it does not cross the membrane barrier with ease.

The active transport of drugs is an energy-dependent process in which drugs can cross membranes against concentration gradients. Active transport is specific and depends on the three-dimensional structure of a drug. Active transport of drugs can be blocked specifically by chemicals whose structure is closely related to that of the drug, by drugs that affect receptors, or, nonspecifically, by drugs that interfere with energy production in cells. Although the complex biochemical processes involved in active transport have not yet been established with certainty, the fate of a drug in the body depends both on its chemical and physical properties and on the special characteristics of passive and active transport involved at the cell membranes through which the drug must pass.

Enteral and parenteral administration

Administration and distribution. Drugs are administered by two general methods, enteral and parenteral. Enteral (*i.e.*, involving the alimentary canal) methods of drug administration include oral (swallowing the drug), sublingual (placing the drug under the tongue), and rectal

(placing the drug in the lower section of the intestinal canal or rectum). Parenteral (*i.e.*, involving channels outside the alimentary canal) methods of drug administration include intravenous (direct injection into a vein), subcutaneous (injection under the skin), intramuscular (injection into muscles), and inhalation (through the lungs). The rate of absorption into the blood following drug administration by any route depends on the physical and chemical properties of the drug itself. After oral administration, absorption of a drug into the bloodstream occurs in the stomach and the intestines; although enteral absorption usually involves passive transfer through membranes, special active transport mechanisms may be involved with some substances (*e.g.*, glucose absorption from the intestines). In the intravenous route of drug administration, the effects of a drug may become evident within seconds; this may be especially valuable for emergency treatment. Rapid intravenous injections, however, may involve serious circulatory and respiratory complications. After subcutaneous or intramuscular injections, the tissue distribution of a drug whose absorption involves passive transfer also depends on the local blood supply. Inhalation is used with gases, and passage of the gas into the blood occurs rapidly by passive transfer through the epithelial cell layer of the lungs.

In blood plasma (see Figure 2), drugs are found either free in solution or bound to the plasma. The ultimate distribution of a drug to the tissues depends on the degree of binding of a drug to the plasma. Transfer of a drug through the walls of blood vessels is rapid and occurs by passive transfer. The supply of most drugs to tissues, therefore, is limited by the rate of blood flow rather than by diffusion through blood capillaries, with the exception of supply to the tissues of the central nervous system. In this case, the blood–brain barrier appears to permit selectively the passage of drugs into the brain.

Metabolism. After its distribution to tissues, a drug may undergo various chemical transformations in the body. These changes result in the formation of some new chemical, or metabolite, which usually is less active than the original drug. In some cases, however, the metabolite may be more active than the original form of the drug; some drugs, in fact, require metabolic transformation for their action. The metabolites of most drugs are more water soluble than the original drug and, therefore, can be excreted more easily in urine than can the drug itself.

Liver enzymes

The metabolism of drugs is catalyzed by enzymes found in a subcellular network in the liver known as the endoplasmic reticulum. The amount of drug-metabolizing enzymes in the liver varies among individuals of the same species and in different species. These differences also correlate with variations in the duration of action of the same drug in different individuals and species. Individual variability in rates of drug metabolism may also contribute to differences in individual susceptibility. Hereditary defects in some enzymes (*e.g.*, plasma cholinesterase) are responsible for increased sensitivity of some individuals to certain substances. Liver enzymes are not present in newborn children and appear very slowly during early childhood. This lack of enzymes may be responsible for the increased sensitivity and toxicity of drugs in infants and young children.

The metabolism of a specific drug may be blocked or enhanced by other drugs that affect liver enzymes, which are not drug specific; the effects of a specific drug may be altered, therefore, if drug combinations are used. Inhibition of the action of metabolizing enzymes (*e.g.*, monoamine oxidase) of one drug by another may cause serious drug reactions, even death; monoamine oxidase, for example, is involved in the metabolism of several classes of drugs. On the other hand, repeated administration of drugs that are inactivated by liver enzymes may induce an increase in enzyme synthesis; therefore, certain drugs may stimulate their own metabolism, and prolonged administration may result in a progressive decrease in the effectiveness of a drug. In addition, increased enzyme activity induced by one drug can cause enhanced destruction of other drugs normally metabolized by the same enzyme. The induction of liver enzymes by one drug, therefore, can render other drugs less effective.

Elimination. Elimination of drugs from the body is mainly by excretion through the kidneys. Some drugs also may be eliminated through the liver into the intestines; in such cases, however, reabsorption from the intestines and subsequent kidney excretion may occur. Gaseous drugs are eliminated mainly through the lungs, although small quantities of these drugs may occur in sweat, saliva, and milk. Drugs are excreted through the kidneys in their unchanged form or as metabolites and may be ionized or un-ionized; the rate of drug excretion through the kidneys is dependent on all of these factors. During renal excretion, passive transfer of the drug occurs first through the membranes of kidney structures called glomeruli. The rate of membrane transfer depends primarily on the quantity of drug in the blood; however, the amount of a drug that is bound (to plasma) may interfere with its rate of passage through the glomerular membranes. Drugs are excreted from the kidney tubules both by passive transfer and by specific active transport mechanisms. The rate of drug excretion may be modified by changes in the hydrogen ion concentration (pH) in the kidney tubules.

Importance of the kidney

Absorption, distribution, metabolism, and excretion influence the rate of disappearance of a drug from the bloodstream and are significant in predicting the duration of action of a drug. After a single dose of a drug, its concentration in the blood rapidly attains a maximum, then progressively declines with time. The overall rate of disappearance of a drug from the blood, defined as the time required to produce a 50 percent decrease in the concentration of the drug in the blood, is referred to as the biological half-life or $t^{1/2}$ of the drug. If the blood concentration of a drug declines rapidly (*i.e.*, the drug has a short biological half-life), it must be administered at shorter intervals than a drug whose blood concentration declines slowly. The frequency of administration and the dose required to achieve prolonged and effective blood levels often can be deduced from the biological half-life of a drug. Absorption, storage, metabolism, and excretion of a specific drug, however, complicate the relationship between dose and blood concentration.

ACTIONS OF VARIOUS CLASSES OF DRUGS ON MAJOR PHYSIOLOGICAL SYSTEMS

Cardiovascular system. Drugs affecting the cardiovascular system, or cardiovascular drugs, include several drug classes. Cardiac glycosides, or the digitalis group (*e.g.*, digitalis leaves, digoxin, digitoxin, ouabain), increase the force of heart muscle contractions and slow the rate of the heartbeat. Increased contractility, a direct action of these drugs on the heart muscle, may be caused by an increased calcium-ion effect that is concerned with the processes of muscle excitation and contraction, or by inhibition of the enzyme adenosine triphosphatase, which affects the transport of sodium ions and potassium ions through membranes. Irregular heart contractions, which also may occur with digitalis, presumably are caused by decreased intracellular concentrations of potassium ions; slowing of the rate of the heartbeat, however, is an indirect effect involving nerves that innervate the heart. Digitalis also causes nausea and vomiting as a result of stimulation of a specific region of the central nervous system. The digitalis group of drugs is used to treat chronic heart failure; as heart function improves, tissue swelling is reduced, and urine output is increased.

Drugs that affect the heart

Anti-arrhythmic drugs (*e.g.*, quinidine, procaine amide), cardiac depressants used to treat heart irregularities, depress excitability of the heart, prolong the normal recovery time between contractions of the heart, and slow the rate of conduction of nerve impulses through the heart. These direct actions on heart tissues may involve changes in the ability of the membrane to permit passage of certain ions. Other anti-arrhythmic drugs belong to various drug groups; *e.g.*, lidocaine is a local anesthetic, propranolol is a blocking agent, and diphenylhydantoin is an anticonvulsant.

Anti-anginal drugs (*e.g.*, glyceryltrinitrate, other nitrites and nitrates referred to as "the nitrites") are used for relief of chest pain in coronary heart disease, angina pectoris. These drugs act by relaxing the coronary blood vessel, as well as others, so that coronary blood flow increases and blood pressure falls. In high doses, anti-anginal drugs also depress heart muscle contractions and decrease the heart rate. The mechanism of anti-anginal action has not yet been established with certainty but is probably indirect, either an improvement in the oxygen supply to the heart or a reduction in the oxygen required by the heart to meet its work demands.

Drugs that affect the blood include anti-anemia agents (*e.g.*, ferrous salts, vitamin B_{12}, folic acid), anticoagulants (*e.g.*, heparin, dicumarol), and coagulants (*e.g.*, vitamin K, menadione). Ferrous salts act by supplying iron to iron-deficient individuals; vitamin B_{12} and folic acid are used to treat deficiency diseases (*e.g.*, pernicious anemia) and to promote red blood cell formation in bone marrow. Anticoagulants, which prevent clotting and depress protein synthesis in the liver, probably act by forming inactive complexes with special blood-clotting factors. Coagulants supply vitamin K and are useful in controlling excessive bleeding (see CARDIOVASCULAR SYSTEM DISEASES AND DISORDERS).

Smooth and skeletal muscle systems. Several groups of drugs affect smooth muscle structures and skeletal muscle. Uterine muscle stimulants (*e.g.*, ergot, ergot alkaloids such as ergotamine and ergonovine, oxytocin) selectively increase force and rate of the contractions of uterine muscles by exerting a direct action. General smooth muscle stimulants (*e.g.*, histamine, serotonin, angiotensin, vasopressin) are all naturally occurring tissue substances that cause contractions of all smooth muscles (*e.g.*, in arterioles, intestines, bronchioles, uterus). Angiotensin, the most potent vasoconstrictor known, produces an intense and sustained increase in blood pressure. Vasopressin, a potent drug, causes constriction of the coronary blood vessel and deleterious effects on the heart in high doses.

Drugs that affect muscles

General smooth muscle depressants include caffeine, aminophylline, papaverine, and bradykinin; all cause expansion of small blood vessels, called arterioles, and a fall in blood pressure, which is the depressor effect. Bradykinin produces an increase in the permeability of blood vessels to various substances, as does histamine, and stimulates pain receptors in the skin. Smooth muscle stimulants and depressants both exert direct actions on smooth muscle, probably by altering membrane permeability, although the mechanism of their action has not yet been established with certainty.

Skeletal neuromuscular blocking agents, or curare-like drugs, include two main groups. The drugs tubocurarine and gallamine are called competitive blockers; succinylcholine and decamethonium are referred to as depolarizing blockers. These drugs are used to increase muscle relaxation and to overcome skeletal muscle spasms during general anesthesia. All curare-like drugs cause weakness, relaxation, and, ultimately, temporary paralysis of all skeletal muscles. Tubocurarine and related competitive blocking drugs act at receptors on neuromuscular junctions by competing with acetylcholine, a chemical involved in the transmission of nerve impulses, thus blocking its action; the effects of these drugs are antagonized by neostigmine and other drugs. Succinylcholine and decamethonium, the depolarizing blocking drugs, cause a persistent depolarization, or leakage of charge, of muscle membranes; this is responsible for the blocking actions of these drugs, since the membrane can no longer transmit impulses. Depressants of the central nervous system (*e.g.*, minor tranquillizers) also interrupt nerve impulse transmission and thus can cause skeletal muscle relaxation (see MUSCLE DISEASES).

Central nervous system. *Anesthetics.* Drugs affecting the central nervous system (CNS) comprise several large groups of drugs. In the first group are drugs that prevent the perception of inflicted pain; these drugs, called anesthetics, are used to prevent pain during surgery. Two types of anesthetics are inhalation anesthetics (*e.g.*, ether, nitrous oxide, cyclopropane) and intravenous anesthetics (*e.g.*, thiopental sodium). General anesthetics cause rapid partial loss of consciousness and of pain sensation; other effects may include varying degrees of euphoria, excitement, or struggling. Excitatory effects caused by anesthetics are the results of a reduction in the activity of the brainstem and removal of normal inhibitions. Surgical anesthesia results in complete loss of consciousness, analgesia, loss of spinal cord functions, cessation of cell movements, and muscular relaxation; respiration and circulation, however, are maintained. Characteristic changes in the electrical activity of the brain also occur. General anesthetics, whose action involves sequential reduction in the activity of certain regions of the brain and the spinal cord, are soluble in lipids and relatively insoluble in water. Lipid solubility generally correlates with potency; *i.e.*, the more lipid-soluble an anesthetic, the more potent it is.

Local anesthetics

A group of agents known as local anesthetics (*e.g.*, procaine, lidocaine) also are used to prevent the pain of minor surgery (*e.g.*, the extraction of a tooth). Injected locally, these anesthetics cause local nerve blockage. Sensations of pain and temperature and pressure changes are blocked progressively. The actions of local anesthetics depend on the diffusion of molecules into nerve cell membranes and subsequent interference of membrane function (see ANESTHETIC).

Analgesics. Pain-reducing drugs known as analgesics affect the central nervous system. There are two types of analgesics, nonnarcotic analgesics (*e.g.*, acetylsalicylic acid or aspirin, related synthetic agents) and narcotic analgesics, whose use may lead to dependence or addiction. Nonnarcotic analgesics relieve mild pains (*e.g.*, headache pain, rheumatic, joint, or muscle pains), reduce body temperature in fevers, relieve inflammatory pain, and increase uric acid excretion without depressing the CNS. Nonnarcotic analgesics, however, are ineffective in alleviating severe pain caused by injury or diseases in the chest or abdomen. The mechanism of their analgesic action has not yet been established with certainty but may be an effect on the CNS (see ANALGESIC). The narcotic analgesics produce drowsiness and an increased state of well-being (euphoria) or anxiety (dysphoria); relief of pain is followed by sleep. Narcotic analgesics also may elevate the threshold at which pain is perceived, depress psychological responses associated with pain, depress coughing and breathing, and cause nausea, vomiting, narrowing of the pupils of the eyes, increased spinal reflexes, constipation, and contraction of smooth muscles. Narcotic analgesics, therefore, cause both depression and stimulation of brain functions (see below *Drugs that produce stupor and narcosis*).

Tranquillizers and antidepressants. Drugs that stabilize behaviour and reduce anxiety are of two types. Major tranquillizers (*e.g.*, chlorpromazine, reserpine) and minor tranquillizers (*e.g.*, meprobamate, chlordiazepoxide) constitute one type; antidepressants (*e.g.*, imipramine, monoamine oxidase inhibitors) make up the second type. Both major and minor tranquillizers have a calming effect, reduce aggression, and produce withdrawal and drowsiness, but not sleep; these effects are associated with changes in numerous physiological systems. The major tranquillizers are used to treat psychoses (*e.g.*, schizophrenia); the minor ones are used as anti-anxiety drugs. Antidepressant drugs increase mental alertness and elevate mood. The main action of both tranquillizers and antidepressants is a selective depression or stimulation of important functions in the midbrain (see TRANQUILLIZER).

Sedatives and hypnotics. Drugs that quiet a patient or induce sleep are called sedative-hypnotic drugs. They include barbiturates (*i.e.*, derivatives of barbituric acid such as barbital, phenobarbital, amobarbital, secobarbital, pentobarbital) and nonbarbiturates (*e.g.*, chloral hydrate, paraldehyde, inorganic bromides, glutethimide, methyprylon). At one dose level, these drugs calm a patient and relieve tension and anxiety; this is called a

Barbiturates

sedative effect. In higher doses, however, this group of drugs produces sleep; this is called a hypnotic effect. The effects of sedative-hypnotic drugs result from a decrease in certain brain functions equivalent to those that occur during normal sleep; excessive doses lead, however, to general anesthesia, and death may result from respiratory failure (see SEDATIVE-HYPNOTIC DRUGS).

Drugs that produce stupor and narcosis. This group of drugs causes drug dependence; *i.e.*, a state of psychological and/or physical dependency, manifested either by the desire or compulsion to take a drug following its periodic or continued use. Drugs that produce dependence include both CNS depressants (*e.g.*, opiate narcotics, barbiturates, nonbarbiturate sedative-hypnotic drugs, minor tranquillizers, ethyl alcohol) and CNS stimulants (*e.g.*, cocaine, amphetamines, cannabis, mescaline, caffeine, nicotine). The characteristic feature of drug dependence is the development of withdrawal symptoms if use of the drug is discontinued (see NARCOTIC; DRUG PROBLEMS).

Drugs that stimulate consciousness. The drugs that excite or stimulate consciousness are of two types. Adrenergic CNS stimulants, so-called because they stimulate the secretion of an adrenalin-like substance at certain nerve endings, include amphetamines and related compounds. General CNS stimulants, also called convulsants (*e.g.*, strychnine, pentylenetetrazol, picrotoxin), can antagonize excessive CNS depression (see STIMULANT).

Hallucinogens. Drugs that profoundly alter mental and psychological functions are known as hallucinogens or psychotomimetic or psychedelic drugs. Included in this group are LSD (lysergic acid diethylamide), cannabis, tetrahydrocannabinol (THC), mescaline, psilocybin, and various tryptamine derivatives. LSD, which has been studied most extensively, causes variable psychotic-like mental aberrations, delusions, and hallucinations but has only minimal adrenergic-like effects. Cannabis is a mild intoxicant in low doses, and mescaline has been used in religious ceremonies (see HALLUCINOGEN).

Autonomic nervous system. Drugs that affect the autonomic nervous system, which consists of the parasympathetic nervous system and the sympathetic nervous system, can be separated into five groups. Cholinergic drugs (*e.g.*, acetylcholine, muscarine, methacholine, pilocarpine) resemble acetylcholine in their physiological action in that they excite receptors sensitive to acetylcholine at nerve endings; on the other hand, anticholinesterase drugs (*e.g.*, neostigmine, physostigmine) inhibit the reaction in which acetylcholine is broken down at nerve endings. This group of drugs slows the rate of the heartbeat, relaxes vascular smooth muscles, contracts smooth muscles in the intestinal tract, eyes, and lungs, and increases glandular secretions (*e.g.*, sweat, saliva) by directly or indirectly exciting receptors at certain nerve endings. Cholinergic blocking drugs (*e.g.*, atropine, hyoscine or scopolamine, homatropine, various antihistamines, and tranquillizers) cause decreased glandular secretions, decreased muscle contractions in the intestinal tract, eyes, and lungs, constriction of blood vessels, and blockage of skeletal muscle functions by blocking the action of acetylcholine at nerve endings. Adrenergic or sympathomimetic drugs (*e.g.*, levarterenol or norepinephrine, epinephrine or adrenaline, isoproterenol) may act directly at adrenergic receptors or indirectly, either by stimulating the release of the physiological adrenergic mediators epinephrine and norepinephrine or by inhibiting the enzymes involved in their metabolism. The effects of these drugs include an increase in heart rate, contractions and dilation of bronchioles in the lungs, constriction of blood vessels of the skin, mucous membranes, lungs, and brain, and an increase in glandular secretions. Adrenergic-blocking drugs include dibenamine, phenoxybenzamine, phentolamine, ergot alkaloids, dichloroisoproterenol, and propranolol. Indirect blocking agents (*e.g.*, reserpine, guanethidine, monoamine oxidase inhibitors) decrease the functions of the sympathetic nervous system by interfering with storage, binding, release, or metabolism of norepinephrine and epinephrine. Such interference blocks the receptors of the sympathetic nervous system. Ganglionic drugs affect both sympathetic and parasympathetic nerves and therefore cause numerous adverse side effects (see NERVOUS SYSTEM DISEASES).

Epinephrine and norepinephrine

Eyes. Drugs that affect pupillary size and visual accommodation are cholinergic, cholinergic-blocking, and adrenergic agents, which exert local actions on the eyes. Cholinergic drugs and the anticholinesterases cause the following: pupillary constriction due to contraction of circular muscle fibres in the iris, spasm of accommodation resulting from contraction of a ciliary muscle, and increased outflow of fluid from the eye. Cholinergic-blocking agents cause expansion of the pupils by inhibiting circular muscle fibres in the iris, and paralysis of accommodation by relaxing ciliary muscles. Adrenergic or sympathomimetic agents such as epinephrine cause constriction of blood vessels around the eyes, thus reducing reddening of the eyes; these agents cause expansion of the pupils by stimulating contraction of certain muscle fibres of the iris (see EYE DISEASES AND VISUAL DISORDERS).

Excretory system. Drugs that affect the excretory system include diuretics (*e.g.*, chlorothiazide, hydrochlorothiazide, mersalyl), which increase both sodium-ion excretion and water excretion by the kidneys. The diuretics act directly in inhibiting reabsorption of sodium ions and chlorine ions in the kidney tubules. Other diuretics act in different ways. Acetazolamide, for example, increases hydrogen-ion concentration in the kidney tubules by inhibiting an enzyme important in the formation of uric acid. Ammonium chloride and other acid-forming diuretics excreted in urine increase acidity in the tubules; osmotic agents (*e.g.*, mannitol) act after excretion of urine to increase the passage of water into the tubules. Antidiuretics (*e.g.*, vasopressin) decrease urine excretion; by increasing the permeability of the tubule membranes, reabsorption of water is increased. The uricosurics, or gout remedies, may act by increasing the removal of uric acid from blood or by inhibiting an enzyme involved in the normal synthesis of uric acid (see EXCRETORY SYSTEM DISEASES).

Digestive system. Drugs that affect the digestive system are of several types. One type acts on glandular secretion in the stomach; for example, antacids such as sodium bicarbonate, magnesium oxide, and aluminum oxide neutralize excess acid in the stomach. In the liver and biliary tract, bile salts increase bile secretion. Emetic drugs (*e.g.*, strong salt solutions, apomorphine) induce vomiting, act to stop irritation of the stomach, or affect a specific region of the brain. Anti-emetic drugs (*e.g.*, chlorpromazine, anticholinergic and antihistamine drugs) may depress functions of the CNS. Purgatives (*e.g.*, bulk laxatives such as agar and bran) absorb water to increase the bulk of the intestinal contents. Saline laxatives (*e.g.*, sodium sulfate, magnesium sulfate, magnesium oxide), on the other hand, contain poorly-absorbed ions that increase the amount of water retained in the intestines, thus causing reflex intestinal motility or peristalsis (see DIGESTIVE SYSTEM DISEASES).

Reproductive system. Sex hormones include estrogens and progesterones, the female sex hormones, and androgens, the male sex hormones. Estrogens (*e.g.*, estradiol, estrone, ethinyl estradiol, diethylstilbestrol) and progesterones (*e.g.*, progesterone, hydroxyprogesterone, norethindrone, norethylnodrel) are used to treat hormone deficiencies. The purpose of oral contraceptive drugs used by women is the prevention of the release of the egg by suppression of the activity of certain hormones secreted by the pituitary gland. A variety of combinations of progesterone-like drugs (*e.g.*, norethylnodrel) and of estrogens (*e.g.*, ethinylestradiol) is used in oral contraceptives. The androgens (*e.g.*, testosterone, methyl testosterone) are responsible for the development of male sex characteristics (see REPRODUCTIVE SYSTEM DISEASES).

Histamine response system. The drugs that affect the action of histamine are called antihistamines. Many are synthetic agents (*e.g.*, diphenhydramine, tripelennamine, pyrilamine, chlorpheniramine, promethazine). Although

Antihistamines

they are effective in antagonizing the effects of histamine on smooth muscle contractions, blood capillary permeability, and salivary secretion, antihistamines are ineffective in stopping the increased gastric secretion caused by histamine and are not effective in relieving bronchial asthma. Antihistamines do provide some protection, however, from allergy and anaphylaxis, both of which are hypersensitivity reactions. The exact functions of histamine are not yet established. Other histamine-like substances (*e.g.*, serotonin, bradykinin) also are involved in histamine responses, and some antihistamines also show, for example, antiserotonin activity. The antihistamines act as competitive antagonists of histamine; that is, they compete with histamine for specific receptors. This competition, however, does not occur at all histamine receptor sites (see HISTAMINE AND ANTIHISTAMINES).

Immune response system. Drugs affecting the immune response system are called immunosuppressive drugs. Purine analogs (*e.g.*, 6-mercaptopurine, azathioprine) depress the formation of bone marrow, destroy the mucosal lining and blood capillaries of the gastrointestinal tract, and may interfere with nucleic acid synthesis in cells. Corticosteroids (*e.g.*, hydrocortisone, cortisone, prednisone) also exert numerous actions. Antilymphocytic serum (ALS) can also suppress the immune response system. The exact mechanism by which this group of drugs acts is not yet known with certainty, but they do increase the opportunities for viral invasion of tissues (see IMMUNITY).

ACTIONS OF VARIOUS CLASSES OF DRUGS AGAINST INVADING PATHOGENS

Chemotherapy

The primary objective of chemotherapy (*i.e.*, the treatment of disease with chemicals) is to develop chemicals that act selectively against disease-producing micro-organisms (*i.e.*, pathogens) without affecting the diseased host. The actions of a drug on a pathogen may be determined either by establishing the minimal effective concentration of an agent against a specific pathogen or by establishing the ratio (called the chemotherapeutic coefficient) between the dose of a substance that effectively destroys a pathogen and the dose that is toxic to an experimental host that has the disease.

Drugs that affect pathogenic bacteria, termed antimicrobial agents, include the antibiotics, which are substances produced by living organisms such as molds and bacteria, and the synthetic chemotherapeutic agents (*e.g.*, sulfonamides). These drugs may exert bacteriostatic action (*i.e.*, inhibit the growth of bacteria) or bacteriocidal action (*i.e.*, kill bacteria). Differences in susceptibility may allow one bacterial species to resist certain concentrations of a drug that would kill another; the first species thus has natural resistance to the drug and may multiply to produce resistant pathogens (*e.g.*, penicillin-resistant bacteria). Drugs that exert a bacteriostatic action may cause a temporary inhibition of bacterial growth, but natural defense mechanisms of a diseased host may be necessary to destroy the organisms. If the defense mechanisms fail and pathogens survive and multiply, they will have acquired resistance to a specific drug or to a group of drugs (called cross resistance).

Antibiotics. An important group of antibiotics is the penicillins, which includes natural penicillins (*e.g.*, penicillin G, obtained from the growth medium of certain molds), derived or semisynthetic penicillins (*e.g.*, penicillin V, obtained by chemical modification of the growth medium of certain molds), and synthetic penicillins (*e.g.*, cloxacillin, ampicillin, carbenicillin). Penicillin substitutes include cephalosporins, erythromycins, and lincomycin. Other antibiotics include the tetracyclines (*e.g.*, chlortetracycline, oxytetracycline, tetracycline), chloramphenicol, and streptomycin.

The antibiotics are effective against various species of bacteria, and each antibiotic group has a more or less characteristic antimicrobial spectrum; *i.e.*, the penicillins are effective against specific types of micro-organisms that may not be affected by other groups of antibiotics. The mechanism of action of the penicillins involves an interference with a constituent of the bacterial cell wall, which protects the bacterial membrane. If a bacterial cell wall is damaged by a penicillin, the resulting injury to the membrane causes death. Certain bacteria, however, rapidly acquire resistance to the penicillins by synthesizing enzymes that destroy penicillin molecules. Derived and synthetic penicillins are less easily destroyed by these enzymes than are natural penicillins. Penicillin-resistant bacteria are a problem because they retain the ability to cause disease.

Penicillins cause little damage to the cells of host tissues and, therefore, have an extremely high chemotherapeutic coefficient. A disadvantage of natural penicillins (in addition to bacterial resistance) is hypersensitivity (*i.e.*, allergic reactions) that may occur in individuals receiving them; sometimes these reactions result in death. Semisynthetic and synthetic penicillins, however, are less likely to cause severe allergic reactions.

Some antibiotics (*e.g.*, polymyxins, nystatin, amphotericin) act directly on the cell membrane to destroy it. The tetracyclines, chloramphenicol, and streptomycins are effective against many bacteria, certain viruses, and some protozoans. These agents, generally referred to as broad spectrum antibiotics, exert a bacteriostatic action that prevents the multiplication of an invading bacterial species by interfering with various aspects of the process of protein synthesis (see ANTIBIOTIC).

Synthetic chemotherapeutic agents. The sulfonamides include a large number of chemically related synthetic agents (*e.g.*, sulfadiazine, sulfamethoxypyridazine, succinylsulfathiazole, sulfisomidine) with a wide spectrum of antimicrobial activity; they exert both bacteriostatic and bacteriocidal actions and are used mainly to treat urinary infections. Sulfonamides prevent the synthesis of an important metabolite, folic acid, in certain bacteria; *i.e.*, sulfonamides act as antimetabolites. Since sulfonamides do not interfere with bacterial utilization of folic acid, bacteria that can accumulate folic acid from their environment are not affected. In addition, use of sulfonamides may cause hypersensitivity reactions involving the skin and the bone marrow.

Numerous other synthetic chemicals are used in the chemotherapy of bacterial infections (*e.g.*, paraaminosalicylic acid and isoniazid in tuberculosis), protozoal infections (*e.g.*, quinine and chloroquine in malaria), and metazoal infections (*e.g.*, piperazine in roundworm infestations [see CHEMOTHERAPEUTIC DRUGS]).

ACTIONS OF VARIOUS CLASSES OF DRUGS AGAINST CANCER CELLS

Adequate evidence for specific biological distinctions between cancer cells and normal cells has not yet been established with certainty. Although cancer cells show a rapid rate of growth, some normal tissues (*e.g.*, the intestinal mucosa and bone marrow) also have remarkably rapid growth rates; cancer cells, however, also have a higher than normal content of nucleic acids. There are seven main classes of drugs that act against cancer cells. Alkylating agents contain alkyl radicals that attach themselves to the molecules of proteins, nucleic acids, and amino acids in cells. Among the alkylating agents are the nitrogen mustards (*e.g.*, mechlorethamine, chlorambucil, cyclophosphamide), the ethylenimines (*e.g.*, thio-TEPA, triethylenemelamine), and the alkyl sulfonates (*e.g.*, busulfan). All of them block the synthesis of deoxyribonucleic acid (DNA), the hereditary material of the cell. Antimetabolites such as methotrexate and 5-fluorouracil act by interfering with the synthesis of certain chemicals and enzymes necessary to form nucleic acids. Anti-tumour antibiotics (*e.g.*, actinomycin D, mithramycin, mitomycin C) also block nucleic acid synthesis. Radioactive isotopes of iodine, gold, and phosphorus produce destructive effects on tissues and may also induce bone marrow damage. Hormones also may induce tumour regression in some cases. Alkaloids (*e.g.*, vinblastine, vincristine) stop cell division and induce other cellular changes, especially in the bone marrow and hair follicles. Miscellaneous agents (*e.g.*, urethan, colchine) also influence cell division

in various ways. Most drugs employed in cancer chemotherapy lead to immunosuppression (see above *Chemotherapy*) and reduce antibody reactions associated with tissue transplantations, both with respect to the responses of the host against the graft and vice versa. The therapeutic value of such agents is limited by their toxic actions on cells (see CANCER).

BIBLIOGRAPHY

Classical textbooks: J.R. DIPALMA (ed.), *Drill's Pharmacology in Medicine*, 3rd ed. (1965); L.S. GOODMAN and A.Z. GILMAN (eds.), *The Pharmacological Basis of Therapeutics*, 4th ed. (1970); J.C. KRANTZ, JR., C.J. CARR, and B.N. LA DU, JR., *Pharmacologic Principles of Medical Practice*, 7th ed. (1969).

Concise and less complex organization of the general topic may be found in the following works: A.S.V. BURGEN and J.F. MITCHELL, *Gaddum's Pharmacology*, 6th ed. (1968); W.C. CUTTING, *Handbook of Pharmacology: The Actions and Uses of Drugs*, 3rd ed. (1967); A. GOTH, *Medical Pharmacology*, 5th ed. (1970); and J.J. LEWIS, *An Introduction to Pharmacology*, 3rd ed. (1964). Detailed and specialized monographs concerning the fundamental basis of drug action include: E.J. ARIENS (ed.), *Molecular Pharmacology*, 2 vol. (1964); A. GOLDSTEIN, L. ARONOW, and S.M. KALMAN, *Principles of Drug Action: The Basis of Pharmacology* (1968), a recent and up-to-date treatise specifically concerned with basic problems of drug action; and W.C. HOLLAND, R.L. KLEIN, and A.H. BRIGGS, *Introduction to Molecular Pharmacology* (1964).

(K.I.M.)

Drug Problems

Man has been experimenting for thousands of years with a variety of naturally occurring substances that act on his nervous tissues: alcohol to intoxicate a weary mind, belladonna to calm an angry intestine or to poison an adversary, opium to overcome worry and strain. The relief of pain, in particular, is an age-old aim of mankind, and various narcotic and sleep-producing agents were probably used by primitive man. But for many men there is another kind of pain—the pain of being—and from time immemorial some men have been trying to expand their vision, enhance their appreciation of their world, change their mood, alter their inner existence, or stupefy their awareness with such drugs as alcohol, opium, and cannabis. It is written in Genesis (9:20) that Noah planted a vineyard, "and he drank of the wine, and became drunk, and lay uncovered in his tent." Alcohol has been used by many cultures and has been worshipped as a god. Opium has also been used extensively, at least since the time of ancient Greece. Homer tells how some of Odysseus' crew succumbed to forgetfulness in the land of the Lotus eaters, and the ancient Vedic philosophers of India spoke of soma, a mysterious and probably mythical plant. Coca, coffee, and tobacco have also played their parts in history.

The article is divided into the following sections:

I. Characteristics of drug use and abuse
- The functions of psychotropic drugs
- The nature of drug addiction and dependence
  - Popular misconceptions
  - Physiological effects of addiction
  - Addiction, habituation, and dependence
  - Psychological dependence
- History of drug control
  - International controls
  - National controls
  - Extent of contemporary drug abuse

II. The varieties of psychotropic drugs
- Opium, morphine, heroin, and related synthetics
  - History
  - Physiological effects
  - Types of users
  - Means of administration and their effects
  - Therapy
- Hallucinogenic drugs
  - Types of hallucinogens
  - History
  - Physiological and psychological effects
  - Types of users
- Barbiturates, stimulants, and tranquillizers
  - Barbiturates
  - Cocaine
  - Amphetamines
  - Tranquillizers
- Cannabis
  - Types of cannabis preparations
  - History
  - Physiological and psychological effects

III. Social and ethical issues of drug use
- Conflicting values in drug use
- Youth and drugs

I. Characteristics of drug use and abuse

THE FUNCTIONS OF PSYCHOTROPIC DRUGS

To consider drugs only as medicinal agents or to insist that drugs be confined to prescribed medical practice is to fail to understand man. The remarks of the American sociologist Bernard Barber are poignant in this regard:

> Not only can nearly anything be called a "drug," but things so called turn out to have an enormous variety of psychological and social functions—not only religious and therapeutic and "addictive," but political and aesthetic and ideological and aphrodisiac and so on. Indeed, this has been the case since the beginning of human society. It seems that always and everywhere drugs have been involved in just about every psychological and social function there is, just as they are involved in every physiological function.

Anesthetic and hallucinogenic functions

The enhancement of aesthetic experience is regarded by many as a noble pursuit of man. Although there is no general agreement on either the nature or the substance of aesthetics, certain kinds of experience have been highly valued for their aesthetic quality. To Schopenhauer (*The World as Will and Representation*), contemplation was the one requisite of aesthetic experience; a kind of contemplation that enables one to become so absorbed in the quality of what is being presented to the senses that the "Will" becomes still and all needs of the body silent. Drugs reportedly foster this kind of Nirvana and are so used by many today. For Nietzsche (*Birth of Tragedy*), man is able to lose his futile individuality in the mystic ecstasy of universal life under the Dionysiac spell of music, rhythm, and dance. The American Indian with his peyote and the modern jazz musician with his marijuana have discovered this kind of Dionysian ecstasy without formal knowledge of aesthetics.

Love is a highly valued human emotion. Thus, not surprisingly, there has been a great deal of preoccupation with the feeling of love and with those conditions believed to enhance the attainment of love. Little is known concerning the aphrodisiac action of certain foods and drugs, but both have been associated in the mind of man with the increased capacity for love. Though the physiological effects may be doubtful, the ultimate effect in terms of one's feeling of love is probably a potent incentive for the repetition of the experience and for those conditions believed to have produced the experience. Hallucinogenic substances such as LSD are said by many to induce a feeling of lovingness. But what the drug user regards as love and what persons around him regard as love in terms of the customary visible signs and proofs often do not coincide. Even so, it is plausible that the dissipation of tensions, the blurring of the sense of competition, the subsidance of hostility and overt acts of aggression—all have their concomitant effect on the balance between the positive and negative forces within the individual, and, if nothing else, the ability of drugs to remove some of the hindrances to loving is valued by the user.

Native societies of the Western Hemisphere have utilized, apparently for thousands of years, plants containing hallucinogenic substances. The sacred mushrooms of Mexico were called "God's flesh" by the Aztecs. During the 19th century, the Mescalero Apaches of the southwestern United States practiced a peyote rite that was adopted by many of the Plains tribes. Psychedelic drugs have the unusual ability to evoke at least one kind of a mystical-religious experience, and positive change in religious feeling is a common finding in studies of the use of these drugs. Whether they are also capable of producing religious lives is an open question. Their supporters argue that the drugs appear to enhance personal security and that from self-trust may spring trust of others and

that this may be the psychological soil for trust in God. In the words of Aldous Huxley (*The Doors of Perception*): "When, for whatever reason, men and women fail to transcend themselves by means of worship, good works and spiritual exercise, they are apt to resort to religion's chemical surrogates" (see also PHARMACOLOGICAL CULTS).

Stress-reducing and supportive functions

William James (*The Varieties of Religious Experience*) observed at the turn of the century that "Our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different." Some people deliberately seek this other consciousness through the use of drugs; others come upon them by accident while on drugs. Only certain people ever have such a consciousness-expanding (psychedelic) experience in its fullest meaning, and the question of its value to the individual must be entirely subjective. For many people, the search for the psychedelic experience is less a noble aim and more the simple need of a psychic jolt or lift. Man is a paradox of sorts. Although he goes to great lengths to produce order and stability in his life, he also goes to great lengths to disrupt his sense of equanimity, sometimes briefly, sometimes for extended periods of time. It has been asserted that there are moments in everyone's life when uncertainty and a lack of structure are a source of threat and discomfort, and moments when things are so structured and certain that unexpectedness can be a welcome relief. Whatever the reason, people everywhere and throughout history have deliberately disrupted their own consciousness, the functioning of their own ego. Alcohol is and has been a favourite tool for this purpose. With the rediscovery of some old drugs and the discovery of some new ones, man now has a wider variety of means for achieving this end.

Many persons face situations with which, for one reason or another, they cannot cope successfully, and in the pressure of which they cannot function effectively. Either the stresses are greater than usual or the individual's adaptive abilities are less than sufficient. In either instance, there are a variety of tranquillizing and energizing drugs that can provide psychological support. This is not chemotherapy in its more ideal sense, but it does enable large numbers of persons to face problems that they might not have otherwise been able to face. Some situations or stresses are beyond the control of the individual, and some individuals simply find themselves far more human and productive with drugs than without drugs. An enormous amount of drug support goes on by way of such familiar home remedies as the aspirin bottle, the luncheon cocktail, and the customary evening drink without anyone calling it that. There is no clear dividing line between drug support and drug therapy. It is all therapy of sorts, but deliberate drug manipulation is a cut different from drug buffering, and much of the psychological support function is just that—taking the "raw edge" off of stress and stabilizing responses.

Therapeutic functions

The therapeutic use of drugs is so obvious as to require little explanation. Many of the chemical agents that affect living protoplasm are not capable of acting on the brain, but some of those that do are important in medical therapeutics. Examples are alcohol, the general anesthetics, the analgesic (pain-killing) opiates, and the hypnotics, which produce sleep—all classified as central-nervous-system depressants. Certain other drugs, such as strychnine, nicotine, picrotoxin, caffeine, cocaine, and the amphetamines, stimulate the nervous system. Most drugs truly useful in the treatment of mental illness, however, were unknown to science until the middle of the 20th century. With the discovery of reserpine and chlorpromazine, some of the major forms of mental illness, especially the schizophrenias, became amenable to chemotherapeutic treatment. These tranquillizing drugs seem to reduce the incidence of certain kinds of behaviour, particularly hyperactivity and agitation. A second group of drugs has achieved popularity in the management of milder psychiatric conditions, particularly those in which patients manifest anxiety. This group includes drugs that have a mild calming or sedative effect and that are also useful in inducing sleep. Not all drugs in psychiatric use have a tranquillizing action. The management of depression requires a different pharmacological effect, and the drugs of choice have been described as being euphorizing, mood-elevating, or antidepressant, depending on their particularly pharmacological properties. There are also drugs useful in overactive states such as epilepsy and Parkinsonism. The so-called psychedelic drugs also may have therapeutic uses.

Drugs have other functions that are not so intimately related to individual use. Several important early studies in physiology were directed toward understanding the site and mode of action of some of these agents. Such studies have proved indispensable to the understanding of basic physiology, and drugs continue to be a powerful research tool of the physiologist. The ability of drugs to alter mental processes and behaviour affords the scientist the unique opportunity to manipulate mental states or behaviour in a controlled fashion. The use of LSD to investigate psychosis and the use of scopolamine to study the retention of learning are examples. The use of drugs as potential instruments of chemical and biological warfare has received wide public attention and scorn. The political use of drugs is a frightening possibility. Whether as a "truth serum," a "brainwashing" technique, a way of destroying certain stable elements of culture, or a way of reducing entire societies to a tranquil slavery, this aspect of drug use should be viewed with alarm because all such uses are obviously possible.

THE NATURE OF DRUG ADDICTION AND DEPENDENCE

If opium were the only drug of abuse, and the only kind of abuse were one of habitual, compulsive use, discussion of addiction might be a simple matter. But opium is not the only drug of abuse, and there are probably as many kinds of abuse as there are drugs to abuse, or, indeed, as maybe there are persons who abuse. Various substances are used in so many different ways by so many different people for so many different purposes that no one view or one definition could possibly embrace all the medical, psychiatric, psychological, sociological, cultural, economic, religious, ethical, and legal considerations that have an important bearing on addiction. Prejudice and ignorance have led to the labelling of all use of nonsanctioned drugs as addiction and of all drugs, when misused, as narcotics. The continued practice of treating addiction as a single entity is dictated by custom and law, not by the facts of addiction.

Confusion of drug abuse and narcotic addiction

The tradition of equating drug abuse with narcotic addiction originally had some basis in fact. Until recent times, questions of addiction centred on the misuse of opiates, the various concoctions prepared from powdered opium. Then various alkaloids of opium, such as morphine and heroin, were isolated and introduced into use. Being the more active principles of opium, their addictions were simply more severe. More recently, new drugs such as methadone and demarol were synthesized but their effects were still sufficiently similar to those of opium and its derivatives to be included in the older concept of addiction. With the introduction of various barbiturates in the form of sedatives and sleeping pills, the homogeneity of addictions began to break down. Then came various tranquillizers, stimulants, new and old hallucinogens, and the various combinations of each. At this point, the unitary consideration of addiction became untenable. Legal attempts at control often forced the inclusion of some nonaddicting drugs into old, established categories—such as the practice of calling marijuana a narcotic. Problems also arose in attempting to broaden addiction to include habituation and, finally, drug dependence. Unitary conceptions cannot embrace the diverse and heterogeneous drugs currently in use.

Popular misconceptions. The bewilderment that the public manifests whenever a serious attempt is made to differentiate states of addiction or degrees of abuse probably stems from two all-too-common misconceptions

concerning drug addiction. The first involves the stereotype that a drug user is a socially unacceptable criminal. The carry-over of this conception from olden times is easy to understand but not very easy to accept today. Ironically, the so-called dope fiend, if indeed one does exist, is likely to be a person who is not using an opiate. The depressant action of opium and its derivatives is simply not consistent with the stereotype. The second misconception involves the naïve belief that there is something magically druglike about a drug, which makes a drug a drug. Many substances are capable of acting on a biological system, and whether a particular substance comes to be considered a drug depends, in large measure, upon whether it is capable of eliciting a "druglike" effect that is valued by the user. There is nothing intrinsic to the substances themselves that sets one active substance apart from other active substances; its attribute as a drug is imparted to it by use. Caffeine, nicotine, and alcohol are clearly drugs, and the habitual, excessive use of coffee, tobacco, or an alcoholic drink is clearly drug dependence if not addiction. The same could be extended to cover tea, chocolates, or powdered sugar, if society wished to use and consider them that way. The task of defining addiction, then, is the task of being able to distinguish between opium and powdered sugar while at the same time being able to embrace the fact that both can be subject to abuse. This requires a frame of reference that recognizes that almost any substance can be considered a drug, that almost any drug is capable of abuse, that one kind of abuse may differ appreciably from another kind of abuse, and that the effect valued by the user will differ from one individual to the next for a particular drug, or from one drug to the next drug for a particular individual. This kind of reference would still leave unanswered various questions of availability, public sanction, and other considerations that lead people to value and abuse one kind of effect rather than another at a particular moment in history, but it does at least acknowledge that drug addiction is not a unitary condition.

Physiological effects of addiction. Certain physiological effects are so closely associated with the heavy use of opium and its derivatives that they have come to be considered characteristic of addictions in general. Some understanding of these physiological effects is necessary in order to appreciate the difficulties that are encountered in trying to include all drugs under a unitary definition that takes as its model opium. *Tolerance* is a physiological phenomenon that requires the individual to use more and more of the drug in repeated efforts to achieve the same effect. At a cellular level this is characterized by a diminishing response to a foreign substance (drug) as a result of adaptation. Although opiates are the prototype, a wide variety of drugs elicit the phenomenon of tolerance, and drugs vary greatly in their ability to develop tolerance. Opium derivatives rapidly produce a high level of tolerance; alcohol and the barbiturates a very low level of tolerance. Tolerance is characteristic for morphine and heroin and, consequently, is considered a cardinal characteristic of narcotic addiction. In the first stage of tolerance, the duration of the effects shrinks, requiring the individual to take the drug either more often or in greater amounts to achieve the effect desired. This stage is soon followed by a loss of effects, both desired and undesired. Each new level quickly reduces effects until the individual arrives at a very high level of drug with a correspondingly high level of tolerance. Man can become almost completely tolerant to 5,000 milligrams of morphine per day, even though a "normal," clinically effective dosage for the relief of pain would fall in the 5 to 20 milligram range. An addict can achieve a daily level that is nearly 200 times the dose that would be dangerous for a normal, pain-free adult.

Characteristics of the development of tolerance

Tolerance for a drug may be completely independent of the drug's ability to produce *physical dependence*. There is no wholly acceptable explanation for physical dependence. It is thought to be associated with central-nervous-system depressants, although the distinction between depressants and stimulants is not as clear as it was once thought to be. Physical dependence manifests itself by the signs and symptoms of abstinence when the drug is withdrawn. All levels of the central nervous system appear to be involved, but a classic feature of physical dependence is the "abstinence" or "withdrawal" syndrome. If the addict is abruptly deprived of a drug upon which the body has physical dependence, there will ensue a set of reactions, the intensity of which will depend on the amount and length of time that the drug has been used. If the addiction is to morphine or heroin, the reaction will begin within a few hours of the last dose and will reach its peak in one to two days. Initially, there is yawning, tears, a running nose, and perspiration. The addict lapses into a restless, fitful sleep and, upon awakening, experiences a contraction of pupils, gooseflesh, hot and cold flashes, severe leg pains, generalized body aches and constant movement. The addict then experiences severe insomnia, nausea, vomiting, and diarrhea. At this time he has a fever, mild high blood pressure, loss of appetite, dehydration, and a considerable loss of body weight. These symptoms continue through the third day and then decline over the period of the next week. There are variations in the withdrawal reaction for other drugs; in the case of the barbiturates, minor tranquillizers, and alcohol, withdrawal may be more dangerous and severe. During withdrawal, drug tolerance is lost rapidly. The withdrawal syndrome may be terminated at any time by an appropriate dose of the addicting drug.

Physical dependence

Addiction, habituation, and dependence. The traditional distinction between "addiction" and "habituation" centres on the ability of a drug to produce tolerance and physical dependence. The opiates clearly possess the potential to massively challenge the body's resources, and, if so challenged, the body will make the corresponding biochemical, physiological, and psychological readjustment to the stress. At this point, the cellular response has so altered itself as to require the continued presence of the foreign substance (drug) to maintain normal function. When the substance is abruptly withdrawn or blocked, the cellular response becomes abnormal for a time until a new readjustment is made. The key to this kind of conception is the massive challenge that requires radical adaptation. Some drugs challenge easily, but it is not so much whether a drug can challenge easily as it is whether the drug was actually taken in such a way as to present the challenge. Drugs such as caffeine, nicotine, bromide, the salicylates, cocaine, amphetamine and other stimulants, certain tranquillizers and sedatives are normally not taken in sufficient amounts to present the challenge. They typically but not necessarily induce a strong need or craving emotionally or psychologically without producing the physical dependence that is associated with "hard" addiction. Consequently, their propensity for potential danger is judged to be less, so that continued use would lead one to expect habituation but not addiction. The key word here is *expect*. These drugs, in fact, are used excessively on occasion and, when so used, do produce tolerance and withdrawal signs. Morphine, heroin, other synthetic opiates, and to a lesser extent codeine, alcohol, and the barbiturates, all carry a high propensity for potential danger in that all are easily capable of presenting a bodily challenge. Consequently, they are judged to be addicting under *continued* use. The ultimate effect of a particular drug, in any event, depends as much or more on the setting, the expectation of the user, his personality, and the social forces that play upon him, as it does on the pharmacological properties of the drug itself.

Distinctions between addiction and habituation

Enormous difficulties have been encountered in trying to apply these definitions of *addiction* and *habituation* because of the wide variations in the pattern of use. (The one common denominator in drug use is variability.) As a result, in 1964 the World Health Organization recommended a new standard that replaces both the term drug addiction and the term drug habituation with the term drug dependence. Drug dependence is defined as a state arising from the repeated administration of a drug on a periodic or continual basis. Its characteristics will vary with the agent involved, and this must be made clear by

designating drug dependence as being of a particular type —that is, drug dependence of morphine type, of cannabis type, of barbiturate type, and so forth. As an example, drug dependence of a cannabis (marijuana) type is described as a state involving repeated administration, either periodic or continual. Its characteristics include (1) a desire or need for repetition of the drug for its subjective effects and the feeling of enhancement of one's capabilities that it effects, (2) little or no tendency to increase the dose since there is little or no tolerance development, (3) a psychic dependence on the effects of the drug related to subjective and individual appreciation of those effects, and (4) absence of physical dependence so that there is no definite and characteristic abstinence syndrome when the drug is discontinued.

Drug dependence

Considerations of tolerance and physical dependence are not prominent in this new definition, although they are still conspicuously present. Instead, the emphasis tends to be shifted in the direction of the psychological or psychiatric makeup of the individual and the pattern of use of the individual and his subculture. Several considerations are involved here. There is the concept of psychological reliance in terms of both a sense of well-being and a permanent or semipermanent pattern of behaviour. There is also the concept of gratification by chemical means that has been substituted for other means of gratification. In brief, the drug has been substituted for adaptive behaviour. Terms such as hunger, need, craving, emotional dependence, habituation or psychological dependence tend to connote a reliance on a drug as a substitute gratification in the place of adaptive behaviour.

Psychological dependence. Several explanations have been advanced to account for the psychological dependence on drugs, but as there is no one entity called addiction, so there is no one picture of the drug user. The great majority of addicts display "defects" in personality. Several legitimate motives of man can be fulfilled by the use of drugs. There is the relief of anxiety, the seeking of elation, the avoidance of depression, and the relief of pain. For these purposes, the several potent drugs are equivalent, but they do differ in the complications that ensue. Should the user develop physical dependence, euphoric effects become difficult to attain, and the continued use of the drug is apt to be aimed primarily at preventing withdrawal symptoms.

Predispositions to drug use

It has been suggested that drug use can represent a primitive search for euphoria, an expression of prohibited infantile cravings, or the release of hostility and of contempt; the measure of self-destruction that follows can constitute punishment and the act of expiation. This type of psychodynamic explanation assumes that the individual is predisposed to this type of psychological adjustment prior to any actual experience with drugs. It has also been suggested that the type of drug used will be strongly influenced by the individual's characteristic way of relating to the world. The detached type of person might be expected to choose the "hard" narcotics to facilitate indifference and withdrawal from the world. Passive and ambivalent types might be expected to select sedatives to assure a serene dependency. Passive types of persons who value independence might be expected to enlarge their world without social involvement through the use of hallucinogenic drugs, whereas the dependent type of person who is geared to activity might seek stimulants. Various types of persons might experiment with drugs simply in order to play along with the group that uses drugs; such group identification may be joined with youthful rebellion against society as a whole. Obviously, the above descriptions are highly speculative because of the paucity of controlled clinical studies. The quest of the addict may be the quest to feel full, sexually satisfied, without aggressive strivings, and free of pain and anxiety. Utopia would be to feel normal, and this is about the best that the narcotic addict can achieve by way of drugs.

Although many societies associate addiction with criminality, most civilized countries regard addiction as a medical problem to be dealt with in appropriate therapeutic ways. Furthermore, narcotics fulfill several socially useful functions in those countries that do not prohibit or necessarily censure the possession of narcotics. An old League of Nations report said: "The social and hygienic conditions under which a great part of the working classes in the Far East live are of so low a standard that these classes of people strive to find some form of diversion permitting them to forget at least for some moments the hardships of life." In addition to relieving mental or physical pain, opiates have been used medicinally in tropical countries where large segments of the population suffer from dysentery and fever.

HISTORY OF DRUG CONTROL

The first major national efforts to control the distribution of narcotic and other dangerous drugs were the efforts of the Chinese in the 19th century. Commerce in poppy (opium) and coca leaf (cocaine) developed on an organized basis during the 1700s. The Manchus of China attempted to discourage opium importation and use, but the English East India Company, which maintained an official monopoly over British trade in China, was engaged in the profitable export of opium from India to China. This monopoly of the China trade was eventually abolished in 1839–42, and friction increased between the British and the Chinese over the importation of opium. Foreign merchants, including those from France and the United States, were bringing in ever-increasing quantities of opium. Finally, the Manchu government required all foreign merchants to surrender their stocks of opium for destruction. The British objected, and the Opium War (1839–42) between the Chinese and the British followed. The Chinese lost and were forced into a series of treaties with England and other countries that took advantage of the British victory. In 1858, the importation of opium into China was legalized by the Treaty of Tientsin, which fixed a tariff rate for opium importation. Further difficulties followed. An illegal opium trade carried on by smugglers in south China encouraged gangsterism and piracy, and the activity eventually became linked with powerful secret societies in the south of China.

Chinese opium traffic

International controls. Throughout the 1800s, the Chinese government regarded opium as an important moral and economic question, but obviously China needed international help. In 1909, U.S. President Theodore Roosevelt proposed an international investigation of the opium problem, and a meeting of 13 nations held in Shanghai in the same year resulted in recommendations that formed the basis of the first opium convention at The Hague in 1912. Ratification of the Hague Convention occurred during the meetings of 1913 and 1914. Although further regulatory activity was suspended during World War I, ratification of the Versailles peace treaties of 1919–20 also constituted a ratification of the Hague Convention of 1912. The League of Nations was then given responsibility to supervise agreements with regard to the traffic in opium and other dangerous drugs. A further important development in drug control was the convention of 1925, which placed further restrictions on the production and manufacture of narcotics. Six more international conventions and agreements were concluded between 1912 and 1936.

League of Nations and United Nations efforts

Under a Protocol on Narcotic Drugs of December 1946 the functions of the League of Nations and of the Office International d'Hygiène Publique were transferred to the United Nations and to the World Health Organization. In 1948 a protocol extended the control system to synthetic and natural drugs outside the scope of the earlier conventions. In 1953 a further protocol was adopted to limit and regulate the cultivation of the poppy plant and the production of, or international and wholesale trade in, and use of opium. Before the protocol became operative in 1963 the international control organs found a need for codifying and strengthening the existing treaties, and a Single Convention on Narcotic Drugs was drawn up in New York in 1961. This Convention drew into one comprehensive control regime all the earlier agreements, limited the use of coca leaves and cannabis

to medical and scientific needs, and paved the way for the International Narcotics Control Board. The Convention came into force in 1964, and the new board began duty in 1968.

National controls. The United States is perhaps the nation most preoccupied with drug control, and it is largely the "Americanized" countries that have made narcotics regulation a matter of public policy with the consequent network of laws, criminal detection agencies, and derived social effects. The principal U.S. legislation has been the Harrison Narcotics Act of 1914, the Opium Poppy Control Act of 1942, and the Narcotic Drug Control Act of 1956; the Drug Abuse Control Amendment of 1965 added controls over depressant, stimulant, and hallucinogenic drugs not covered under the other narcotic control acts. Manufacturers and distributors are required to register with the U.S. Food and Drug Administration, retail dealers are required to keep inventories, and physicians are required to limit the period of prescription and the number of refills permitted. Heroin manufacture was prohibited in the United States in 1924, and by 1956 all heroin legally held in the United States was surrendered to the government. The *legal* use of heroin is practically nonexistent today anywhere in the world—largely because of the action of the League of Nations.

Great Britain controls the manufacture, distribution, and sale of narcotics through the Dangerous Drug Act of 1920. The British system, however, is based on a public policy position different from that of the United States, and narcotic addiction has remained a minor social problem. In 1967, England placed the prescription of narcotics under the control of the National Health Service and its associated clinics. Canada and Japan attempt to control narcotics in much the same manner as the United States with much the same consequence in terms of high rates of addiction. Opium traffic appears to still flourish in Asia, but the East has officially gone through a period of regulation, governmental monopoly over cultivation, and finally prohibition of use.

Extent of contemporary drug abuse. Complete and reliable data on the extent of drug abuse in recent years is simply not available. To specify the size and extent of the drug problem, accurate information as to manufacture, distribution, and sale of drugs would be needed. Complete evaluation would also require knowledge of the incidence of habituation and addiction in the general population, the number of persons admitted to hospitals because of drug intoxication, and the number of arrests for drug sales that do not conform to the law. This kind of determination is not possible under existing laws even for the legitimate sources of drugs. Unfortunately, much of the drug traffic is from uncontrolled, illicit sources, and here there is an almost total absence of reliable information. Black market diversion of drugs may occur at any point from the manufacture of basic chemicals used to synthesize the drugs, through the process of actually preparing the drug, to the distribution of the final drug form, to the retail drug store, or even the physician. This is a complex chain involving chemical brokers, exporters, and dealers in addition to those who have more direct involvement with the drug process. Finally, there is the problem of currentness. Time is involved in reporting at each level of information, and the final data may be no more recent than three to five years, and a basic source or reference work may contain figures that have suffered a decade of delay from the actual occurrence of the drug abuse. Drug abuse patterns change over a relatively short time. In the 1960s, the youthful drug abuser tended to use a drug that increased the level of consciousness. In contrast, it had been only a short time earlier that youthful drug abuse involved only the hypnotics and alcohol, which depress consciousness and blunt experience.

Problems in measuring distribution and use

The U.S. Federal Bureau of Narcotics maintains a register of so-called active addicts who in the early 1970s numbered approximately 60,000 persons; the actual number of narcotic addicts in the United States was probably closer to 120,000 persons. England has about 400 addicts but this number has been steadily growing as a result of an influx of narcotic addicts from Canada. With the exception of Germany, other European countries have no more than a few hundred narcotic addicts each. West Germany has close to 5,000 addicts. Canada has from 3,000 to 3,500 addicts. In the Middle and Far East, the highest rates are found in Egypt, India, and Iran, with lesser rates for Borneo, Burma, China, Japan, and Korea. Opium cultivation has been declining in each of the last several years and is found chiefly in Asia. In 1965, India (55.8 percent), the U.S.S.R. (25.8 percent), and Turkey (17.7 percent) accounted for almost the total output of opium as reported to the United Nations. Bolivia and Peru are the chief producers of coca leaf (for cocaine).

The extent of problems involving other drugs can only be guessed from certain superficial indicators. By 1966, for example, approximately 8,000,000,000 amphetamine tablets were being produced yearly in the United States. This would give each adult and child 35 doses of a five-milligram tablet each year. It is probable that about half the production is diverted into illegal channels, and that 90 percent of the illegal traffic emanates from restaurants, truck stops, gasoline stations, and bars. The barbiturate figures are older (1957), but production was estimated to be 864,000 pounds or 26 doses of 100 milligrams for each adult and child yearly. Barbiturates are the leading mode of suicide in the U.S., accounting for over 3,000 deaths each year. A 1967 report to the U.S. president on medical care prices indicated that retail sales of sedatives and tranquillizers had increased 535 percent between the years 1953 and 1965. Marijuana and the hallucinogens are illegal, and no good data are available.

II. The varieties of psychotropic drugs

OPIUM, MORPHINE, HEROIN, AND RELATED SYNTHETICS

Types of opiates and related synthetics

The opiates are unrivalled in their ability to relieve pain. Opium is the dried milky exudate obtained from the unripe seed pods of the poppy plant (*Papaver somniferum*), which grows naturally throughout most of Asia Minor. Of the 20 or more alkaloids found in opium, only a few are pharmacologically active. The important constituents of opium are morphine (10%), papaverine (1%), codeine (0.5%), and thebaine (0.2%). (Papaverine is pharmacologically distinct from the narcotic agents and is essentially devoid of effects on the central nervous system.) In 1803, a young German apothecary's assistant named Sertürner isolated crystalline morphine as the active analgesic principle of opium. Codeine is considerably less potent (⅙) and is obtained from morphine. Diacetylmorphine—or heroin—was developed from morphine by the Bayer Company of Germany in 1898 and is five to ten times as potent as morphine itself. Opiates are not medically ideal. Tolerance is developed quite rapidly and completely in the more important members of the group, morphine and heroin, and they are highly addictive. In addition, they produce respiratory depression and frequently cause nausea and emesis. As a result, there has been a constant search for synthetic substitutes: meperidine (Demerol), first synthesized in Germany in 1939, is a significant addition to the group of analgesics, being one-tenth as potent as morphine; alphaprodine (Nisentil) is one-fifth as potent as morphine but is rapid-acting; methadone, synthesized in Germany during World War II, is comparable to morphine in potency; levorphanol (Levo-Dromoran) is an important synthetic with five times the potency of morphine. These synthetics exhibit a more favourable tolerance factor than the more potent of the opiates, but they fall short of an ideal analgesic in being addictive. Of this entire series, codeine has the least addiction potential and heroin has the greatest (see also NARCOTIC).

History. The narcotic and sleep-producing qualities of the poppy have been known to man throughout recorded history. Sumerian records from the time of Mesopotamia (5000 to 4000 BC) refer to the poppy, and medicinal reference to opium is contained in Assyrian medical tablets. Homer's writings indicate Greek usage at least by 900

BC; Hippocrates (*c.* 400 BC) made extensive use of medicinal herbs including opium. The Romans probably learned of opium during their conquest of the eastern Mediterranean; Galen (AD 130–200) was an enthusiastic advocate of the virtues of opium, and his books became the supreme authority for hundreds of years. The art of medicinals was preserved by the Islāmic civilization following the decline of the Roman Empire; opium was introduced by the Arabs to Persia, China, and India. Paracelsus (1493–1541), professor at the University of Basle, introduced laudanum, the modern tincture of opium. Le Mort, a professor of chemistry at the University of Leyden (1702–18), discovered paregoric for the control of diarrhea by combining camphor with tincture of opium.

19th-century medical uses of opiates

There is no adequate comprehensive history of the addictive aspects of opium use in spite of the fact that it has been known since antiquity. Because there were few alternative therapeutics or pain-killers until the 19th century, opium was somewhat of a medical panacea. Thus, although at least one account in 1701, by a London physician named Jones, spoke of an excessive use of opium, there appears to have been no real history of concern until recent times, and opiates were easily available in the West in the 19th century, for instance, in a variety of patent medicines. Physicians prescribed them freely, they were easy to obtain without prescription, and they were used by all social classes. At one time, the extensive use of these medicines for various gynecological difficulties probably accounted for the high addiction rate among women (three times the rate among men). Today, in the United States, only one addict out of six is a woman. The invention of the hypodermic needle in the mid-19th century and its subsequent use to administer opiates during wartime produced large numbers of addicted soldiers (about 400,000 during the U.S. Civil War alone); it was thought mistakenly that if opiates were administered by vein, no hunger or addiction would develop, since the narcotic did not reach the stomach. Toward the end of the 19th century, various "undesirables" such as gamblers and prostitutes began to be associated with the use of opiates, and narcotics became identified more with the so-called criminal element than with medical therapy. By the turn of the 20th century, narcotic use had become a worldwide problem, and various national and international regulatory bodies sought to control traffic in opium from the Near and Far East.

20th-century patterns of use

In the 20th century, until recently, narcotic use was largely associated with metropolitan slums, principally among poor and culturally deprived. Currently, narcotic use has begun to spread to middle class youth, and, interestingly, there is evidence that the middle class is now beginning to look at narcotic addiction as a mental health problem. When it was confined to the slums, it was considered a police problem.

Physiological effects. The various opiates and related synthetics all produce about the same physiological effects. All are qualitatively similar to morphine in action and differ from each other mainly in degree. The most long-lasting and conspicuous physiological responses are obtained from the central nervous system and the smooth muscle of the gastrointestinal tract. These effects, while restricted, are complex and vary with the dosage and the route of administration (intravenous, subcutaneous, oral). Both depressant and stimulant effects are elicited. The depressant action involves the cerebral cortex, with a consequent narcosis, general depression, and reduction in pain perception; it also involves the hypothalamus and brain stem, inducing sedation, the medulla, with associated effects on respiration, the cough reflex, and the vomiting centre (late effect). The stimulant action involves the spinal cord and its reflexes, the vomiting centre (early effect), the tenth cranial nerve with a consequent slowing of the heart, and the third cranial nerve resulting in pupil constriction. Associated effects of these various actions include nausea, vomiting, constipation, itchiness of the facial region, yawning, sweating, flushing of skin, a warm sensation in the stomach, fall in body temperature, diminished respiration, and heaviness in the limbs.

Opiates as analgesics and depressants

The most outstanding effect of the opiates is one of analgesia. All types of pain perception are affected, but the best analgesic response is obtained in relieving dull pain. The analgesic effects increase with increasing doses until a limit is reached beyond which no further improvement is obtained. This point may fall just short of complete relief.

Depression of cortical function results in a euphoric response involving a reduction of fear and apprehension, a lessening of inhibitions, an expansion of ego, and an elevation of mood that combine to enhance the general sense of well-being. Occasionally in pain-free individuals, the opposite effect, dysphoria, occurs and there is anxiety, fear, and some depression. In addition to analgesia and associated euphoria, there is drowsiness, mental and physical impairment, a clouding of consciousness, poor concentration and attention, reduced hunger or sex drives, and sometimes apathy.

Apart from their addiction liability, respiratory depression leading to respiratory failure and death is the chief hazard of these drugs. All of the more potent opiates and synthetics produce rapid tolerance, and tolerance to one member of this group always is associated with tolerance to the other members of the group (cross-tolerance). The more potent members of the group have a very great addiction liability with the associated physical dependence and abstinence syndrome.

Types of users. There is no single narcotic addict personality type: addiction is not a unitary phenomenon occurring in a single type. The great variation in addiction rates and classes of addicts in various countries caution against placing too great an emphasis on personality variables as major causative factors. Even within the United States, there is great danger in generalizing from the cases of the patients found at the public health service hospitals in Lexington, Kentucky, and in Fort Worth, Texas. These inmates are a highly select group of adults who have spent previous time in correctional institutions. They are not representative of the adolescent addict or the adult addict who has not had continual difficulty with the law. The United States has recently experienced a new type of slum-dwelling addict who is a member of a closely knit adolescent gang. This subculture is highly tolerant of drug abuse, and the members have ready access to narcotic drugs. They do not actively seek the opportunity to try heroin; neither are they deliberately "hooked" on heroin by adult drug peddlers. They are initiated to narcotic use by friends, gang members, or neighbourhood acquaintances, and the opportunity for such use is almost always casual but ever present. This "kicks" user is apt to abandon narcotics when gang membership is abandoned. The chronic user is more likely to be the immature adolescent at the periphery of gang activities who uses narcotics for their adjustive value in terms of his deep-seated personality problems. He does not abandon drug use for the more conventional pursuits when he enters adulthood. Instead, old ties are severed; interest in previous friendships is withdrawn; athletic and scholastic strivings are abandoned; competitive, sexual, and aggressive behaviour becomes markedly reduced, and he retreats further into his drug-induced state. Identification is now with the addict group: a special culture with a special language. The addict's world revolves around obtaining drugs.

Emotional and physical feelings of heroin use

Means of administration and their effects. Most persistent users follow a classic progression from sniffing (similar to the oral route), to "skin popping" (subcutaneous route), to "mainlining" (intravenous route)—each step bringing more intense experience, a higher addiction liability. With mainlining, the initial "thrill" is more immediate. Within seconds, a warm, glowing sensation spreads over the body, most intense in the stomach and intestines, comparable to sexual release. This intense "rush" is then followed by a deep sense of relaxation and contentment. The user is "high" and momentarily free. It is this initial state of intense pleasure that presumably brings the novice to repeat the experience, and it is this mode of administration that hastens him on the way to

drug tolerance and physical dependence. Soon he finds that the effects are not quite there. Instead, his body is beginning to experience new miseries. At this juncture, he "shoots" to avoid discomfort. The euphoria is gone. He now spends every waking moment in obtaining further supplies to prevent the inevitable withdrawal symptoms should he fail. Habits are expensive. If indigent, the addict must spend all his time "hustling" for drugs—which means that he must steal and dispose of $50,000 worth of goods to support a habit that may cost $10,000 a year or raise money by other means such as prostitution, procuring, or small-time narcotics peddling. An addict is judged by his success in supporting his habit. The addict always faces the danger of withdrawal, the danger of arrest, the danger of loss of available supply, the danger of infection, of collapsed veins, or of death from overdosage. Very few individuals are still addicted by the age of 40. They have either died, somehow freed themselves from their addiction, or sought treatment.

Therapy. Drug dependence can be viewed as an ethical problem: Is it right and permissible to need a narcotic agent? How one answers this question dictates the position one will take in regard to addiction therapy. In general, the addict can be given his drug, he can be placed on a substitute drug, or drugs can be barred from him. Narcotic maintenance, which gives the addict his drug, is the system employed in the management of opiate dependence in England. Methadone treatment is a drug substitution therapy that replaces opiate addiction with methadone addiction in order that the addict might become a socially useful citizen. Synanon, a therapeutic community movement involving an intensive program of family-like resocialization, is an approach with total abstinence as its goal. Psychological approaches to total abstinence through re-education involve psychotherapy, hypnosis, and various conditioning techniques that attempt to attach unpleasant or aversive associations to the thoughts and actions accompanying drug use. Each of these approaches has had successes and has limitations.

Narcotic maintenance

Great Britain began to control the use of narcotics in 1950 but, unlike the United States, has embraced the principle of drug maintenance. Supporters of the approach have insisted that, at least until recently, narcotic addiction in Great Britain remained a very minor problem because addiction is considered an illness rather than a crime. (In recent years, however, addiction has apparently become more widespread, for uncertain reasons.) The British physician was earlier allowed to prescribe maintenance doses of a narcotic if, in his professional judgment, the addict was unable to lead a useful life without the drug. But in 1967 the British government took the right to prescribe for maintenance addiction away from the general practitioner and placed it in the hands of drug-treatment clinics. Although some addicts must obtain legal supplies from the clinic, others are allowed to obtain supplies from a neighbourhood pharmacy and medicate themselves. These clinics also provide social and re-educative services such as psychotherapy for the addict. The general experience among these clinics has been that a large proportion of the addicts are becoming productive, socially useful members of the community.

There are two major drawbacks to the maintenance use of narcotic drugs such as employed by the British. Both the physical and the social health of the user remains unsatisfactory. A high incidence of hepatitis, bacterial endocarditis, abcesses, and, on occasion, fatal overdosage accompanies the self-administration of opiates. Socially, the addict on self-administration also tends to remain less productive than his peers—the reason apparently being that the individual on narcotic maintenance is still very preoccupied with certain aspects of narcotic use. Narcotic addiction is a two-faceted problem: the yearning for the "high" and the felt sense of not being physiologically normal. The addict on narcotic maintenance often attempts to obtain or retain both drug effects: frequent intravenous use prevents the feeling of drug hunger and maximizes his attempt to experience euphoria.

Methadone therapy

Methadone therapy aims to block the abnormal reactions associated with narcotic addiction while permitting the addict to live a normal, useful life as a fully participating member of the community. Methadone provides a "narcotic blockade" in that it is possible to increase methadone medication to a point at which large oral doses will induce a state of cross-tolerance in which the euphoric effects of other narcotics cannot be felt even in very high doses. Additionally methadone has the ability to allay the feeling of not being right physically, which the addict finds he can correct only by repeated narcotic use. Methadone treatment, then, rests on these two pharmacological actions: the blockade of euphoric effects and the relief of "narcotic hunger." Methadone is not successful in every case (10–15% failure), but results, to date, have been dramatic. In a 1970 study conducted in New York City on 2,862 addicts who entered a treatment program, 2,279 remained in the program, and none returned to daily use of heroin. The majority either accepted employment or started school, and previous patterns of antisocial behaviour were either eliminated or significantly reduced. Methadone is a drug of addiction in its own right, but it does not have some of the more serious undesirable consequences associated with heroin.

Synanon

Synanon is a social movement founded in 1958 by Charles Dederich to teach self-reliance and promote personal growth. Dederich is a former alcoholic who brought certain principles that he had found helpful in Alcoholics Anonymous to the treatment of drug addiction. This is a type of moral therapy that has produced encouraging results. As a social climate, Synanon combines certain elements of the autocratic family with the kind of structure one might find in a primitive tribe. The individual is expected to perform tasks as part of the group. His peers who have preceded him in the Synanon experience have a certain status of "elders," and there is a group leader who has the characteristics of the "head" of the household or tribal chief. The central figures in this simulated family structure attempt to teach values and enhance the spiritual aspects of the individual. Emphasis is directed toward the offending behaviour, which is the use of drugs, and there is a rigid requirement that there be total abstinence from addictive substances. The "synanon" per se is a form of group therapy conducted by a recovered drug addict who relies heavily on his own experiences in his work with the newcomer to the group. Initially, he is authoritarian and will employ ridicule and hostile attacks. There is also a technique that places the addict in a position in which he must accept vigorous, rough feedback on his own behaviour. Generally, however, the setting is supportive and permissive. The addict is judged in terms of present behaviour, not past. The basic philosophy is to regard him as a rational person. At no time is addiction viewed as a sickness.

In such countries as the United States, where the addict is treated as a criminal, physicians are prevented from administering opiates for the maintenance of addiction. Acceptable treatment includes enforced institutionalization for about four months, strict regulation against ambulatory care until the person is drug free, and the total prohibition of self-administration of drugs even under a physician's care. Estimates of cures based upon decades of these government-regulated procedures range from 1 to 15 percent.

HALLUCINOGENIC DRUGS

Problems of terminology and definition in hallucinogens

It is difficult to find a suitable generic name for a class of drugs having as many diverse effects as have been reported for "hallucinogens." Abnormal behaviour as profound as the swings in mood, disturbances in thinking, perceptual distortions, delusions, and feelings of strangeness that sometimes occur with these drugs is usually indicative of a major mental disorder; consequently these substances are often called psychotomimetic to indicate that their effects mimic the symptoms of a naturally occurring psychosis. There are indeed points of similarity between the drug states and the natural psychoses, but there are also many dissimilarities—so many as to make

the resemblance quite superficial. Such substances as the bromides, heavy metals, belladonna alkaloids, and intoxicants can, however, cause abnormal behaviour to a degree sometimes described as psychotic, and if the list is extended to include the drugs being discussed here, then the objection—that the term psychotomimetic should refer only to the mimicking of a natural psychosis—is no longer valid. Taking this point of view, some investigators prefer the term psychotogenic ("psychosis causing"). One of the most conspicuous features of this kind of drug experience is the occurrence of the distinctive change in perception called hallucination. For this reason the term hallucinogenic is sometimes used. Most people are aware, however, even while under the influence of the drug, that their unusual perceptions have no basis in reality; so this is not a very accurate use of the term. Strictly speaking, very few people truly hallucinate as a result of taking a hallucinogen.

All these terms are borrowed from medicine and are closely identified with pathology. In this sense, all are negative. It has been suggested that these drugs be called psychedelic ("mind manifesting"). This term shifts the emphasis to that aspect of the drug experience that involves an increased awareness of one's surroundings and also of one's own bodily processes—in brief, an expansion of consciousness. The term also shifts emphasis from the medical or therapeutic aspect to the educational or mystical-religious aspect of drug experience. Only certain people, however, ever have a psychedelic experience in its fullest meaning, and the question of its value to the individual is entirely subjective. The possibility of dangerous consequences, too, may be masked by such a benign term. None of these terms, then, is entirely satisfactory, and one or two are distinctly misleading. (These terms are used interchangeably henceforth with no particular intent other than to indicate membership in the LSD-type family of drugs.)

Types of hallucinogens. The psychoactive substances that have aroused widespread interest and bitter controversy are the LSD-type drugs that produce marked aberrations of behaviour. The most important of these are (1) d-lysergic acid diethylamide, commonly known as LSD-25, which originally was derived from ergot (*Claviceps purpurea*), a fungus on rye and wheat; (2) mescaline, the active principle of the peyote cactus (*Lophophora williamsii*), which grows in the southwestern United States and Mexico; and (3) psilocybin and psilocin, which come from Mexican mushrooms (notably *Psilocybe mexicana* and *Stropharia cubensis*). Bufotenine, originally isolated from the skin of toads, is the alleged hallucinogenic agent contained in banana peels. It has also been isolated in the plant *Piptadenia peregrina*, the mushroom *Amanita muscaria*, and is thought to be the active principle of the hallucinogenic snuff called cohoba and yopo and used by the Indians of Trinidad and by the Otamac Indians of the Orinoco Valley. Harmine is an alkaloid found in the seed coats of a plant (*Peganum harmala*) of the Mediterranean region and the Near East and also in a South American vine (*Banisteriopsis caapi*). There are some amides of lysergic acid contained in the seeds of the morning glory (*Rivea corymbosa*), which the Aztecs call *ololiuqui*. Synthetic compounds of interest are DMT (dimethyltryptamine) and STP (dimethoxyphenylethylamine; DOM). Cannabis (discussed separately below) is not usually included in this group of hallucinogenic drugs, but there is no particular justification for its exclusion. It is a resin obtained from the leaves and tops of the hemp plant (*Cannabis sativa;* see also HALLUCINOGEN).

LSD, mescaline, and psilocybin

History. Native societies of the Western Hemisphere have for 2,000 years utilized various naturally occurring materials such as the "sacred" mushroom of Mexico and the peyote cactus of the southwestern United States. Scientific interest in the hallucinogenic drugs developed slowly. A neurologist wrote about his experience with peyote before the turn of the 20th century, and his account attracted the serious attention of two distinguished psychologists, Havelock Ellis and William James. Mescaline was isolated as the active principle of peyote in 1896, and its structural resemblance to the adrenal hormone epinephrine was recognized by 1919. There followed some interest in model psychoses (that is, drug-induced simulations of abnormal behaviour patterns), but it was not until 1943, when a Swiss chemist accidentally ingested a synthetic preparation of lysergic acid diethylamide and experienced its psychedelic effects, that the search for a natural substance responsible for schizophrenia became widespread. An American mycologist called attention to the powers of the Mexican mushroom in 1953, and the active principle was quickly found to be psilocybin.

Physiological and psychological effects. The psychedelics are capable of producing a wide range of subjective and objective effects; however, there is apparently no reaction that is distinctive for a particular drug. Subjects are unable to distinguish among LSD, mescaline, and psilocybin when they have no prior knowledge of the identity of the drug ingested. These drugs induce a physiological response that is consistent with the type of effect expected of a central-nervous-system stimulant. Usually there is elevation of the systolic blood pressure, dilatation of the pupils, some facilitation of the spinal reflexes, and excitation of the sympathetic nervous system and the brain.

There is considerable difference in the potency of these drugs. A grown man requires about 500 milligrams of mescaline or 20 milligrams of psilocybin or only 0.1 milligram of LSD for full clinical effects when the substances are ingested orally. The active principle in the seeds of the morning glory is about one-tenth as potent as LSD. There are also differences in the time of onset and the duration of effects. Psilocybin acts within 20 to 30 minutes, and the effects last about five to six hours. LSD acts within 30 to 60 minutes, and the effects usually last for eight to ten hours, although occasionally some effects persist for several days. Mescaline requires two to three hours for onset, but the effects last for more than 12 hours. All psychedelics presumably are lethal if taken in large enough quantities, but the effective dose is so low as compared to the lethal dose that death has not been a factor in experimental studies. Physiological tolerance for these drugs develops quite rapidly in man—fastest for LSD, somewhat more slowly and less completely for psilocybin and mescaline. The effects for a particular dose level of LSD are lost within three days of repeated administration, but the original sensitivity is quickly regained if several days are allowed to intervene. Cross-tolerance has been demonstrated for LSD, mescaline, psilocybin, and certain of the lysergic acid derivatives. Tolerance to one of the drugs reduces the effectiveness of an equivalent dose of a second drug, thus suggesting a common mode of action for the group.

Most persons regard the experience with one of these drugs as something totally removed from anything they have ever encountered in normal, everyday experience. The subjective effects vary greatly among individuals and, for a particular person, even from one drug session to the next. The variations seem to reflect such factors as the mood and personality of the subject, the setting in which the drug is administered, the user's expectation of a certain kind of experience, the meaning for the individual of the act of taking the drug, and his interpretation of the motives of the person who is administering the drug to him. Nevertheless, certain invariant reactions seem to stand out. The one most easily described by users is the effect of being "flooded" with visual experience, as much when the eyes are closed as when they are open. Light is greatly intensified; colours are vivid and seem to glow; images are numerous and persistent, yielding a wide range of illusions and hallucinations; details are sharp; perception of space is enhanced; and music may evoke visual impressions, or light may give the impression of sounds. A second important aspect, which people have more difficulty in describing, involves a change in the feelings and the awareness of the self. The sense of personal identity is altered. There may be a fusion of subject and object; legs may seem to shrink or be-

Subjective experiences under the influence of hallucinogens

come extended, and the body to float; space may become boundless and the passage of time very slow; and the person may feel completely empty inside, or he may believe that he is the universe. This type of reaction has been called depersonalization, detachment, or dissociation. Increased suspiciousness of the intentions and motives of others may also become a factor. At times the mood shifts. Descriptions of rapture, ecstasy, and an enhanced sense of beauty are readily elicited; but there can also be a "hellish" terror, gloom, and the feeling of complete isolation. For some people the experience is so disturbing that psychiatric hospitalization is required. Studies of performance on standardized tests show some reduction in reasoning and memory, but the motivation of the subject probably accounts for much of the performance decrement, since many people are uncooperative in this type of structured setting while under the influence of a drug.

Interest in these drugs was routinely scientific for the first few years following the discovery of LSD, but in the 1950s some professional groups began to explore the use of the psychedelics as adjuncts to psychotherapy and also for certain purposes of creativity. It was at this juncture, when the drugs were employed to "change" people, that they became a centre of controversy. LSD is not an approved drug in most countries; consequently its therapeutic applications can only be regarded as experimental. In the 1960s, LSD was proposed as an aid in the treatment of neurosis with special interest in cases recalcitrant to the more conventional psychotherapeutic procedures. LSD also was being given serious trial in the treatment of alcoholism, particularly in Canada, where experimentation is not heavily restricted. LSD has also been employed to reduce the suffering of terminally ill cancer patients. The drug was also under study as an adjunct in the treatment of narcotic addiction, of autistic children, and of the so-called psychopathic personality; and the use of various hallucinogens continue to be advocated in the experimental study of abnormal behaviour because of the degree of control that they offer.

Undesirable or dangerous effects

LSD can be dangerous when used improperly. Swings of mood, time and space distortion, "hallucinations," and impulsive behaviour are complications especially hazardous to the individual when he is alone. Driving while under the influence of one of these drugs is particularly dangerous. Acts of aggression are rare but do occur. In the 1960s, the suicide rate was not high in the various investigational (legal use) groups, but the rate of serious untoward psychological effects requiring psychiatric attention climbed steadily. These drugs do induce psychotic reactions that may last several months or longer. Negative reactions, sometimes called bad trips, are most apt to occur in unstable persons or in other persons taking very large amounts of a drug or taking it under strange conditions or in unfamiliar settings. So far as is known, these drugs are nontoxic, and there are no permanent physical effects associated with their use. There is no physical dependence or withdrawal symptom associated with long-term use, but certain individuals may become psychologically dependent on the drug, become deeply preoccupied with its use, and radically change their life style with continued use.

A new dimension was added to the LSD controversy in 1967, when laboratory studies began to appear in the scientific literature that linked LSD to chromosomal and genetic damage, thus intimating that future generations of the LSD user might be subject to the fearful issue of malformation and genetic illness. Unfortunately, there remains only the poorest understanding of exactly what has been found, to date, in such studies. The findings are neither clear nor conclusive, and moreover they involve not only LSD but also several classes of drugs in rather common use, such as aspirin, caffeine, tranquillizers, and antibiotics. The LSD chromosome story, then, is the story not just of LSD-induced changes but also of possible chromosomal and genetic damage that might be induced by wide classes of drugs in general use. If the gene pool of the LSD user is in jeopardy, as these studies suggest, then, the gene pool of the whole population is also in jeopardy, as these studies also suggest. The danger of drugs is the excessive reliance on drugs; the culprit is everywhere present.

The several types of research upon which these chromosomal and genetic findings are based are wrought with difficulties. Genetic studies that attempt to produce structural malformation use, of necessity, the experimental animal; and there is thus the basic problem of evaluating the extent to which these findings can be generalized to the human. The conditions of the experiments in general do not sufficiently parallel the conditions of natural LSD use to render the data very meaningful. The chromosomal studies present equal difficulties—trying to infer from the behaviour of a cell in the test tube how a mass of cells that are part of a living organism will act.

Types of users. Prior to the mid-1960s, LSD-type drugs were taken by several different types of persons including many who were respected, successful, and well-established socially. Intellectuals, educators, medical and mental health professionals, volunteer research subjects, psychiatric patients, theological students, and participants in special drug-centre communities were some of the first users of these hallucinogenic substances. Beginning in 1966, experimentation in most countries was severely restricted, and subsequent use has been almost entirely of a black market type. LSD use is markedly on the decline, being replaced by cannabis and the amphetamines. Currently, most users tend to be of the middle class—either young, college-educated persons or those who have drifted to the fringe of society. Drug initiation is typically by way of a personal friend or acquaintance. Employers or teachers also have a powerful influence over subordinates and students in terms of drug acceptance. The user of LSD seems often to have an almost fanatic need to proselytize others to drug use. Those who have taken a hallucinogenic substance generally have had experience with other drugs prior to the LSD experience, and there is also a tendency on the part of those who take these drugs to repeat the drug experience and to experiment with other drugs. The special language, method of proselytizing, and psychological dependence surrounding the use of psychedelics bear striking resemblance to the context of narcotics addiction. The chronic LSD user tends to be introverted and passive. Motives for LSD use are many: psychological insight; expansion of consciousness; the desire to become more loving, more creative, open, religious; a desire for new experience, profound personality change, and simple "kicks."

Proselytization among LSD users

BARBITURATES, STIMULANTS, AND TRANQUILLIZERS

There are many sanctioned uses for drugs that exert an effect on the central nervous system. Consequently, there are several classes of nonnarcotic drugs that have come into extensive use as sleeping aids, sedatives, hypnotics, energizers, mood elevators, stimulants, and tranquillizers.

Sedatives and hypnotics differ from general anesthetics only in degree. All are capable of producing central-nervous-system depression, loss of consciousness, and death (see also SEDATIVE-HYPNOTIC DRUGS).

The barbiturates, bromides, chloral hydrate, and paraldehyde are well-known drugs—with the barbiturates being of greatest interest because of the increasing number of middle and upper class individuals who have come to rely on them for immediate relaxation, mild euphoria, and an improved sense of well-being. But alcohol has been and continues to be the drug of choice for these same effects (see also SEDATIVE-HYPNOTIC DRUGS).

Of the drugs that excite the nervous system, nicotine, caffeine, the amphetamines, and the potentially addicting cocaine are well known. The use of stimulants to facilitate attention, sustain wakefulness, and mask fatigue has made the amphetamines an increasingly popular drug by students and those who engage in mental work. Originally the drug of truck drivers, amphetamine is now a common cause of arrest among teen-agers and young adults who commit drug offenses. Cocaine has always been a potentially dangerous drug, but it is not in wide use among the middle and upper classes. Stimulants do not

create energy, and the energy mobilized by these drugs is eventually depleted with serious consequences (see also STIMULANT).

The tranquillizers are a heterogeneous group, as are the behaviours that they are employed to alter. In general, tranquillizing drugs reduce hyperactivity, agitation, and anxiety, which tend to cause a loss of behavioral control. Tranquillizing drugs do not characteristically produce general anesthesia, no matter what the dose; this attribute tends to distinguish tranquillizing drugs from the barbiturates (see also TRANQUILLIZERS).

Dangers of barbiturates, stimulants, and tranquillizers

All the barbiturates, stimulants, and tranquillizers are widely prescribed by physicians, and all these drugs are available through nonmedical (illegal) sources. Most of these drugs are classified as "habit-forming." The minor tranquillizers are commonly associated with habituation and may induce physical dependence and severe withdrawal symptoms. The amphetamines and cocaine intoxicate at high dosages, and both are capable of inducing serious toxic and psychotic reactions under heavy use. The barbiturates are the leading cause of death by suicide. They are judged to be a danger to health by both the World Health Organization Expert Committee and the United Nations Commission on Narcotic Drugs, which have recommended strict control on their production, distribution, and use. The nonnarcotic drugs in widespread use among middle and upper class citizenry manifest considerable untoward consequences for the individual and for society when abused—thus placing their problem in a different perspective than that normally associated with the opiates, LSD, and marijuana.

Barbiturates. The barbiturates relieve tension and anxiety at low dose levels without causing drowsiness, although some tendency toward drowsiness may be an initial reaction for the first few days on the drug. These drugs exert a selective action in small amounts on higher cortical (brain) centres, particularly those centres that are involved in the inhibitory or restraining mechanisms of behaviour. As a consequence, there is an increase in uninhibitedness such as talkativeness and unrestricted social interaction following the taking of the drug. There is also an impairment of function at low dose levels. All the barbiturates are capable of inducing sleep when given in sufficient amounts. They do not affect the perception of pain as do the analgesics, but they do alter the individual's response to pain (*e.g.*, decreasing his anxiety) and are useful in this regard. Infrequently, the barbiturates produce undesirable reactions ranging from simple nervousness, anxiety, nausea, and diarrhea to mental confusion, euphoria, and delirium. Some tolerance is developed to these drugs, but no physical dependence occurs in the drug range (100 to 200 milligrams) normally employed clinically. Prolonged use may lead to drug habituation and psychic dependence. When the drug is used chronically in higher amounts (400 milligrams per day), physical dependence may develop. Sudden withdrawal of a barbiturate following chronic use is frequently associated with withdrawal symptoms that are more severe than those produced by the opiates. A barbiturate should never be withdrawn abruptly following long continued use. The barbiturate addict shows many of the symptoms associated with chronic alcoholism, including blackouts, irrationality, slurred speech, poor motor coordination, emotional deterioration, mood swings, and psychosis.

Cocaine. Cocaine is an alkaloid derived from the leaves of the coca plant (*Erythroxylon coca*), a bush that is natural to Bolivia, Chile, and Peru along the western slopes of the Andes Mountains. Cocaine has a pronounced excitant action on the central nervous system and, in small doses, produces a pleasurable state of well-being associated with relief from fatigue, increased mental alertness, physical strength, and a reduction of hunger. In greater amounts, cocaine is an intoxicant that produces excitement, mental confusion, and convulsions. The Incas were acquainted with the ability of cocaine to produce euphoria, hyperexcitability, and hallucinations; the practice of chewing the coca leaf as part of religious ceremonials was an established custom at the time of the Spanish conquest in the 16th century. The natives who now work the mines high in the Andes chew coca leaves for increased strength and endurance. Coca plants are under cultivation in Sri Lanka (formerly Ceylon), India, and Java. The alkaloid, tropacocaine, is chemically related to cocaine and is obtained from the Java coca plant.

Mild excitation or intoxication

Cocaine is habit forming and may also be physically addicting in some individuals, but not to the extent of the opiates. Only certain persons display abstinence symptoms on withdrawal. Significant physiological tolerance does not develop. Chronic use is associated with severe personality disturbances, inability to sleep, loss of appetite, emaciation, an increased tendency to violence, and antisocial acts. When a toxic psychosis develops, it is characteristically accompanied by paranoid delusions. Hallucinations are prominent with continued use of cocaine, particularly the tactile hallucinations that give the impression that bugs are under the skin. The drug is a white crystalline powder in pure form and the practice of "snuffing" cocaine was common in Europe at the turn of the 20th century. It is less potent when taken by mouth. When injected by vein, a favourite method in the United States, the effects are rapid in onset, intense, but of short duration. This is followed by a correspondingly deep depression that prompts the user to repeat the dose to restore the sense of well-being. Cocaine is sometimes mixed with heroin to dampen any extreme excitability produced by the cocaine. The great number of undesired effects that come on continued use frequently prompts the cocaine user to turn to other drugs.

Amphetamines. These stimulants are of three types having closely related actions on the nervous system: amphetamine proper (Benzedrine), one of its isomers (Dexedrine), and methamphetamine (Methedrine). The amphetamines have been used to alleviate depression, fatigue, the hyperkinetic behaviour disturbances of children, postencephalitic Parkinsonism, enuresis, nausea of pregnancy, and obesity. More recently, the amphetamines have been used in combination with one of the barbiturates, such as amobarbital or phenobarbital, to produce mood elevating effects. It is the effects of the amphetamines on mood that have led to their widespread abuse. A toxic psychosis with hallucinations and paranoid delusions may be produced by a single dose as low as 50 milligrams if no drug tolerance is present. Although the normal lethal dose for adult humans is estimated to be around 900 milligrams, habitual use may increase adult tolerance up to 1,000 milligrams per day.

Mood elevation and induced psychosis

The ability of amphetamine to produce a psychosis having paranoid features was first reported in 1938, shortly after its introduction as a central stimulant. Sporadic reports of psychosis followed, and in 1958, a monograph on the subject of amphetamine psychosis included these statements:

> Psychosis associated with amphetamine usage is much more frequent than would be expected from the reports in the literature. . . . The clinical picture is primarily a paranoid psychosis with ideas of reference, delusions of persecution, auditory and visual hallucinations in a setting of clear consciousness. . . . The mental picture may be indistinguishable from acute or chronic paranoid schizophrenia. . . . Patients with amphetamine psychosis recover within a week unless there is demonstrable cause for continuance of symptoms; *e.g.*, continued excretion of the drug or hysterical prolongation of symptoms.

There have been subsequent attempts to distinguish between amphetamine psychosis and paranoid schizophrenia. Whatever the outcome, amphetamine induces a psychosis that comes closer to mimicking schizophrenia than any of the other drugs of abuse, including LSD. Some behavioral symptoms such as loss of initiative, apathy, and emotional blunting may persist long after the patient stops taking the drug. Methamphetamine was used extensively by the Japanese during World War II, and by 1953 the habitual users of the drug in Japan numbered about 500,000 persons. This large-scale usage created such a serious social problem that the amphetamines were placed under governmental control in Japan in 1954. This Japanese experience has provided the op-

portunity for systematic studies on chronic methamphetamine intoxication. One group of 492 addicts who had been institutionalized showed a 14 percent rate of chronic psychosis with evidence of permanent organic brain damage. In the language of the street, "Meth is death." The amphetamines produce habituation, drug dependency, physiological tolerance, and toxic effects, but no physical addiction.

Tranquillizers. Serendipity has played a major role in the discovery of tranquillizers (as it has in all facets of medicine). Tranquillizers were unknown to medical science until the middle of the 20th century when the therapeutic value of reserpine and chlorpromazine in psychiatry was discovered by chance. Reserpine was originally derived in the 1930s from *Rauwolfia serpentina*, a woody plant that grows in the tropical areas of the world, but it has since been synthesized. Because this drug has many undesirable side effects such as low blood pressure, ulcers, weakness, nightmares, nasal congestion, and depression, however, it has been largely replaced in psychiatric practice by chlorpromazine (Thorazine) and a number of other phenothiazine derivatives synthesized in the 1950s. These phenothiazines are inexpensive, easily available, produce little immediate pleasurable effects, can usually be taken in large amounts without harm, and are not physically addicting. They are used extensively in the treatment of various hyperactive and agitated states, and as antipsychotic agents. These drugs, however, may produce jaundice, dermatitis, or, infrequently, convulsive seizures, and they do not combine well with the drinking of alcohol. Chlorpromazine is effective in reversing "bad trips" such as an LSD-induced panic reaction, but it tends to strengthen rather than reverse the powerful hallucinogenic effects of STP (DOM). There is a second group of drugs, inappropriately termed minor tranquillizers, which have achieved popularity in the management of milder psychiatric conditions, particularly anxiety and tension. The major form is meprobamate (Miltown, Equanil). Although these minor tranquillizers are considered to be entirely safe in terms of side effects, they do produce serious complications, for they are commonly associated with habituation and psychic dependence. Heavy, prolonged use may result in physical dependence and severe withdrawal symptoms including insomnia, tremors, hallucinations, and convulsions.

CANNABIS

Marijuana, hashish, ghanja, and bhang

Cannabis is the general term applied internationally to the Indian hemp plant, *Cannabis sativa*, when the plant is used for its pleasure-giving effects. The plant may grow to a height of 16 feet, but the strains used for drug-producing effects are typically short stemmed and extremely branched. The resinous exudate is the most valued part of the plant because it contains the highest concentration of tetrahydrocannabinol (THC), an active hallucinogenic principle associated with the plant's potency. The term cannabis also encompasses the use of the flowering tops, fruit, seeds, leaves, stems, and bark of the hemp plant even though the potency of these plant parts is considerably less than that of the pure resin itself. Hemp grows freely throughout the temperate zones of the world, but the content of the resin in the plant differs appreciably according to the geographic origin of the plant and the climate of the region in which the plant is grown. A hot, dry, upland climate is considered most favourable in terms of the potency of the plant. Careful cultivation is also considered to be an important factor in resin production. The prevention of pollination and the trimming of top leaves to produce dwarfing enhances the content of resin at plant maturity.

Types of cannabis preparations. Marijuana, hashish, charas, ghanja, bhang, kef, and dagga are names that have been applied to various varieties and preparations of the hemp plant. Hashish, named after the Persian founder of the Assassins of the 11th century (Ḥasan-e Ṣabbāḥ), is the most potent of the cannabis preparations, being about eight times as strong as the marijuana used in the United States. Very few geographic areas are capable of producing a plant rich enough in resins to produce hashish. Unless sifted and powdered, hashish appears in a hardened, brownish form with the degree of darkness indicating strength. The North African either eats it in a confection or smokes it, the water pipe often being used to cool the smoke. The effects are more difficult to regulate when hashish is either ingested as a confection or drunk. In India, this resinous preparation of cannabis is called charas.

Ghanja is a less active form of cannabis. Whereas hashish and charas are made from the pure resin, ghanja is prepared from the flowering tops, stems, leaves, and twigs, which have less resin and thus less potency. Ghanja is nevertheless one of the more potent forms of cannabis. It is prepared from specially cultivated plants in India, and the flowering tops have a relatively generous resinous exudate. Ghanja is consumed much in the manner of charas.

Bhang is the least potent of the cannabis preparations used in India. It does not contain the flowering tops found in ghanja. As a result, bhang contains only a small amount of resin (5 percent). It is either drunk or smoked. When drunk, the leaves are reduced to a fine powder, brewed, and then filtered for use. Bhang is also drunk in Hindu religious ceremonials.

Marijuana is the variety of cannabis grown in the Western Hemisphere. Considered mild in comparison to other forms of cannabis, it is similar in potency to the bhang used in India. Typically, it is smoked, but occasionally it is brewed as a tea or baked into cakes. Marijuana varies considerably in potency, with most American marijuana being grown in Mexico.

Antiquity of cannabis use

History. Cannabis is an ancient plant in terms of use, having been known in central Asia and China as early as 3000 BC and in India and the Near East shortly thereafter. Its introduction to Europe and the Western Hemisphere was probably by way of Africa. Historically, cannabis has been regarded as having medicinal value and it was used as a folk medicine prior to the 1900s. Reportedly, it was considered valuable as an analgesic, topical anesthetic, antispasmodic, antidepressant, appetite stimulant, antiasthmatic, and antibiotic. In the 20th century the pattern of pleasure-giving use spread from the lower classes to the middle classes in the West, particularly among intellectuals. In the 1960s and 1970s it spread throughout various student populations from universities and colleges to secondary schools, finally reaching the elementary schools. This spread to "fad" proportions almost totally obscured the historic use of cannabis as a medicine. Although there are no established therapeutic uses for cannabis at present, cannabis may prove to be of some value in the treatment of depression, loss of appetite, high blood pressure, anxiety, migraine, and various gynecological and menstrual problems.

Physiological and psychological effects. The effects of cannabis are difficult to specify because of the wide variations in the potency of the various preparations of the hemp plant. Hashish or charas would be expected to produce a greater degree of intoxication than marijuana or bhang. It would also make a difference whether the drug is smoked, drunk, eaten, or received as an administration of synthetic tetrahydrocannabinol (THC). In general, hashish produces effects similar to those of mescaline or, in sufficient quantity, to those of LSD—extreme intoxication being more typical when the substance is swallowed. Marijuana, on the other hand, is more apt to produce effects at the opposite or mild end of the continuum from those of LSD. When smoked, physiological manifestations are apparent within minutes. These include dizziness, lightheadedness, disturbances in coordination and movement, a heavy sensation in the arms and legs, dryness of mouth and throat, redness and irritation of the eyes, blurred vision, quickened heartbeat, tightness around the chest, and peculiarities in the sense of hearing such as ringing, buzzing, a feeling of pressure in the ears, or altered sounds. Occasionally drug use is accompanied by nausea and an urge to urinate or defecate. There is also a feeling of hunger that may be associated with a craving

for sweets. Toxic manifestations are rare and include motor restlessness, tremor, ataxia, congestion of the conjunctivae of the eye, abnormal dilation of the pupil, visual hallucinations, and unpleasant delusions. Marijuana is not a drug of addiction. Use does not lead to physical dependence, and there are no withdrawal symptoms when the drug is discontinued. Psychological dependence does occur among certain types of users. Infrequently, a "cannabis psychosis" may occur, but generally this type of psychiatric reaction is associated only with heavy, long-term use of hashish, such as in India and Morocco. Other effects of chronic hashish use are a debilitation of the will and mental deterioration.

Psychological manifestations are even more variable in response to cannabis. Alterations in mood may include giggling, hilarity, and euphoria. Perceptual distortions may also occur, involving space, time, sense of distance, and sense of the organization of one's own body image. Thought processes may also become disorganized, with fragmentation, disturbances of memory, and frequent shifts of attention acting to disrupt the orderly flow of ideas. One may also experience some loss of reality contact in terms of not feeling involved in what one is doing; this may lead to considerable detachment and depersonalization. On the more positive side, there may be an enhancement in the sense of personal worth and increased sociability. Undesired subjective experiences include fear, anxiety, or panic. These effects vary considerably with practice and with the setting in which the drug is taken.

Inconclusive data on effects of cannabis

Many articles have been written on the subject of cannabis, but there is precious little worthwhile data to support any kind of a conclusion with regard to its use. One carefully controlled study on marijuana suggests that it is a very mild substance that requires considerable practice before its full (desired) effects are achieved. Alcohol clearly appears more potent and far more deliterious. If, however, one spends any great amount of time around young people, it does not take long to notice the personal tragedies that are associated with marijuana use. One does not even need to look for them; they are there, and the personal accounts of these young people are very impelling (scientific literature notwithstanding). Perhaps their lives would have been tragedies in any event.

From the point of view of those who favour the legalization of marijuana, the drug is a mild hallucinogen that bears no similarity to the narcotics. They feel that the evidence clearly indicates that marijuana is not a stepping stone to heroin and that its use is not associated with major crimes. As a means of reducing tension and achieving a sense of well-being, they believe that it is probably more beneficial and considerably safer than alcohol. The current hysteria over the use of marijuana and the harsh penalties that are imposed are perceived by users as a greater threat to society than would be a more rational and realistic approach to drug use.

III. Social and ethical issues of drug use

CONFLICTING VALUES IN DRUG USE

Modern industrialized societies are certainly not neutral with regard to the voluntary nonmedical use of psychotropic drugs. Whether one simply takes the position of psychologist Erich Fromm that people are brought up to desire and value the kinds of behaviour required by their economic and social system or whether one goes further and speaks of the Protestant ethic in the sense that Max Weber used it to delineate the industrialist's quest for salvation through worldly work alone, it is simply judged not "right," "good," or "proper" for man to achieve pleasure or salvation chemically. It is accepted that the only legitimate earthly rewards are those that have been "earned" through striving, hard work, personal sacrifice, and an overriding sense of duty to one's country, the existing social order, and family. This orientation is believed to be fairly coincident with the requirements of industrialization as it has been known up to the middle of the 20th century. But the social and economic requirements of modern society may have undergone a radical change in the last few decades, even though the inertia of the existing social character, its desires and its values, will be felt for some time to come. In one major sense, current drug controversies are a reflection of this cultural lag with all of the consequent conflict of wishes and values that result from the lack of good correspondence between traditional teachings and the view of the world as it is now being perceived by large numbers within society. Modern society is in a state of rapid transition, and this transition is not without its untoward consequences in terms of stability.

Cultural transitions notwithstanding, the dominant social order has strong negative feelings about any nonsanctioned use of drugs that contradicts its existing value system. Can society succeed if individuals are allowed unrestrained self-indulgence? Is it right to dwell in one's inner experience and glorify it at the expense of the necessary ordinary daily pursuits? Is it bad to rely on something so much that one cannot exist without it? Is it legitimate to take drugs if one is not sick? Does one have the right to decide for oneself what one needs? Does society have the right to punish someone if he has done no harm to himself or to others? These are difficult questions that do not admit to ready answers. One can guess what the answers would be to the nonsanctioned use of drugs. The traditional ethic dictates harsh responses to conduct that is "self-indulgent" or "abusive of pleasure." But how does one account for the quantities of the drugs being manufactured and consumed today by the general public? It is one thing to talk of the few hundred thousand or so "hard" narcotic users who are principally addicted to the opiates. One might still feel comfortable in disparaging the widespread illicit use of hallucinogenic substances; these are still the "other guys." But the sedatives, tranquillizers, sleeping remedies, stimulants, alcohol, coffee, tea, and tobacco are complications that trap the advocate in some glaring inconsistencies. It may be asked by partisans whether the cosmetic use of stimulants for weight control is any more legitimate than the use of stimulants to "get with it?"; whether the conflict-ridden businessman or the conflict-ridden housewife is any more entitled to relax chemically (alcohol, tranquillizers, sleeping aids, sedatives) than the conflict-ridden adolescent?; whether physical pain is any less bearable than mental pain or anguish? Billions of pills and capsules of a nonnarcotic type are manufactured yearly.

Massive public use of psychotropic drugs

Sedatives and tranquillizers account for somewhere around 12 to 20 percent of all doctor's prescriptions. In addition there are about 150 different sleeping aids that are available for sale without a prescription. The alcoholic beverage industry produces countless millions of gallons of wine and spirits and countless millions of barrels of beer each year. One might conclude that there is a whole drug culture; that the problem is not confined to the young, the poor, the disadvantaged, or even to the criminal; that existing attitudes are at least inconsistent, possibly hypocritical. One always justifies one's own drug use, but one tends to view the other fellow who uses the same drugs as an abuser who is weak and undesirable. It must be recognized that the social consensus in regard to drug use and abuse is limited, conflict ridden, and often glaringly inconsistent. The problem is not one of insufficient facts but one of multiple objectives that at the present moment appear unreconcilable.

YOUTH AND DRUGS

Young people seem to find great solace in the fact that the "establishment" is a drug user. One cannot deny that many countries today are drug-oriented societies, but the implications of drug use are not necessarily the same for the adult as they are for the adolescent. The adult has already acquired some sense of identity and purpose in life; he has come to grips with the problems of love and sex; he has some degree of economic and social skill; and he has been integrated or at least assimilated into some dominant social order. Whereas the adult may turn to drugs and alcohol for many of the same reasons as the adolescent, drug use does not prevent the adult from remaining productive, discharging his obligations, main-

taining his emotional and occupational ties, acknowledging the rights and authority of others, accepting restrictions, and planning for the future. The adolescent, in contrast, is apt to become ethnocentric and egocentric with drug usage. He withdraws within his narrow drug culture and within himself. Drug usage for many adolescents becomes a preposterous "cop-out" at a time when more important developmental experiences are required. To quote one observer:

Escapism in drug use

> It all seemed really quite benign in an earlier time of more moderate drug use, except for the three percent who became crazy and the ten percent we described as socially disabled. Since then, however, more and more disturbed kids have been attracted to the drug world, resulting in more unhappy and dangerous behavior. Increasingly younger kids have come into the scene. Individuals who, in psychoanalytic terms, are simply lesser people, with less structure, less ego, less integration, and hence, are less likely to be able to cope with the drugs. Adolescents are at a crisis period in their lives, and when you intrude regularly at this point with powerful chemicals, the potential to solve these problems of growing up by living them through, working them out, is stopped.

But it would appear that the "establishment" is a drug user, and this has important implications in terms of the expectations, roles, values, and rewards of the social order; but the "establishment" does not "cop-out" on drugs, and this is a fact of fundamental importance in terms of youth. Drugs may be physiologically "safe," but the drug experience can be very nonproductive and costly in terms of the individual's chances of becoming a fully participating adult.

BIBLIOGRAPHY. BERNARD BARBER, *Drugs and Society* (1967), provides an excellent introduction to the general topic of drugs, esp. ch. 1, 5, and 6. RICHARD BLUM *et al.* (eds.), *Utopiates: The Use and Users of LSD-25* (1964), should also be consulted, with particular reference to ch. 1, 11, and 13–14. These two works complement each other to give the reader a rounded picture. For more extensive coverage of this topic, ALFRED R. LINDESMITH, *The Addict and the Law* (1965), offers a broad analysis of the narcotic problem; while DAVID SOLOMON (ed.), *LSD: The Consciousness-Expanding Drug* (1964), provides the reader with some of the history, rationale, subjective accounts, and mystique that launched the drug movement. More specialized papers of general interest are those of F. BARRON, M.E. JARVIK, and S. BUNNELL, "The Hallucinogenic Drugs," *Scient. Am.*, 210:29–37 (1964); and A.T. WEIL, N.E. ZINBERG, and J.M. NELSEN, "Clinical and Psychological Effects of Marihuana in Man," *Science*, 162:1234–42 (1968). RICHARD R. LINGEMAN, *Drugs from A to Z: A Dictionary* (1969), offers an impressive array of information of general interest on almost all aspects of drug use. Technical works covering the same broad scope are J.R. DIPALMA (ed.), *Drill's Pharmacology in Medicine*, 3rd ed. (1965); L.S. GOODMAN and A.Z. GILMAN (eds.), *The Pharmacological Basis of Therapeutics*, 4th ed. (1970); and DANIEL H. EFRON (ed.), *Psychopharmacology: A Review of Progress, 1957–1967* (1968). A summary of psychological and psychiatric views may be found in JAMES C. COLEMAN, *Abnormal Psychology and Modern Life*, 3rd ed. (1964); THEODORE MILLON, *Modern Psychopathology* (1969); and SILVANO ARIETI (ed.), *American Handbook of Psychiatry*, 3 vol. (1959–66).

(W.G.St.)

Druzes

The Druzes, a numerically small Middle Eastern Islāmic sect known for their belligerence, have attracted the attention of European scholars, travellers, and politicians. The Druzes are of particular significance because of the mystery in which they shroud their religious beliefs. They profess *tawḥīd* (pure monotheism) and call themselves *muwaḥḥidūn* (monotheists). The Druzes believe in the divinity of al-Ḥākim bī-Amr Allāh (Ruler by the Command of Allāh), the sixth caliph (996–1021) of the Fāṭimid dynasty of Egypt, whom they call al-Ḥākim bī-Amrih (Ruler by His Own Command).

The initiates and the non-initiated

Nature and significance. Not all Druzes, however, know the secret doctrines of their sect, the Ḥākimīyah. They are divided into two groups: the *ʿuqqāl*—*i.e.*, the sages initiated into the secret teachings of the *ḥikmah*, the Druze religious doctrine—and the *juhhāl*—*i.e.*, those ignorant of the *ḥikmah*. The *ʿuqqāl* are divided into several grades; those who achieve the highest degrees of perfection after long periods of meditation, study, ascetical practices, and seclusion are called the *ajāwīd*, the generous. Only the *ajāwīd* know the innermost secrets of the Ḥākimīyah.

To ask a Druze about his religious beliefs would be considered improper, since the *juhhāl* are ignorant of them, and the *ʿuqqāl* are not allowed to reveal the teachings to the *juhhāl* until they have been initiated into the *ʿuqqāl*. All Druzes, however, united by a religious covenant, have a religious duty to help one another, especially against non-Druzes.

Though Druzes are considered to be very conservative, women hold a respected position in their community and are accepted into the *ʿuqqāl*. Strict monogamy is practiced, but divorce is possible. The divorced, however, cannot remarry their former spouses. Sexual modesty, nonsmoking, and abstinence from drinking wine are strictly observed especially by the *ʿuqqāl*.

The Druzes live in villages that are scattered throughout southern Syria, Lebanon, Israel, and Jordan. These villages often have mixed populations of Muslims and Christians; in former times, Jews also inhabited many of these villages. There are many villages composed entirely of Druzes, however, especially in Syria. Though the exact number of Druzes is not known, estimates place their numbers at approximately 300,000 to 320,000.

Nearly half of them live in the Ḥawrān area in Syria. This area, known as Jabal Druze, has a Druze majority that comprises 88 percent of the population. About 110,000 live in Lebanon, 35,000 in Israel, and 10,000 in Jordan.

Sources. Though the Druzes cloak their religion in secrecy, the source materials for outsiders' knowledge of Druze history and religion are not as scarce as might be expected. There is, of course, information from non-Druze sources, though these sources are not always accurate. Muslim and Christian accounts of historical events, as well as comments and assessments on the Druze religion and customs, should be treated with reserve and caution because of their polemical character. Jewish sources contemporary with the first half of the reign of al-Ḥākim present him in a favourable light.

Even though it would be unacceptable for a Druze sage, a member of the *ʿuqqāl*, to show a non-Druze one of the documents containing the secrets of their beliefs, by the end of the 18th century approximately 120 manuscripts of the Druze religion were to be found in European libraries. The most important and authoritative source of the Druze religion is the *Kitāb al-ḥikmah*, a book containing the fundamental doctrines of the sect. Some other writings are attributed to al-Ḥākim or his earliest disciples. The *at-Taʿlīm* ("Instruction"), a type of catechism written by an anonymous author, presents most aspects of the *ḥikmah* in the form of questions and answers. The author of *at-Taʿlīm* quotes from various sacred texts, such as *Risālat at-taḥdhīr wa at-tanbīh* ("The Epistle of Warning and Reproof," by Ḥamzah), *Risālat an-nisāʾ* ("Epistle to Women"), *Kitāb al-mīthāq* ("Book of the Covenant"), *Risālat ar-riḍā wa at-taslīm* ("Epistle of Recognition," by Ḥamzah), and *as-Sijill al-muʿallaq* (proclamations after the disappearance of al-Ḥākim). The *at-Taʿlīm* is regarded as representative of the teachings of the Druzes.

At-Taʿlīm ("Instruction")

History. *Origins.* The Druzes trace their origins to the times of the Fāṭimid caliph of Egypt, al-Ḥākim, the founder of their religion. The catechism notes that the years AD 1009–21 (or AH 400–411; *i.e.*, 400 to 411 years after the Hijrah, the flight of Muḥammad from Mecca to Medina) were decisive for their development.

In the year AH 400 (AD 1009–10) al-Ḥākim began a program of harassment against Muslims, especially against the Ismāʿīlī sect of the Shīʿah, and also against Christians, whose Church of the Holy Sepulchre in Jerusalem was burned to the ground. After 1012 Jews also were persecuted. Only in the last years of the reign of al-Ḥākim did these persecutions cease. During the years 1017–20

he allowed the rebuilding of the places of worship of other religions. Those persons who had been forced to convert to the religion of al-Ḥākim were allowed to return to their former religion. In the year 1017 al-Ḥākim was publicly proclaimed as the incarnation of God.

Even before 1017 (year one of Ḥamzah, the first year of the Druze era), however, propagation of the tenets of the new religion had begun. Ḥamzah ibn ʿAlī az-Zūzanī (from Zūzan, Persia), regarded as the first disciple of al-Ḥākim, converted the Shīʾite missionary Muḥammad ad-Darazī to the belief in the divinity of al-Ḥākim. From ad-Darazī's name the new religion took its name, Druze. Muḥammad ad-Darazī preached the new religion in the Lebanese plain, an area rife with extreme Shīʿite sects; in 1019, however, he reportedly was killed, having been reproached by his opponents for allowing various vices to be practiced, such as sexual excesses and wine drinking. Allegedly, Ḥamzah was ad-Darazī's opponent. In 1021 al-Ḥākim disappeared. Though he apparently was murdered, his followers believed that he had gone into hiding and would appear a thousand years after he made his first appearance. After ad-Darazī's death, Ḥamzah, with the aid of Ismāʿīl at-Tamīmī, Salāmah as-Samūrī, Muḥammad ibn al-Wahb, and ʿAlī (surnamed Bahāʾ ad-Dīn), developed the doctrines of the new sect. Bahāʾ ad-Dīn played an important role as the author of several sacred works. The last of his writings concludes the formative phase of the *tawḥīd* (1041).

Millennial return of al-Ḥākim

The period following the founding of the Druze religion is obscure. Testimony about their beliefs and practices is to be found in *The Itinerary* (in Hebrew) of Benjamin of Tudela, who travelled in Syria and Lebanon about 1167. During the period of the Crusades, Druzes reportedly inhabited areas near Banyas and Beaufort, and in the 14th century they were known to have inhabited Safed.

The Ottoman period. The Ottoman conquest of the area initiated a new phase in the Druze religion. Sultan Selim, the conqueror of Syria and Egypt (1516–17), confirmed Fakhr ad-Dīn (d. 1544) of the House of Maʿn as emir of the Druzes. Although his son Qurqomāz (1544–85) revolted against the Ottomans, Fakhr ad-Dīn II (1590–1635), son of the rebel, became the emir. He expanded his sphere of influence and made contacts with the Grand Duke of Tuscany. Because the Sublime Porte of the Turks received information about a possible establishment of a crusader state in Syria, Palestine, and Cyprus, he ordered the *wālī* of Damascus to attack Fakhr ad-Dīn. The emir of the Druzes fled to Tuscany, returning to his home in 1618 after a five-year exile in Italy. With aid promised by the Grand Duke of Tuscany, Fakhr ad-Dīn continued his activities against the Porte. In 1635 he was defeated by the Turks at Ḥāṣbayyā and with his two sons was executed in Constantinople.

Houses of Maʿn, Shihāb, Jānbulāṭ, and Aṭrash

The House of Maʿn died out in 1697. The daughter of the last Maʿnī emir married a member of the House of Shihāb. The Shihābs enhanced their position after the defeat of their opponents, the Yaman faction, in the Battle of ʿAyn Dārā in 1711. The Yaman faction migrated to Ḥawrān, which later became known as Jabal Druze. An important emir of the Shihābs, Bashīr II (1788–1840), became a Christian (in name only) and gained the support of the British. He was later deposed by the Sublime Porte, as was his son (officially a Maronite Christian) after a year's reign. Because of antagonisms between the Maronites, who were supported by the French, and the Druzes, who had begun to assert themselves under Bashīr with British support, two *qa'im-maqāms* (subgovernors), one a Maronite and the other a Druze, were appointed. Leadership of the Druzes passed on to the House of Jānbulāṭ, a Kurdish family that came from Aleppo and joined the Druzes rather late.

In 1860 a war broke out between the Druzes and the Christians. The Druzes were supported by the British and the Sublime Porte, the Christians by the French. An estimated 11,000 Christians were massacred in battle and 4,000 died because of destitution. A French landing force restored order. By the organic statute of 1861 (revised in 1864), an autonomous system of government was set up in Lebanon under a Roman Catholic governor general, and the people of the land enjoyed an era of prosperity and relative tranquillity that lasted until 1918.

The modern period. In the Jabal Druze the Aṭrash family rose to prominence and power. The post-World War I division of much of the Middle East into mandated territories allowed the sultan al-Aṭrash possibilities for potential political power. Supported by the French, the spiritual leaders of the Druzes and part of the Aṭrash family proclaimed the independence of the Jabal Druze within the framework of the French mandate over Syria. Sultan al-Aṭrash, however, negotiated with the British. Because the French were unable to control the situation, a revolt of the Druzes in 1925 spread over Syria and part of Lebanon. Damascus and Ḥaṣbayyā were taken by the Druzes, but the revolt failed because of a lack of support from the Lebanese Druzes and the Mutawālī, a Twelver Shīʿah sect. Sultan al-Aṭrash fled to Transjordan, returning only in 1938.

Though Druzes have inhabited Palestine, their stay there has had little historical significance. Some Druze settlements in Upper Galilee were reinforced in the 17th and 18th centuries but were abandoned in part after the defeat of Ibrāhīm Pasha in 1840. Under the Turks and during the British mandate, Druzes were recognized as Muslims and subject to Islāmic religious authorities in matters of personal status and property management. In Israel they are recognized as an independent community with its own autonomous leadership and religious courts.

Beliefs, practices, and institutions. *Doctrines.* The religious beliefs of the Druzes developed out of Ismāʿīlī teachings. Various Jewish, Christian, Gnostic, Neoplatonic, and Iranian elements, however, are combined under a doctrine of strict monotheism.

In Ismāʿīlī theology, the Neoplatonic doctrine of cosmic emanations finds its historical counterpart. With each periodical manifestation, *ʿaql al-kull* (Universal Intelligence) is revealed more completely. Under such a system of successive stages of mystical insight, Ismāʿīlī doctrines were successfully propagated.

Divinity of al-Ḥākim and his emanations

The Druzes, as adherents of the Ḥākimīyah, are regarded as the most extreme of the Ismāʿīlī sects. They believe that al-Ḥākim was the last incarnation of the Deity and that he was preceded by nine previous incarnations called *maqām* (place).

From the light of the Creator (*al-bārī*) is successively emanated Universal Intelligence (*al-ʿaql*), Universal Soul (*an-nafs*), the Word (*al-kalimah*), the Right Wing (*al-janāḥ al-ayman*), and the Left Wing (*al-janāḥ al-aysar*). Called *al-ḥudūd* (boundaries, laws), these five emanations were historically manifested in the five highest ranking disciples of al-Ḥākim; *i.e.*, Ḥamzah, Ismāʿīl at-Tamīmī, Salāma as-Samūrī, Muḥammad ibn al-Wahb, and ʿAlī (Bahāʾ ad-Dīn). Just as there were previous incarnations of the Deity, so also there have been earlier incarnations of the *ḥudūd*. Before Ḥamzah, Universal Intelligence had been incarnated in Shuʿayb (Jethro) in the days of Moses, in the Messiah of True Justice in the days of Jesus, and in Salmān al-Fārisī in the days of Muḥammad. To know the five *ḥudūd* is an obligation in the Druze religion.

Below the five ministers of the Deity, in successively lower ranks, are the emissaries: *dāʿī* (missionary), *maʾdhūn* (authorized agent), and *mukāsir* (way opener) or *naqīb* (one in charge of a district). All of these ranks have the duty of spreading knowledge of *tawḥīd* (monotheism) among the believers. They are to bring this knowledge to women and the *juhhāl* who have reached physical and moral maturity. Because the number of believers has been determined from the time of the Creation, unbelievers are excluded from the true religion.

Transmigration of souls

In light of this teaching, Druzes believe in *tanāsukh* (metempsychosis, or transmigration of souls). They believe that whenever a Druze dies another Druze is born, and the soul of the dying enters the body of the born. Except for the period of al-Ḥākim's appearance, when it was possible for unbelievers to accept the *tawḥīd*, they are not allowed to learn the *ḥikmah* (religious doctrines),

which is regarded as an instrument of salvation. The Jānbulāṭ family, however, joined the Druzes long after the time of al-Ḥākim.

Relationships with other religious groups helped to shape the Druze doctrine of the Last Judgment. In *at-Taʿlīm* there is a detailed description of the Last Times: al-Ḥākim will come again and judge the world twice. In the second and final judgment Christians, Jews, Muslims, and *muwaḥḥidūn* (monotheists) will be judged. The righteous *muwaḥḥidūn* will receive the power and the kingdom, gold, silver, and property, but the Matāwilah, Muslims, and Christians will receive a harsh judgment, and Jews a comparatively lighter punishment.

Practices. The five religious duties (*arkān*) of Islām were abolished by Ḥamzah. Though the Fast of Ramaḍān and the ʿĪd al-Fiṭr were not observed by Druzes, they apparently secretly observed the Muslim ʿĪd al-Aḍḥā (Feast of the Offerings). Ḥamzah instead proclaimed seven fundamental duties for the Druzes: (1) recognition of al-Ḥākim and adherence to *tawḥīd*, (2) negation of all non-Druze tenets, (3) eschewal of the devil, (4) acceptance of God's acts, (5) submission to God, (6) truthfulness, and (7) mutual help and solidarity between fellow Druzes. The order and the number of the duties vary in different texts.

The seven duties

The duty of truthfulness in religious matters applies only to relations with fellow Druzes. Toward non-Druzes strict secrecy (*kitmān*, *taqīyah*) is to be practiced. In times of persecution, however, a Druze is allowed to deny his faith outwardly if his life is in danger. This concession is allowed according to *at-Taʿlīm:* "Our Master commanded us to hide under the wings of the majority religion. When among Christians, we should act like Christians, when among Muslims we should act like Muslims, and so on. For our Master, al-Ḥākim bī-Amrih, said: 'If any community prevails over you, follow its lead [ostensibly], and keep me in your hearts.' "

Worship. The Druze house of worship, the *khilwah*, though not secret, is often difficult to find in a Druze village. The *khilwah* is a rather austere building without architectural embellishments and without furniture, except for small lecterns on which one may lay books during a period of meditation. It is usually located outside the village, and some are in the mountains. Thursday night is the usual time for a Druze meeting, which even the *juhhāl* may attend. When prayer, study, and meditation begin, however, they must leave. Women who belong to the *ʿuqqāl* participate in the worship services.

The highest ranking of the *ʿuqqāl*, the *ajāwīd*, retire for prayer and devotional exercises in the *khilawāt* (plural of *khilwah*). Some of the more important *khilawāt* are located at Qanawāt, the seat of the highest ranking *jawwād* in the Jabal Druze, in Bayyaḍā, the seat of the two spiritual leaders in Lebanon, and at the site of the grave of Nabi Shuʿayb (Jethro) near Tiberias in Israel. At this last site a mass pilgrimage and popular festival are held between April 23 and 25. Druzes from Lebanon and Jabal Druze have attended the festival in the past. Other minor pilgrimages are made to various *khilawāt* that are sites of holy graves.

Conclusions. Just as the whole Middle East population is in transition, so also is the Druze society. During the last generation signs of social changes have been increasing. After World War II the great feudal families were forced to concede political positions to men of other social classes. In Lebanon Majid Arslan was minister of defense (1964) and Kamal Jumblat (Jānbulāṭ) was home minister (1970); many new men, however, came into leading positions. In Israel the revolt of the young progressive forces against the conservative religious leaders was conspicuous. Many of the progressives in Israel have served in the army and are university graduates, having attended coeducational schools and colleges. The continuing changes in Israel and Lebanon will probably affect the Druze societies in Syria and Jabal Druze.

BIBLIOGRAPHY. SILVESTRE DE SACY, *Exposé de la religion des Druzes, tiré des livres religieux de cette secte*, 2 vol. (1838), still the classic work on the Druzes, containing an important list of approximately 120 manuscripts, with many excerpts; NARCISSE BOURON, *Les Druzes* (1930; Eng. trans., *Druze History*, 1952), an account of the 1925–27 revolt; C.H. CHURCHILL, *The Druzes and the Maronites Under the Turkish Rule from 1840 to 1860* (1862), an account by a resident of Lebanon of the massacre of the Christians in 1860; P.K. HITTI, *The Origins of the Druze People and Religion; with Extracts from Their Sacred Writings* (1928), a brief account written for the general reader; H.Z. (J.W.) HIRSCHBERG, "The Druzes," in A.J. ARBERRY (ed.), *Religion in the Middle East*, vol. 2 (1969), a short survey of the history and religious teachings of the Druzes, especially in Palestine, up to 1965.

(H.Z.H.)

Dryden, John

The greatest English poet of the later 17th century, John Dryden was the writer of almost 30 tragedies, comedies, and dramatic operas. He also made a valuable and permanent contribution to English literature in his amiable, civilized, and intelligent commentaries on poetry and drama, which are sufficiently extensive and original to entitle him to be considered, in the words of Dr. Samuel Johnson, as "the father of English criticism." He so dominated the literature of his time that it is frequently referred to by literary historians as "the age of Dryden."

By courtesy of the National Portrait Gallery, London



Dryden, oil painting by Sir Godfrey Kneller (1646–1723). In the National Portrait Gallery, London.

Heritage, youth, and education. The son of a country gentleman, Dryden was born on August 9 (August 19, new style), 1631, in the Northamptonshire village of Aldwinkle. Growing up in the country, he was 11 years old when the Civil War broke out, and most of his early manhood was spent in the protectorate of Oliver Cromwell. When the Civil War compelled Englishmen to stand up and be counted, both his father's and his mother's families sided with the Parliament against the King; but just where Dryden's own sympathies lay at that time there are no means of knowing.

About 1644 he was admitted to Westminster School, where John Locke was one of his contemporaries. Under the celebrated Richard Busby, who did not spare the rod but who seems to have retained the affection and respect of his pupils, the education Dryden received at this famous London school was predominantly classical. His easy and lifelong familiarity with classical literature and his ability to render it into idiomatic English undoubtedly stem from his tutelage under Busby.

In 1650 he entered Trinity College, Cambridge, where he took his B.A. degree in 1654. About the only record of his Cambridge years concerns his being disciplined in 1652 for disobedience to the vice master, and "his contumacy in taking of his punishment inflicted by him." College records, like newspaper reports, are apt to chronicle the less reputable happenings in a man's life: there is little reason to suppose that Dryden was an angry young man and still less that he neglected his studies. Cam-

bridge in the 1650s was full of new ideas; the traditional academic curriculum was being challenged, and, although classical studies were still the basis of the educational system for arts students, some colleges—notably Trinity—were now showing a considerable interest in science. It is significant that in 1663 Dryden, not yet 32 years old, was elected a fellow of the recently founded Royal Society. At Cambridge he could well have acquired his pronounced interest in science and that skeptical and undogmatic cast of mind with which Dryden afterward approached the problems of the day, always excepting political ones.

Fellow of the Royal Society

Early poetic endeavours. What he did between leaving the university in 1654 and the restoration of Charles II is not certainly known. If he was writing poetry, he had almost nothing to show for it, except some ambitious but cluttered verses contributed to a memorial volume on a young Lord Hastings who had died of smallpox (1649) and some complimentary lines the following year to one John Hoddesdon, who had published a volume of epigrams. When in 1659 he contributed a set of "heroic stanzas" to a memorial volume on Oliver Cromwell, he emerged as a poet who might be worth watching. Dryden's poem was mature, considered, sonorous, and sprinkled with those classical and scientific allusions that were to be characteristic of his later verse. This kind of public poetry was always one of the things that he did best.

In May 1660 came the restoration of the monarchy, which was no colonels' revolt, but something that the great majority of Englishmen, weary of the Cromwellian mixture of godliness and ruthlessness, now thoroughly approved. In common with most of his fellow countrymen, Dryden set about adapting himself to the new regime, and he made the necessary adjustments fast and apparently with complete conviction. The poets of the day, great and small, welcomed the young king with loyal effusions, and in June Dryden produced his tribute, *Astraea Redux*, a poem of over 300 lines in rhymed couplets. For the coronation, which took place in April 1661, he was ready with another set of verses, *To His Sacred Majesty*. Those two panegyrical and magniloquent poems were designed to dignify and strengthen the new establishment and to invest the young monarch with an aura of majesty, permanence, and even divinity. Dryden has often been called a timeserver; but his welcome to the King and the new order that came with him seems to have been genuine. At all events, whatever had been holding back his literary development disappeared completely after 1660, and from then on his productivity was remarkable and his touch almost invariably confident and sure.

He had come to know Sir Robert Howard, a son of Thomas Howard, 1st earl of Berkshire, and on December 1, 1663, he was married to the Earl's youngest daughter, Elizabeth. In due course she became the mother of his three sons. In 1666 he was at work on *Annus Mirabilis*, by far his longest poem to date, written in measured quatrains to celebrate two stubborn battles fought out by the English and Dutch fleets in the summer and the Great Fire that destroyed a large part of the City of London in September. The two naval engagements were seen by Dryden as English victories and even the fire was a kind of victory—a triumph not only for the brave Londoners who could "take it" but also for their devoted king ("the father of the people"), who personally directed the fire-fighting operations and whose devout prayers to the Almighty had diverted the wind that was fanning the flames. Dryden was once again gilding the royal image and reinforcing the concept of a loyal nation united under the best of kings.

Appointment as poet laureate

It was hardly surprising that, when Sir William Davenant died in 1668, Dryden was at once appointed poet laureate in his place and historiographer royal two years later. His salary was £200 (increased to £300 in 1677), but the payments were always in arrears, and the king died owing him more than £1,000.

Restoration dramas. One of the first casualties of the Civil War had been the theatres, which were closed in 1642. Soon after Charles II returned to London, he issued patents to Davenant and Thomas Killigrew to form two companies of actors and erect two playhouses, and, before the end of 1660, both theatres had opened. For some time they had to rely on old plays; but Restoration theatregoers, as fashion conscious as any generation before or since, soon began to want new plays that would mirror their own way of life and that were not written in what was by now an antiquated language. In February 1663 Dryden joined the little band of new dramatists with his first play, *The Wild Gallant*, a farcical comedy with some strokes of humour and a good deal of licentious dialogue. It was a comparative failure, but in January 1664 he had some share in the success of *The Indian Queen*, a heroic tragedy in rhymed couplets in which he had collaborated with Sir Robert Howard. Dryden was soon to exploit very successfully this new and popular genre, with its conflicts between love and honour and its lovely heroines before whose charms the blustering heroes sank down in awed submission. In the spring of 1665 Dryden had his own first outstanding success with *The Indian Emperour*, a play that was a sequel to *The Indian Queen*.

Owing to the spread of the Great Plague of London, the theatres were closed by proclamation in June and did not reopen until November 1666. When they did, Dryden had another remarkable hit with a tragicomedy, *Secret Love, or the Maiden Queen*, which appealed particularly to the king. One of the innovations of the Restoration theatre was the substitution of actresses for the boys who had always played women's parts, and in Dryden's new play the part of Florimel, a gay and witty maid of honour, was played to perfection by the king's latest mistress, Nell Gwyn. Samuel Pepys was completely captivated and recorded in his diary: "I never can hope ever to see the like done again, by man or woman." In Florimel's rattling exchanges with Celadon, the Restoration aptitude for witty repartee reached a new level of accomplishment. Another comedy that Pepys laughed at till his head ached was *Sir Martin Mar-all* (1667), based on a translation by the aged William Cavendish, duke of Newcastle, of Molière's *L'Étourdi* (*The Blunderer*, 1762). Dryden revised the Duke's play, and because he ultimately allowed it to be published under his own name, he must have had a considerable share in it, but this rollicking comedy of humours was not in his usual mode. He had now tried his hand at various kinds of drama. That he had also thought a good deal about what he was doing became evident in 1668, when he published *Of Dramatick Poesie, an Essay*, a leisurely discussion between four contemporary writers, of whom Dryden (Neander) is one. This commentary is about dramatic principles and techniques and about the old drama and the new and the rival merits of French and English plays. The dialogue form of the *Essay* was well suited to his open-minded approach to the subject and to the easy colloquial style that came naturally to him.

Commentaries on drama

In May 1668 he entered into an agreement to write exclusively for Killigrew's company and to give them three plays a year. In return, he became a shareholder, entitled to roughly one-tenth of the profits, which in a good year brought him in between £300 and £400. Dryden averaged only one play a year, but, even so, the company must have gained by the contract. In June 1669 he gave them *Tyrannick Love* with its blustering and blaspheming hero Maximin; in December of the next year came the first part of *The Conquest of Granada by the Spaniards*, followed by the second part about a month later. All three plays were highly successful; and in the character Almanzor, the intrepid hero of *The Conquest of Granada*, the theme of love and honour reached its climax. But the vein had now been almost worked out. In December 1671 London playgoers were being highly entertained by the witty burlesque of heroic drama *The Rehearsal*, by George Villiers, 2nd duke of Buckingham, in which Dryden (Mr. Bayes) was the main satirical victim. *The Rehearsal* did not kill the heroic play: as late as November 1675, Dryden staged his last example of the genre, *Aureng-Zebe*, in many ways the sanest and most intelligent of them all. In the prologue, however, he told

his audience that he had grown weary of "his long-lov'd mistress, rhyme," and had discovered that passion was "too fierce to be in fetters bound."

Neoclassical plays. In writing those heroic plays, he had been catering to an audience that was prepared to be stunned into admiration by drums and trumpets, rant and extravagance, stage battles, rich costumes, and exotic scenes. Some years later he admitted that there were passages in those plays that he knew to be "bad enough to please, even when I writ them," but for the future he would not seek any reputation from "the applause of fools." How intelligent he could be when he was not thinking exclusively of "pit, box, and gallery" he had shown by 1672, with his brilliant comedy, *Marriage A-la-Mode*, in which the Restoration battle of the sexes was given a sophisticated and civilized expression that only Sir George Etherege and William Congreve at their best would equal. Equally fine in a different mode was his tragedy *All for Love* (1677), based on Shakespeare's *Antony and Cleopatra* and written in a flowing but controlled blank verse. Dryden had now entered what may be called his Neoclassical period, and if his new tragedy was not without some echoes of the old extravagance, it was admirably constructed, with action developing naturally from situation and character. Of this dignified and at times moving play Dryden said some years later: "I never writ anything for myself but *Antony and Cleopatra*."

For some years Killigrew's mismanaged company had been in grave difficulties, and early in 1678 Dryden appears to have been at loggerheads with his fellow shareholders. When, in collaboration with Nathaniel Lee, he had completed the tragedy of *Oedipus*, he offered it to the rival company and must then or a little later have ceased to be a shareholder, with a consequent serious loss of income. In December 1679 he encountered a different sort of trouble: in an ill-lit London alley he was set upon by three ruffians by whom, in the words of an advertisement he inserted in the *London Gazette*, he was "barbarously assaulted and wounded." It is usually assumed that his assailants had been hired by some noble person whom he had offended, but the affair remains a mystery. It is even possible that he was mistaken for somebody else.

Verse satires. Since the publication of *Annus Mirabilis* 12 years earlier, Dryden had given almost all his time to playwriting. If he had died in 1680, it is as a dramatist that he would be chiefly remembered. Now, in the short space of two years, he was to make his name as the greatest verse satirist that England had so far produced. The immediate stimulus for this outburst of poetical activity was the political crisis that had begun to develop in 1678, when Titus Oates testified before a London magistrate that he possessed evidence of a Jesuit plot to kill the King and place his Catholic brother James on the throne. Led by the able and unscrupulous Anthony Ashley Cooper, 1st earl of Shaftesbury, the Whig Party (then in opposition) turned the so-called Popish Plot to political ends and twice introduced, unsuccessfully, a bill to exclude the Catholic Duke of York from the throne. When a new Parliament assembled in March 1681 at Oxford, the Whigs made a third attempt to bring in their exclusion bill, and, in the House of Lords, Shaftesbury proposed to the King that James Scott, the duke of Monmouth, his illegitimate son but a Protestant, should be declared his successor. The King's answer was to dissolve Parliament. He had played a waiting game with courage and skill, and from now on popular feeling began to turn against the Whigs. On July 2 Shaftesbury was sent to the Tower of London on a charge of high treason.

As poet laureate in those critical months Dryden could not stand aside, and in November he came to the support of the King with his *Absalom and Achitophel*, so drawing upon himself the wrath of the Whigs. Adopting as his framework the Old Testament story of King David (Charles II), his favourite son Absalom (Monmouth), and the false Achitophel (Shaftesbury), who persuaded Absalom to revolt against his father, Dryden gave a satirical version of the events of the past few years as seen from the point of view of the King and his Tory ministers, and yet succeeded in maintaining the heroic tone suitable to the King and to the seriousness of the political situation. As anti-Whig propaganda, ridiculing their leaders in a succession of ludicrous satirical portraits, Dryden's poem is a masterpiece of confident denunciation; as pro-Tory propaganda it is equally remarkable for its serene and persuasive affirmation. But Shaftesbury was to survive a little longer. On November 24, one week after the publication of *Absalom and Achitophel*, a London grand jury refused to endorse the indictment of Shaftesbury (a necessary legal preliminary to his trial), and the triumphant Whigs had a medal struck to commemorate the occasion. Dryden went to work again, and he worked fast: in March 1682 he published *The Medall*, a hard-hitting satire of over 300 lines, prefaced by a vigorous and plain-spoken prose "Epistle to the Whigs." Apart from the ridicule of Shaftesbury, the most striking feature is the unsparing invective directed against those subversive "fanatics" who were trying to overthrow the established order. No poem of Dryden's shows more clearly his innate or acquired conservatism and his dread of mob rule. For a continuation of *Absalom and Achitophel* by Nahum Tate, Dryden also contributed a long passage of brilliant denigration of his fellow dramatists Elkanah Settle (Doeg) and Thomas Shadwell (Og).

Political satire

Shadwell had a whole poem to himself when, in October 1682, anonymously and apparently without Dryden's authority, there appeared in print his famous extended lampoon, *Mac Flecknoe*, written about four years earlier. What triggered this devastating attack on Shadwell has never been satisfactorily explained, but with authors competing for the patronage of the great, the literary world was full of jealousy and bickering. It is true that Dryden was a convinced Tory and Shadwell an ardent Whig, but it was not until later that the two men were to clash on those grounds. In *Mac Flecknoe*, it is as a writer that Shadwell is ridiculed, so ludicrously and with such good-humoured contempt that his reputation has suffered ever since. The year 1682 had been Dryden's own "wonderful year" in poetry, and even now there was more to come. In November he published his *Religio Laici or a Laymans Faith*, which is at once the poet's own confession of his Anglican faith and a lucid statement of the Church of England's "middle way" between an infallible Church of Rome and the diverse individualism of the Protestant sects, with what Dryden saw as their excess of private judgment and their claim to "inspiration."

In February 1685 the death of Charles II (commemorated by his poet laureate in a long pindaric ode, *Threnodia Augustalis*) brought the prolonged crisis of the succession to a head. The accession of James II passed off peacefully, and when, in June, the young Duke of Monmouth appeared at the head of an ill-planned rebellion in the west country, he was easily defeated, captured, and promptly executed. Before long the new king was to give mortal offense to the Church of England by showing himself willing to grant toleration to dissenters and by favouring the small Catholic minority. The worst fears of the Whig politicians were coming true, and in the country at large there was a resurgence of the old anti-Catholic feeling. In this hush before the storm Dryden took a step that he may have been contemplating for some time: toward the end of 1685 the author of *Religio Laici*, who had so admirably defended the Church of England, was received into the Church of Rome. Coming when it did, Dryden's change of faith inevitably exposed him to the charge of timeserving, and the attacks upon him were redoubled; but however ill-timed it may have been, his conversion was almost certainly the outcome of deep and long-sought conviction. Certainly he never recanted afterward, although it might have paid him to do so. When Dryden's thoughts had run clear on any subject it was natural and even necessary for him to put those thoughts in writing. What he now wrote was his longest poem, *The Hind and the Panther* (1687), an elaborate beast fable, in which he argued the case for his adopted church against the Church of England and, of course, the dissenting sects. Like *Religio Laici*, it shows

Conversion to Church of Rome

the delight that a great orator takes in putting a case, and it shows too, in its strange mixture of reverence and ribaldry, the conflicting forces in Dryden's diverse nature.

Late works. With the abdication of James II in 1688 Dryden's prospects became bleak. He had inherited a small property from his father, but in the new reign of William and Mary he had, as a Catholic, to pay double taxes. The worst blow of all was the loss of his laureateship, which went, ironically enough, to Shadwell. Now he had to depend almost entirely on his pen, and it was natural that he should turn again to the theatre, although he was apprehensive of the sort of reception he would get there. His fine tragedy *Don Sebastian* (1689) failed to please, but he was more successful the next year with his comedy *Amphitryon*, which was helped by the attractive music of Henry Purcell. He collaborated more fully with Purcell in a dramatic opera *King Arthur* (1691), which met with considerable success. A tragedy, *Cleomenes*, was for some time refused a license owing to some passages considered politically dangerous, and, when in 1694 his tragicomedy *Love Triumphant* failed completely, he gave up writing for the stage.

The theatre, however, was no longer his chief means of support. Since 1684 he had been supervising a series of poetical miscellanies for the successful publisher Jacob Tonson, and some of his own poems, mainly translations, appeared in them. In 1692 Dryden published *Eleonora*, a long memorial poem to the Countess of Abingdon that her sorrowing husband had commissioned and for which he paid the poet handsomely. For another Tonson volume, a translation of Juvenal and Persius (1693), he contributed all the Persius and five satires of Juvenal. But the great work of these later years was his translation of Virgil, for which Tonson contracted in 1694 and which appeared in the summer of 1697. Taking into account what he got from dedications and what Tonson paid him, it has been estimated that the Virgil, in many respects his magnum opus, brought him in about £1,400. Dryden was now the grand old man of English letters and was often to be seen at Will's Coffee-House discoursing amiably with the wits of a younger generation. His last work for Tonson was a volume of *Fables Ancient and Modern* (1700), consisting of verse translations mainly from the works of Ovid, Chaucer, and Boccaccio, as fresh as anything he ever wrote and introduced to the reader by the most urbane and friendly of all his critical prefaces.

For some time he had been in failing health, but he continued to write almost to the last. For a play that was to be performed for his benefit he wrote a "secular masque," which contains the well-known lines:

'Tis well an old age is out,
And time to begin a new.

Dryden, however, never saw the promised land of the 18th century: he died on May 1, 1700, and was buried in Westminster Abbey, between Geoffrey Chaucer and Abraham Cowley in the Poets' Corner.

His reputation remained high for the next 100 years, and even in the Romantic period the reaction against him was never so great as that against Alexander Pope. In the 20th century there has been a notable revival of interest in his poems, plays, and criticism, and much scholarly work has been done upon them. In the 1970s his contemporary reputation probably stands as high as at any time since his death.

MAJOR WORKS

POEMS: *Astraea Redux* (1660), on the restoration of the monarchy; *Annus Mirabilis: the Year of Wonders 1666* (1667); *Absalom and Achitophel* (1681); *The Medall. A Satyre against Sedition* (1682); *Mac Flecknoe, or a Satyr upon the True-Blew-Protestant Poet* (pirated version, 1682; authorized edition, 1684; written *c.* 1679); *Religio Laici or a Laymans Faith* (1682); *Miscellany Poems*, 4 vol. (1684–94); *Threnodia Augustalis: a Funeral-Pindarique Poem Sacred to the Happy Memory of King Charles II* (1685); "To the pious Memory of . . . Mrs. Anne Killigrew, . . . An Ode" (1686); "A Song for St. Cecilia's Day" (1687); *The Hind and the Panther* (1687); "Alexander's Feast" (1697), second ode in honour of St. Cecilia's Day.

PLAYS: *Secret Love, or the Maiden Queen* (1666); *The Indian Emperour* (1667); *Tyrannick Love, or the Royal Martyr* (1670); *An Evening's Love, or the Mock-Astrologer* (1671); *The Conquest Of Granada by the Spaniards* (1672); *Marriage A-la-Mode* (1673); *Aureng-Zebe* (1676); *All for Love: or, The World well Lost* (1678); *Troilus and Cressida, or Truth Found too Late* (1679); *The Spanish Fryar* (1681); *Albion and Albanius* (1685), an opera; *Don Sebastian, King of Portugal* (1690); *Amphitryon; or, The Two Socia's* (1690); *King Arthur: or, The British Worthy* (1691), an opera; *Cleomenes, The Spartan Heroe* (1692); *Love Triumphant; or, Nature will prevail* (1694).

PROSE: *Of Dramatick Poesie, an Essay* (1668); "Defence of the Epilogue, or an Essay on the Dramatique Poetry of the Last Age" (to his play *The Conquest of Granada*, 1672); "Dedication" to *'Examen Poeticum'* (vol. 3 of *Miscellany Poems*); "A Discourse concerning the Original and Progress of Satire" (to his translations of Juvenal; 1693).

TRANSLATIONS: Of Persius and five satires of Juvenal (1693); of Virgil (complete 1697); selections of Horace, Ovid, Homer, Theocritus, and Lucretius.

ADAPTATIONS: "Creator Spirit, by Whose Aid" (1693, from "Veni, Creator Spiritus"); *Fables Ancient and Modern* (1700, paraphrases from Chaucer, Boccaccio, and Ovid).

BIBLIOGRAPHY. The standard bibliography is that of HUGH MACDONALD, *John Dryden: A Bibliography of Early Editions and of Drydeniana* (1939). It may be supplemented by additions and corrections supplied by JAMES M. OSBORN in *Modern Philology* (1941–42). GUY MONTGOMERY has compiled *A Concordance to the Poetical Works of John Dryden* (1957). Surviving Dryden manuscripts are practically confined to his letters, scattered in various collections. The *Letters* have been edited by CHARLES E. WARD (1942). Aside from the great national libraries, the most useful collection of Dryden's works and related material is that in the William Andrews Clark Memorial Library, University of California, Los Angeles.

Editions: Collected editions of the poems and plays were published by JACOB TONSON in the late 17th and early 18th centuries; and in 1800, EDMOND MALONE edited *The Critical and Miscellaneous Prose Works . . .*, 4 vol., with his long and valuable biographical account in vol. 1. The first edition of the complete *Works* is that of SIR WALTER SCOTT, 18 vol. (1808, 1821), which was rather lightly revised by GEORGE SAINTSBURY (1882–92). These will be superseded when *The Works of John Dryden* (University of California, 1956–) has been completed. The best modern editions of the *Poetical Works* are those of GEORGE R. NOYES, rev. ed. (1950); and JAMES KINSLEY, 4 vol. (1958). The *Dramatic Works*, ed. by MONTAGUE SUMMERS, 6 vol. (1931-32), are usefully annotated but textually unreliable. Dryden's criticism was collected by W.P. KER, *Essays of John Dryden*, 2 vol. (1900); and by GEORGE WATSON, *Of Dramatic Poesy, and Other Critical Essays*, 2 vol. (1962). A useful guidebook, J.M. ADEN, *The Critical Opinions of John Dryden* (1963), gives short extracts arranged alphabetically under topics.

Biography: The standard modern biography is that of CHARLES E. WARD, *The Life of John Dryden* (1961). Information about Dryden's life is meagre, and derives mainly from Thomas Birch, Samuel Johnson, Malone (above), W.D. Christie, and more recently from JAMES M. OSBORN, *John Dryden: Some Biographical Facts and Problems*, rev. ed. (1965). The *Life* in Scott's edition (vol. 1) is still the most readable account, and that in the edition of Noyes (above) is closely packed with information.

Critical studies: Critical studies of Dryden's writings have become numerous in recent years. Of the earlier general studies, Samuel Johnson's critical appraisement in his *Lives of the Poets* (1779), is still basic; and GEORGE SAINTSBURY, *John Dryden* (1881), is at once enthusiastic and discriminating. *Essential Articles for the Study of John Dryden*, ed. by H.T. SWEDENBERG, JR. (1966), offers a selection of modern criticism. For the poetry, MARK VAN DOREN, *The Poetry of John Dryden*, rev. ed. (1960), is a humane and sensitive study. Recent general studies of distinction include A.W. HOFFMAN, *John Dryden's Imagery* (1962); ALAN ROPER, *Dryden's Poetic Kingdoms* (1965); and EARL MINER, *Dryden's Poetry* (1967). An important aspect of Dryden's writing is covered by WILLIAM FROST in *Dryden and the Art of Translation* (1953). General studies of the plays include A.C. KIRSCH, *Dryden's Heroic Drama* (1965); BRUCE KING, *Dryden's Major Plays* (1966); and F.H. MOORE, *The Nobler Pleasure: Dryden's Comedy in Theory and Practice* (1963). The standard study of Dryden's thought was for long that of LOUIS I. BREDVOLD, *The Intellectual Milieu of John Dryden* (1934), but this has been in part corrected and superseded by PHILIP HARTH, *Contexts of Dryden's Thought* (1968).

(J.R.Su.)

Dualism, Religious

Dualism is the doctrine that the world (or reality) consists of two basic, opposed, and irreducible principles or substances (*e.g.*, good and evil; mind and matter) that account for all that exists. It has played an important role in the history of thought and of religion.

NATURE AND SIGNIFICANCE

In religion, dualism has meant the belief in two supreme opposed powers or gods, or sets of divine or demonic beings, that control the world. It may conveniently be contrasted with monism, which sees the world as consisting of one principle such as mind (spirit) or matter; with monotheism; or with various pluralisms and polytheisms, which see a multiplicity of principles or powers at work. As is indicated below, however, the situation is not always clear and simple, a matter of one or two or many, for there are monotheistic, monistic, or polytheistic religions with dualistic aspects.

Types of dualism

Various distinctions may be discerned in the types of dualism in general. In the first place, dualism may be either absolute or relative. In a radical or absolute dualism, the two principles are held to exist from eternity; for example, in the Iranian dualisms, Zoroastrianism and Manichaeism, both the bright and beneficent and the sinister and destructive principles are from eternity.

In a mitigated or relative dualism, one of the two principles may be derived from, or presuppose, the other as a basis; for example, the Bogomils, a medieval heretical Christian group, held that the devil is a fallen angel who came from God and was the creator of the human body, into which he managed by trickery to have God infuse a soul. Here the devil is a subordinate being and not coeternal with God, the absolute eternal being. This, then, is clearly a qualified, not a radical, dualism. Both radical and mitigated types of dualism are found among different groups of the late medieval Cathars, a Christian heretical movement closely related to the Bogomils.

Another and perhaps more important distinction is that between dialectical and eschatological dualism. Dialectical dualism involves an eternal dialectic, or tension, of two opposed principles, such as, in Western culture, the One and the many, or Idea and matter (or space, called by Plato "the receptacle"), and, in Indian culture, *maya* (the illusory world of sense experience and multiplicity) and *ātman-brahman* (the essential identity of mind and ultimate reality). Dialectical dualism ordinarily implies a cyclical, or eternally repetitive, view of history. Eschatological dualism—*i.e.*, a dualism concerned with the ultimate destiny of man and the world, how things will be in the "last" times—on the other hand, conceives of a final resolution of the present dualistic state of things, in which evil will be eliminated at the end of a "linear" history constituted of a series of unrepeatable events, instead of a "cyclical," repetitive one. The ancient Iranian religions, Zoroastrianism and Manichaeism, and Gnosticism—a religiophilosophical movement influential in the Hellenistic world—provide examples of eschatological dualism. A type of thought, such as Platonism, that insists on a profound harmony in the cosmos, is thus more radically dualistic, because of its irreducibly dialectical character (see below) than Zoroastrianism and Manichaeism, with their emphasis on the cosmic struggle between two antithetical principles (good and evil). Midway between these extremes is Gnostic dualism, which has an ontology (or theory of being) of an Orphic-Platonic type (for Orphism, see below *Among ancient civilizations and peoples*) but which also affirms the final disappearance and annihilation of evil with the eventual destruction of the material world—and thus comprises both dialectical and eschatological dualism.

In philosophy, dualism is often identified with the doctrine of transcendence—that there is a separate realm or being "above" and "beyond" the world, as opposed to monism, which holds that the ultimate principle is inside the world (immanent). In the disciplines concerned with the study of religions, however, religious dualism refers not to the distinction or separation of God and the world but to the doctrine of two basic principles; a doctrine that, moreover, may easily be compatible with a form of monism (*e.g.*, Orphism or Vedānta) that makes the opposition between the One and the many absolute and sees in multiplicity merely a fragmentation (or illusory obliteration) of the One.

HISTORICAL VARIETIES OF RELIGIOUS DUALISM

Among ancient civilizations and peoples. Dualism is a phenomenon of major importance in the religions of the ancient world. Those of the Middle and Near East will be considered here.

Egypt and Mesopotamia. While there was generally no explicit dualism in ancient Egyptian religion, there was an implicit dualism in the contrast between the god Seth and the god Osiris. Seth, a violent, aggressive, "foreign," sterile god, connected with disorder, the desert, and loneliness, was opposed to Osiris, the god of fertility and life, active in the waters of the Nile. Seth also possessed some typically dualistic marks of a mythological character; his action, as well as his personality itself, was ambivalent; and, as a typical trickster, he was also capable, at times, of constructive action in the cosmos. The myths of Osiris and Seth may be compared in various ways with those recently discovered among the Dogon peoples of the western Sudan, which contrast Nommo, a fertile and happily mated primordial being pictured in fish form, with Yurugu ("Pale Fox"), an unhappy, sterile character who lives in the wilderness without a mate. Yurugu is considered to be the element that makes the universe complete (the same role assigned to Seth in the Egyptian myth).

Osiris and Seth

Dualism, broadly speaking, was also present in ancient Mesopotamian religion. In myths pertaining to the origin of the gods and of the cosmos, the opposition between the primordial deities (Apsu, the Abyss; and Tiamat, the Sea) and the new ones (particularly Marduk, the demiurge, or creator) displayed some dualistic aspects. Though the earlier deities had established the basic reality of the universe—its ontological core—because of their chaotic and selfish nature they resisted their own offspring, who were later to create the now existing, definite order of the cosmos. A dualism of the ontological —basic reality or being—versus the cosmological—the form or order of the material universe—is thus implicitly affirmed.

Greece and the Hellenistic world. Analogous dualistic concepts may be found in the early Greek *Theogony* of Hesiod (fl. *c.* 800 BC) in his myths of the gods Uranus, Cronus, and Zeus, and the conflict between primordial and later gods. It was in the later, classical Greek world, however, that dualism was most evident. Many of the pre-Socratic philosophers (6th and 5th centuries BC) were dualistic in various ways. In the teachings of Parmenides, for example, noted for reducing the world to a static One —a classical instance of monism—there is still a radical opposition between the realms of Being and Opinion—between ultimate reality and the world of human sense experience. On the other hand, in the doctrines of Heracleitus, noted for reducing the world to fiery Change, the conflict of opposites (hot–cold, day–night, beginning–end, the-way-up–the-way-down), called by Heracleitus *polemos* ("war"), was exalted to become a metaphysical principle. Though these opposites are piecemeal dyads ("pairs"), their effect, taken together, is, as a whole, dualistic. The dualism of Empedocles, simultaneously a religious teacher and a natural philosopher, is especially striking, for he viewed the primordial sphere of the universe as undergoing cycles alternately under the dominance of the antithetical principles of Love and Discord, which periodically break and then reconstruct it. In this context there exist *daimones* ("souls"), divine beings that have fallen from a superior world into this world and exist clothed in the "foreign robe of the flesh." These souls are therefore subject to transmigration through a series of vegetable, animal, and human bodies, owing to a primitive accident (for which credit was given "to the furious Discord").

The same antithetical principles are to be found in Or-

phism, a Greek mystical school, which constituted an independent development within Greek religion and philosophy; beginning in the 6th century BC, it was part of a "mysteriosophic" trend that sought to attain the wisdom of secret mystic and cultic doctrines. Orphism is characterized by its *sōma-sēma*, or body-tomb concept, which saw the body as a prison or tomb in which the soul—a divine element, akin to the gods—is incarcerated. In addition to this psychophysical dualism of soul and body, the Orphic idea that "everything comes from the One and returns to the One" demonstrates a typical dialectical dualism, in which an implicit monism is involved. Developing on an analogous level, Pythagorean numerical and mystical speculation, arising from the 6th-century-BC Greek philosopher and religious teacher Pythagoras, also stressed the dualistic opposition of Monad–Dyad (One–Two) and of other dialectical pairs of opposites.

Platonic dualism

Many of these dualistic ideas, especially the Orphic and Pythagorean ones, are also found in writings of the Greek philosopher Plato (428–348/47 BC), such as the *Timaeus*, *Phaedo*, *Gorgias*, and *Cratylus*. In these writings a divine part of the human soul that is directly infused by the divinity and a mortal part (passionate and vegetative) are defined and considered. The mortal part is assigned to man by inferior divinities, charged to do so by the supreme divinity; and the appetitive passions involved, if followed, are held to be responsible for the punishments that the soul will suffer during various periods of habitation in the other world and reincarnations in this one. Thus God remains free of blame for the destiny of man. The mortal or spoiled part of man is further attributed, in Plato's *Laws*, to the "titanic nature" within his makeup—an element of violence and impiety inherited from the primordial rebellious Titans, sons of the Earth.

Plato's notions of man were rooted in both ontology and cosmology; *i.e.*, in views on being and on the orderly structure of the universe. In the *Timaeus* he considers the cosmos as a single harmony, which for the sake of completeness requires the existence of inferior levels that are bound not only to matter but also to Necessity (the realm of things that could not have been otherwise, and that are hence not amenable to divine activity). A different view is found in his *Laws*, which describes two "Souls" of the World, one of which causes good and one evil. The *Politicus* is concerned with two eternally recurring, alternating cycles in the cosmos, with successive epochs guided either by the gods or by men.

Plato's central inspiration, which unifies his metaphysics, his cosmology, his theory of man, and his doctrine of the soul, was basically dualistic (in the sense of dialectical dualism) with two irreducible principles: the Idea and the *chora* (or material "receptacle") in which the Idea impresses itself. All of this world is conditioned by materiality and necessity; and because of this, the descent of souls into bodies is said to be rendered necessary as well.

Neoplatonism, a 3rd-century-AD development from Plato's thought, conceived the cosmos as a harmony with a succession of levels emanating from an ultimate unit. There was in the system, nevertheless, a rupture of the harmony of the cosmos called *tolma* ("the audacity"), which served as an explanation for the descent of Soul into the material world—and thus constituted a dualistic element.

Gnosticism

In Gnosticism, a Hellenistic religious movement that entered original Christianity from earlier pagan sources, and which viewed matter as evil and spirit as good, dualism manifested itself in a more dramatic way. Gnostic dualism cannot be understood without reference to both Judaism and Christianity, and perhaps even to Zoroastrianism, since Gnostic eschatological characteristics were derived from them. Gnosticism was also connected with certain principles of Orphism and Platonism; reflecting the Orphic body-tomb doctrine, for example, Gnosticism adopted a firmly antisomatic stance (against the body), and similarly adopted the concept of the divine soul—the pneumatic, or spiritual, soul, as the Gnostic would say, of the same substance as the divinity—that is destined to free itself from the tyranny of a material, cosmic demiurge (or subordinate deity). Certain Gnostics, moreover, developed a radical anticosmism, in which they registered their animosity against the material universe by cursing the stars—which brought them bitter reproach from Plotinus (*c.* AD 205–269/270), the founder of Neoplatonism. As viewed by the Gnostic Ophite sect, which venerated the *ophis* (or "snake") as a symbol of knowledge, the cosmos comprises three parts: the superior world, the inferior world (material and chaotic), and the intermediate world, or *logos* ("word" or "reason")—the *logos* being depicted as a snake that impresses spiritual forms into the chaotic matter. These forms—life, soul, vital masculine substance—are later freed again, a liberation that completely empties the material world. Such Gnostic views are of two types: Iranian and Syrian-Egyptian. Iranian Gnosticism is characterized by an absolute, radical dualism: light and darkness, *pneuma* ("spirit") and chaotic formless matter, oppose each other from eternity. Syrian-Egyptian Gnosticism is characterized by a dualism that is mitigated (as earlier defined) but also drastic: the inferior world, the chaotic darkness, begins to exist only at a special moment owing to an accident in the divine world; and this accident is usually also identified with an "audacity," a defect in one of the "aeons," or divine entities.

Iran. In the Indo-Iranian period (2nd millennium BC) there were already tendencies toward dualistic thought, especially in myths relating to monstrous and demonic beings who still the movement of the waters and thus make cosmic life impossible; in later-archaic Indian speculation there was also a tendency to oppose *devas* ("gods") to *asuras* ("demons"). Iranian dualism, however, expressed itself most characteristically in Zoroastrianism. In the Zoroastrian religious texts, the *Gāthās*, there is an opposition between two spirits, the Beneficent Spirit (Spenta Mainyu) and the Destructive Spirit (Angra Mainyu, or Ahriman). These two spirits are different, irreducible principles; at the beginning they have chosen life and nonlife, respectively. Though the Beneficent Spirit is almost an hypostasis (the substance) of the divinity (Ahura Mazdā), nothing is said in the *Gāthās* about the origin of the Destructive Spirit. In any case, the very fact that the Destructive Spirit is said to be "twin brother" of the Beneficent One does not imply that he is a son of Ahura Mazdā but only that the two spirits are "symmetrical"; *i.e.*, equal and contrary as to their respective efficacity and orientation.

Ahura Mazdā versus Ahriman

Medieval Zoroastrian treatises present radical and eschatological dualisms in their extreme forms. According to the *Bundahishn* ("Primordial Creation") text, Ormazd (Ahura Mazdā) and Ahriman have always existed. Ormazd is represented as lofty, in the light, full of omniscience and goodness, while Ahriman is represented as debased, in darkness, full of aggressiveness and ignorance. Ormazd's omniscience allows him to conceive and to actualize the Creation and Time, because only these can offer him an arena in which to accost Ahriman and eliminate him.

The medieval Zoroastrian treatises also describe another "dual" formulation, the two realms of creation and of reality: the *mēnōk* ("potential, embryonic, initial, heavenly, and invisible") and the *gētīk* ("realized, final, worldly, concrete, and visible"). But this opposition does not imply a devaluation of the *gētīk*, of this world.

Zurvanism, a Zoroastrian heretical movement (*c.* 3rd/4th century BC–7th century AD), was also dualistic. The very names of Zurvān (Time-Destiny) and the partially synonymous *zamān* ("time") already appear in the later Avesta and in medieval treatises, in which Time is the milieu in which Ormazd and Ahriman fight. Also, a myth attributed to Zoroastrian priests by later, non-Iranian sources speaks of Zurvān as the father of Ormazd and Ahriman. At times "Zurvanite" mythology tends toward formulations of a Gnostic and Manichaean type (women paid allegiance, for example, to Ahriman, who has partial authority in the world). Zurvanism also developed theosophic characteristics (involving mystical insights), such as that which discerned the ambivalence of

Zurvān—viz., that although an evil element (an evil thought or spiritual corruption) has always existed within him, he nonetheless, so it seems, eliminates the evil by expressing it and is thus worthy to be identified with the supreme divinity (Yazdān).

Among religions of the East. Dualisms have also appeared in various forms in the religions of India and China.

India. Indian dualism has involved the opposition of the One and the many: of reality and appearance. In an ancient Hindu hymn (Ṛgveda, 10.90.), Puruṣa, "the Immortal that is in heaven," is opposed to this world; the three quarters of the Puruṣa that comprise the transcendent world are opposed to the other quarter of him (his limbs) that is this world; *i.e.*, the divine foundation, the divine substance of this world, is made out of his limbs. Early speculation on the identity of the *ātman* ("self") and Brahman ("cosmos"), as opposed to the material and visible world that is subject to *māyā* (or "mundane illusion"), has been mentioned above.

The Sāṃkhya school of Indian philosophy presents another, probably later, formulation of dualism based on two eternal and opposed cosmic principles: *prakṛti* ("original matter") and *puruṣa* ("spirit"), the name of the ancient primordial Man, substance of the universe. Matter is differentiated into three different *guṇas* (or "qualities") that articulate the three levels of the being and essential nature of man in hierarchical connection with each other. Spirit, in itself free, eternal, and infinite, becomes involved in matter by the development of the latter. Salvation coincides with the knowledge of the state of things: "I (spirit) am one thing and It (matter) is another."

China. The first words of the Taoist text, the *Tao-te Ching*, express a doctrine that is typical of a pervasive Chinese dualism; *i.e.*, that of the two opposed and complementary principles, the Yin and the Yang (respectively, feminine and masculine, lunar and solar, terrestrial and celestial, passive and active, dark and bright; in short, the entire series of opposites). The dialectics of Yin and Yang are the double manifestation of the one and only eternal, undividable, and transcendent principle: Tao ("the Way").

Yin and Yang

African philosophies and religions. A typically dual composition (involving the coexistence and cooperation of two elements), or even a dualistic opposition (as two opposed elements that function as principles in respect to the actual creation), is found in the Dogon (western Sudanese) notions about Nommo and Yurugu, already mentioned. A series of "words" refers to both principles; *i.e.*, a series of realities and categories can be named that constitute the world in its functional variety, which transcend the simple good–evil opposition, and according to which both Nommo and Yurugu are dualistic "principles" essential to the actual dynamics of the world.

Among religions of the West. Dualisms have appeared in Western religions chiefly under the impact of Gnostic influences.

Judaism. No real dualism is found in Judaism, except in the Gnostic and theosophic forms of Jewish mysticism known as Kabbala. The presence of a vigorous and universal monotheism implies not only faith in a single creative god but also faith in a god who is the uncontested master of history; and neither Satan nor Belial detract from this absolute monotheism. Within these limitations, however, a tendency towards dualistic thought exists in such late noncanonical texts as the *First Book of Enoch* (*c.* 1st century BC), in which certain angels are said to have fallen as a consequence of their wedding with the daughters of men. These angels, it is held, taught mankind the malevolent arts of magic, seduction, and violence, together with such elements of culture as the use of metals and writing. Though there is no dualism in the stricter sense in the *Manual of Discipline*, one of the Qumrān texts of the Dead Sea Scrolls, a certain polarity is nonetheless displayed in a passage that asserts of God that

> he created man to have dominion over the world and made for him two spirits, so that he may walk by them until the time of his visitation: they are the spirits of truth and error. In the dwelling of light are the origins of the truth, and from a spring of darkness are the origins of error. In the hand of the Prince of Lights is dominion over all the children of righteousness, in the ways of light they walk. And in the hand of the angel of darkness is all dominion over the children of error; and in the ways of darkness they walk.

The context of this passage, however, is completely monotheistic. It expresses a doctrine also found in the *Didachē*, a Jewish-Christian work of the early 2nd century AD (better known as the *Teachings of the Twelve Apostles*), that of the two roads on which a man may walk, the good road and the bad, the road of life and that of death, with God leaving the choice of the road to man's free will; and also the later rabbinic doctrine of the struggle between the good and evil inclinations (*yetzer*) within man. There is also no hint of dualism in the two "sources" mentioned in the Qumrān texts, the bright source and the dark. These are hardly dualistic principles (in the ontological sense of the term) but are simply radical (*i.e.*, original) polarities in spiritual orientation. (Not even the "Angel of Darkness," mentioned in the same context is a principle, though he is a person and a power.)

Non-dualistic polarities in late Judaism and Jewish Christianity

There is thus no possible hostile confrontation of two principles such as appears in Iranian Zurvanism. Elements of dualistic thought (in a Platonic sense) are also found in the works of the Jewish Hellenistic philosopher Philo of Alexandria (1st century AD), whose philosophy was dualistic in its doctrines about the universe and man, but without shaking his basic adherence to biblical monotheism.

Christianity. In Christianity dualistic concepts appeared principally in its Gnostic developments. But even in the 2nd-century Judaizing sect of the Encratites, which was not really Gnostic, there were dualistic aspects that had modified some tendencies in later Judaism. These teachings were also particularly prominent in the writings of the supporters of Docetism (the doctrine that Christ, being divine, did not suffer and die; 2nd century), who held that matter is essentially evil and that the soul is a pre-existent substance. According to the Encratites, the pre-existent soul, once it "gets effeminated by concupiscence," drops into the carnal world. Since generation perpetuates the soul's state of decay in this bodily world, they condemned all sexual relations. The dualism of Marcion (a 2nd-century semi-Gnostic Christian heretic) was really a ditheism (a system positing two gods), though the common Gnostic presuppositions—such as antisomatism and anticosmism, the condemnation of the body and the material universe—were also present in his thought. For Marcion, the God of the Old Testament is an inferior and harsh creator demiurge, author of the world and man, who is nonetheless completely distinct from the supreme divinity, who manifested himself in Jesus and is a stranger to this world. For Saturninus (or Satornil) of Antioch, the founder of a 2nd-century Syrian Gnostic group that was commonly connected with the tradition of Simon Magus (reputed leader of an earlier Gnostic sect), the God of the Old Testament is only one of the angels, the martial angel of the Judaic nation, although (as with Marcion) he is distinct from the devil, who is in fact his opponent. According to Saturninus a primordial accident caused a wave of *pneuma* ("spirit") to land in the inferior darkness, where it is said to have remained prisoner and now continues its existence in those who, characterized by the presence in them of this superior element, will later be conducted back to their heavenly origin by Jesus, a messenger coming from above. Conceptions of a similar type are also found in the "Psalm (or Hymn) of the Naassenes" (Naassene is the Hebrew term for Ophite, mentioned above) and in the "Song of the Pearl" in the Gnostic *Acts of Thomas;* here also occurs the concept of a "saviour to be saved," who has been sent from above and was made a prisoner by darkness. This basic concept was developed fully only in Manichaeism. The Gnostic-dualist view survived in late antiquity and into the Middle Ages, both in the East, among the Mandeans. Yezidis ("devil-worshippers"), and

Encratism

some extreme sects within the Shīʿah branch of Islām and in the West among the Bogomils and Cathars. It is still present today in modern theosophy.

Among modern nonliterate religions. Religious dualism also manifests itself among primitive peoples, especially in the concept of a "second" figure, an ambivalent demiurge-trickster who can be both a collaborator and rival of the supreme being and independent of the latter in origin. Such tricksters include the Coyote (in North American Indian mythology), the Raven (among Paleosiberians), or the Crow (among the Southeast Australian tribes). To these animal figures are attributed the origin of such negative aspects of life as death and illness. But they are also credited as benefactors; *e.g.*, in creating utilities in the cosmos and in the invention of fire. The demiurge-trickster is typically ambivalent, tremendously frightful and efficacious, but also frequently limited in power. For example, such tricksters are often incapable of animating the beings that they have molded and must therefore request the help of the supreme being in bringing them to life. They are said to be selfish, lonely, and unhappy and this moves them, despite their arrogance, to attempt to relate themselves to or unite with the supreme being.

Role of the trickster

Other dualistic concepts among primitive peoples posit opposite the supreme being a violent and death-bearing "second" figure of a demiurgical type. The character of Erlik in the mythologies of the Central Asiatic Turks (*e.g.*, among the Altaics) is typical.

Erlik is a king of the dead and master of death who assumes the role of a fraudulent and unfortunate collaborator with the supreme being. In stories about the origin of the universe, he appears as an aquatic bird in charge (under the supreme being) of fishing a little earth from the bottom of the primordial sea—a theme also well-known in East European folklore. In other myths, a similar being spits on human beings at the time they are created by God or breathes his bad spirit into man or woman. Elsewhere there is depicted an opposition of two twin brothers, of whom one is the demiurge-creator of good things and the other of death; both, however, are the sons of a mother goddess of heavenly origin. This pattern is exemplified in the Iroquoian myth of Yoskeha and Tawiskaron—a myth curiously reminiscent of certain aspects of the Iranian Zurvanite mythology.

Other ethnological polarities, or pairs of opposites (eastern–western, celestial–terrestrial, solar–lunar divinities, right–left, full moon–dark moon, etc.) are dualistic in the sense of contrasting principles or creating agencies.

THEMES OF RELIGIOUS DUALISM

The sacred and the profane. Among the various themes of religious dualism the opposition between sacred and profane is also important. This distinction, appearing in some sense in nearly every religion, must be particularly acute, however, to qualify a religion as dualistic. Such an intensification of the sacred–profane opposition to the point at which it becomes a dualism is evident in the mid-20th century historian of religions Mircea Eliade's conception of religion (see also SACRED OR HOLY). This contrasts time (the *illud tempus*, "those times," of the intact, sacred, primordial creation that are periodically restored by ritual) and the historical time (marked by decay, profaneness, and loss of plentitude and significance).

Good and evil. More pertinent (even if not always dualistic) is the opposition between good and evil, in the various meanings of these words. Whenever the problem of the origin of evil is solved by conceiving the real existence of another principle separate from the prime principle of the world, or by affirming an inner ambivalence, limited sovereignty, or inadequacy of the prime principle, or of divine beings, a dualism then emerges; and through this good–evil opposition, the problems of theodicy (*i.e.*, of the doctrine of the justification of divine action in a world in which evil is present) are posed. If evil either is, or comes from, a self-existent principle antithetical to the principle of good, then this provides the divinity with a "justification." Such views are completely different from the justification of God in nondualistic religions, especially the monotheistic ones. In monotheistic religions evil does not originate within the divinity nor in general within a divine world (*plērōma*) as it does in Gnosticism; it arises instead from the improper use of freedom by created beings. In monistic religions—all of which are based on the opposition between the One and the many, seen either as an illusion or as the decay or fragmentation of the One—along with a strong ascetic emphasis, there is a notion of evil as being for man a painful and fatal essence that issues from a metaphysical cause or an ontologically negative principle. For the same reason, it is necessary to distinguish between the nondualistic concept of "original sin" in Christian theology and the concept of "previous sin"—in monistic religions with a dualist aspect; whereas "original sin" arises and spreads within the human sphere, "previous sin" is consummated in some sort of a "prologue in heaven" and generates the very existence of the world and of humanity itself.

The problem of evil

Creation and destruction: life and death. Another important dualistic theme is that which opposes life to death based on two opposing metaphysical principles. A typical example of this dualistic opposition is found in Zoroastrianism. Zoroastrian doctrine is strongly vitalistic: Ahriman's chief acolytes are Aēshma (the fury), the Druj Nasu (the deadly agent of putrefaction), Jēh (the infertile whore), and Apaoša (the demon of sterility)—death-bearing forces. There is also a strong vitalistic formulation of these principles in Gnostic doctrines, especially in the Ophite and Barbelo-Gnostic (worshipping Barbelo as the Great Mother of life) varieties, which identify the *pneuma* and the light with the vital substance. At other times the opposition of life and death is formulated in a dialectical manner as a recurring alternation of the two principles. The complex Egyptian opposition between Osiris, the "dead god," who is nonetheless the principle of fecundity and life, and his counterpart Seth has already been mentioned (see above *Egypt and Mesopotamia*). The same dialectic is typical of the "fecundity cults," in which a god-genius of vegetation, a "dying god," is featured, who undergoes a seasonal disappearance and return (not to be interpreted as a "resurrection"). To such vegetation gods, death- or decay-producing figures are sometimes opposed—as Mot (the Death) opposed to Baal, and an infernal and lethal wild boar opposed to Adonis, and (in German religion and mythology) Loki opposed to Baldr. These figures, the agents for disastrous occurrences, were already implicit in the figure of the dying god himself and in his relation to the seasonal cycle of vegetation. To be sure, the growing season is limited; and the new arrival of vegetation each spring (and the wedding of the fertility god) is terminated in the fall by the god's departure to the netherworld (with appropriate lamentation). But the rise of vegetation, though ephemeral, is nonetheless basically benevolent. This complexity is also manifest in those agricultural religions that present themselves as mystery cults (*e.g.*, the Eleusinian mysteries), bestowing upon the initiate a hope for life after death.

The dying god and vegetation cults

But the dualistic theme is far more evident in "mysteriosophy"; *i.e.*, in the "sophic," or "wise," reinterpretation of mysteries (*e.g.*, Orphism). In this context, the divine soul replaces the dying god in the soul's descent from a superior world into the corporeal world—a concept that was later bequeathed to Gnosticism and is especially apparent in its transposed basic vitalism.

A dialectical formulation of the opposition of life and death is also found in the basic theology of Hinduism: with Viṣṇu (Vishnu) cast as the principle of creation (called Nārāyaṇa) and the sustenance of life and Śiva (Shiva) as the principle of destruction and death. The ambivalence of life–death is also found in a series of Hindu divinities (*e.g.*, Śiva, Kālī) and cults whose death-inflicting characteristics are justified in a paradoxical celebration of the recurring triumph of life.

Polytheistic themes. Among the instances of dualistic structure in polytheistic religions are those that oppose celestial and terrestrial, male and female, actual and mythical primordial-chaotic, "diurnal" and "nocturnal," especially when they do so within the context of mythol-

ogies and cosmogonies belonging to the ancient world's polytheistic "high cultures" (see above *Egypt and Mesopotamia; Greece and the Hellenistic World*). Such pairs of opposites often provide a framework for polytheistic pantheons that would otherwise appear anarchic or less than comprehensive (see also POLYTHEISM).

FUNCTIONS OF RELIGIOUS DUALISM

Cosmological and cosmogonic functions. The essential function of any religious dualism is obviously ontological—to account for a duality of opposed principles in being—even when the two principles are not regarded as coeternal; and this underlies the cosmological–cosmogonic, anthropological, and sociological functions and expressions of dualism. Both dialectical dualism (*e.g.*, in the fertility cults, Orphic mysteriosophy, and Platonism) and eschatological dualism (*e.g.*, in the Zoroastrian and Manichaean notion of the "mixture" between the two creations good and bad) have a basically cosmological function—the explanation of the structure of the universe. Whenever the concept of a distinct creator, transcendent with respect to his work, is missing (as, for example, in monistic formulations of the Indian type or in polytheistic milieus), dualism has a cosmogonic function—the explanation of the origin of the universe.

On a cosmogonic level, dualistic opposition may also be manifest in the celestial world; *e.g.*, in the late Zoroastrian opposition between the beneficent fixed stars and the planets (which are negative, because they are alleged to proceed in the reverse sense); or else between the world of the Heptad (again the seven planets, under the dominion of the tyrannic archons, or rulers, that cause human passions) and the superior heaven of the Ogdoad (the group of eight divine beings or aeons), as in Gnosticism and Mithraism. In this form of dualism the monstrous figure of Leontocephalos (a human figure with a lion's head, belted by a snake with astral signs) represents the power of astral Destiny-Time to be transcended by the soul—a power that is a basic presupposition of astrology and magic. On the other hand, the heaven–earth opposition cannot be regarded as dualistic if the two elements are represented merely as cosmic progenitors (see also CREATION, MYTHS AND DOCTRINES OF).

Anthropological functions. The anthropological functions of dualism (dealing with the nature and destiny of man) are present in all those doctrines that consider man as a duality, or, rather, as an irreconcilable duality of opposed elements. Of particular importance is the opposition between masculine and feminine, in which their opposition involves a remarkable difference in level of being. In mythologies (whether dualistic or not) with a "second" figure, a demiurge, there is frequently a connection between the demiurge and the origin of women (*e.g.*, the myths of Prometheus-Epimetheus in ancient Greece, and of Paliyan in southeast Australia) or between the demiurge and the origin of sexuality (*e.g.*, the myths of the trickster Coyote and of the Gnostic demiurge). In the Platonic theory of man the first incarnation of the soul occurs in a masculine body, and only a subsequent incarnation, marking a later descent of the soul into the world of bodies, is feminine. In Gnosticism (Ophite sects) the vital substance that animates the universe is masculine (active), while the quality of the material world is feminine (passive); and in the last *logion* ("saying") of the Gnostic *Gospel of Thomas*, it is said that Mary will be saved by being made a male; *i.e.*, she will become a "living spirit" (*pneuma*). Gnostic and Manichaean antifeminism, as well as Encratite (and perhaps Orphic) antifeminism, are motivated by their hatred for procreation, which they believe implies the fall of the soul into the material world and its permanent abode there. At other times procreation is explained in terms of a division of a complete, originally androgynous (both male and female) being (as in Plato's *Symposium* and in the Gnostic *Gospel of Philip*). There are other nondualistic doctrines in which woman is considered to be connected in some way with the origins of evil but not as the embodiment of the evil principle (*e.g.*, in Genesis and the apocryphal late-Judaic *Book of Adam*).

Masculine and feminine

Sociological functions. The sociological functions of religious dualism are less relevant. Among some Australian peoples the "totems" of the two classes of a tribe that intermarry are the Falcon-Eagle (Bundjil), the supreme being, and the Crow (Waang), a demiurge-trickster. According to the Menominee Indians, the highest region of the universe is inhabited by benevolent gods (among whom the supreme being is Mate Häwätûk) and the inferior region by bad ones; and these two groups are constantly fighting. The Menominee believe that they come from an alliance of families that once belonged to these two groups, whose respective descendants have particular places in the assembly and clearly differentiated functions.

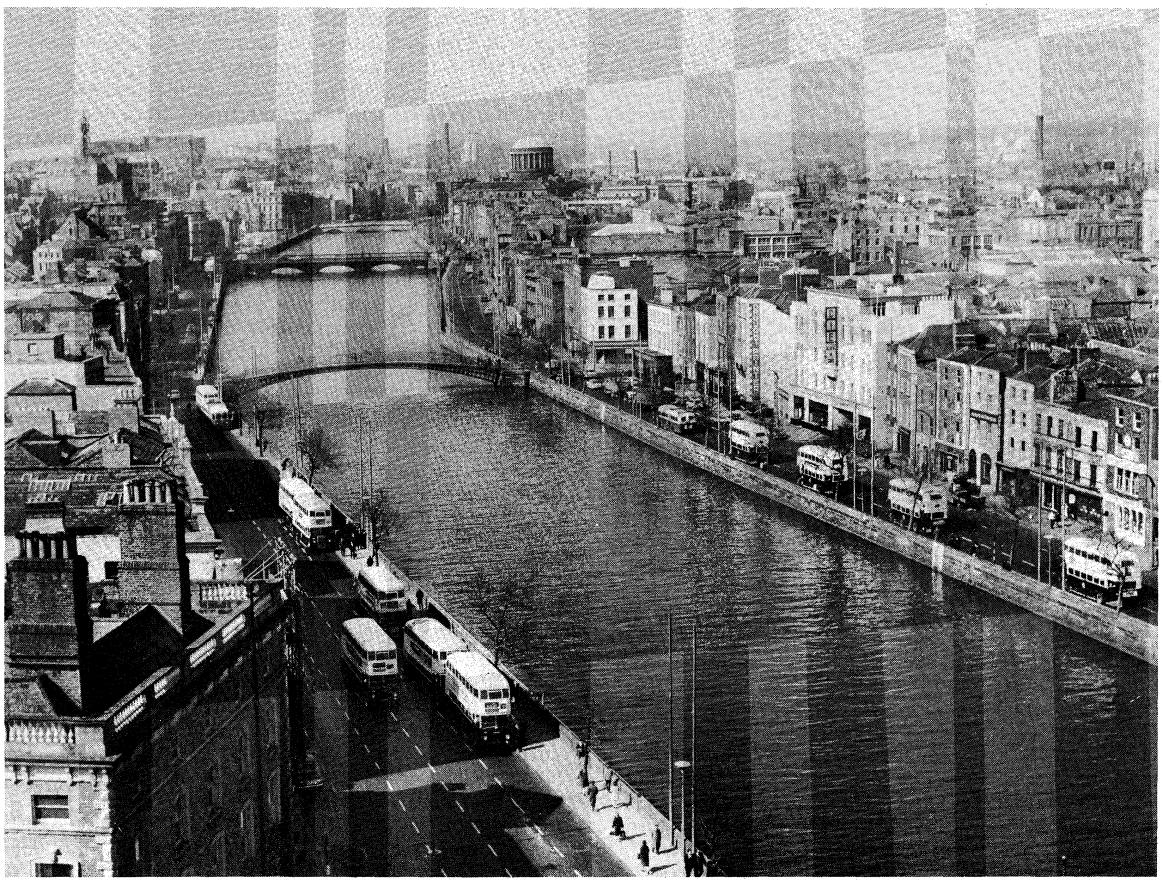
Sociological and economic class oppositions, however, cannot provide a general explication for dualism. All dualities (*e.g.*, in the social structure) are necessarily relevant to religious dualism. On the ethnic level, sociological functions of dualism are found in the Zoroastrian opposition (even if not absolute) between Iran, with its so-called "good religion," and the Turanians, northern plunderers representing the aggressive world of evil. But this can by no means substantiate general hypotheses that explain dualistic oppositions between divinities or groups of divinities as a "projection" of a previously existing opposition between ethnic layers of conquerors and of conquered populations.

BIBLIOGRAPHY. UGO BIANCHI, *Il dualismo religioso: saggio storico ed etnologico* (1958), discusses the "dualistic area" extending from ancient Greece to Iran, east European folklore, north and Central Asia, and North America; see also his "Le dualisme en histoire des religions," *Revue de l'histoire des religions*, 159:1–46 (1961); and *Le origini dello gnosticismo* (1967), a collection of papers (in French, German, English, and Italian) presented at the Colloquium of Messina, April 1966, many of which are devoted to dualism in various religions. MIRCEA ELIADE, "Prolegomenon to Religious Dualism: Dyads and Polarities," *The Quest: History and Meaning in Religion*, ch. 8 (1969), is concerned not only with dualism proper but also with the functions of "duality"; his *De Zalmoxis à Gengis-Khan*, ch. 2–3 (1970), is a study of dualism in folklore and ethnology. SIMONE PETREMENT, *Le Dualisme dans l'histoire de la philosophie et des religions* (1946), and *Le Dualisme chez Platon, les Gnostiques et les Manichéens* (1947), are two important general surveys but do not sufficiently distinguish the different meanings of dualism in the philosophical and the religious-historical terminologies. For an analytic exposition of the Gnostic ideology, with modern analogies, see HANS JONAS, *The Gnostic Religion*, 2nd ed. (1963); for a discussion of the varieties of Iranian dualism, R.C. ZAEHNER, *Zurvan: A Zoroastrian Dilemma* (1955). JACQUES DUCHESNE-GUILLEMIN, *The Western Response to Zoroaster* (1958), deals with the history of the problem of Iranian dualism. Other works include: GEO WIDENGREN, "Der iranische Hintergrund der Gnosis," in *Zeitschrift für Religions- und Geistesgeschichte*, 4:97–114 (1952), on dualism in the Indian *Upaniṣads;* F.K. NUMAZAWA, *Die Weltanfänge in der japanischen Mythologie* . . . (1946), a comparative, ethnological appreciation of the Yin–Yang opposition; MARCEL GRIAULE and GERMAINE DIETERLEN, *Le Renard Pâle*, vol. 1, fasc. 1 (1965), an account of dualism in the ontology and the mythology of the Dogon of West Sudan; HELMER RINGGREN, "Dualism," *The Faith of Qumran*, ch. 2 (1963), a study of Qumrānic dualism; and R.M. GRANT, *Gnosticism and Early Christianity*, rev. ed. (1966).

(U.B.)

Dublin

The pre-eminent city of the republic of Ireland, Dublin (Baile Átha Cliath) is the last capital of the Western world whose winter streets are still smudged with the smoke of home fires, redolent of burning peat. It is also the legendary capital of English conversation. How long such agreeable vestiges of the past will endure is problematical: Dublin, so long a relic of a fading colonialism, is, in the 1970s, moving toward its long-denied future. There have been other occasions when Dublin, roused by some freshening of economic or political winds, has hurried to catch up with the rest of the world. What effect the current surge of modernization will have upon the charm of Dublin depends in large measure on the quality of the urban civilization it is hurrying to embrace.



River Liffey and quays, Dublin. In the background (centre) is the dome of the Four Courts.
Laenderpress

Dublin's charm is architectural, geographical, and social. Its architectural heritage—the great garden squares and magnificent terraces of late-18th-century Georgian houses—has suffered badly, a result achieved by the passage of time, and also by the undoubted profitability of new buildings at old distinguished addresses. While the addresses may be distinguished, most of the new buildings are not: Dublin's modern architects seem to lack the wit of its playwrights, the fancy of its poets, the audacity of its horsemen, or even the snug traditionalism of its publicans.

Geographical advantages

The city's geographical site is superb. At the head of a lovely bay, Dublin straddles the River Liffey where it flows eastward through a hill-ringed plain to the shores of the Irish Sea. (The dark bog water made the "black pool" that gave the city its name, Dubh Linn in Irish, Dyfflin in Norse.) It was this opening from the sea, leading through the mountains to the fruitful central plains of Ireland, that originally tempted the wandering Norse raiders to settle there. In spite of its long historical development, the Dublin of the 1970s remains a physically small city, although its twentieth-century population did expand from 390,000 in 1922 to over 650,000 in 1971 (or more than a fifth of the national total population). From Dublin Castle, it is little more than four miles to the furthest city boundary in any direction.

A new prosperity, free health care, social insurance, and new housing have substantially improved the lot of Dublin's poor. An erroneous impression of continued widespread misery is nevertheless abetted by the dress of manual labourers, who habitually wear worn, soiled street garb on the job instead of special work clothes. The traditional harpers, or ballad singers, have moved off the streets of Dublin and into the singing pubs, but its beggars—usually native gypsies, once called tinkers, now "itinerants"—are still abroad in the streets.

Dublin is a low-built, steepled city, with few buildings dating from before the 17th century. The Catholic churches are 19th- and 20th-century structures. The tallest new building—Liberty Hall, a trade union headquarters—is 17 stories high, and most of the others are not higher than ten stories. Although Dublin is the largest manufacturing city in the Irish Republic, the factories, aside from the breweries and distilleries, are engaged principally in light manufacture, and they do not unduly obtrude.

Thus Dublin remains a city of red brick, gray stone, and green parks, shade trees, front gardens, canal banks, copper cupolas, and front doorways. Each year the city's suburbs jut farther into the country, but the countryside nonetheless persists in trying to reclaim the city. Dublin is the capital of a predominately agricultural nation, and many of its urban vistas close on a rising blue note of wild mountain.

An agricultural heritage

Country folk are often to be seen in town for the day, and from September to April Dubliners costumed as country gentlemen leave town to hunt with nearby packs of hounds. The biggest social and tourist event of the year is the August horseshow at the Royal Dublin Society's grounds. Every winter the municipal government advertises for tenders from stockmen who want to graze cattle on the spring swards of the largest city park—Phoenix—which also houses the zoo and the establishments of the Papal nuncio, the American ambassador, and the president of the republic.

Altogether, the city belongs more to the hills and the fields around it than to the waters before it, even though Dublin is Ireland's largest port, handling half the country's imports and half of its domestic exports. Some ships berth right in the centre of town, and gulls fan out from the Liffey, mewing and wheeling their white wings. But the town's purest gift from the sea is the quality of its light. It enriches the old brick and softens the cold Dublin stone. In long summer twilights its luminosity gilds the cityscape, and it lends even the dreariest February afternoon a silvery sheen.

Another of the city's charms has been the tempo of its life—urban, yet leisurely, with the numerous pubs and

churches offering a ready excuse from more mundane pursuits. Yet the beat of Dublin life has been subtly accelerating. The city's advertising agencies, to take but one gauge of the incursions of modernity, have tripled their business in recent years, and they appear to be the heralds of a newer, swifter hour.

Of all the charms of Dublin, the most pervasive and durable is the citizenry itself, a people notoriously hospitable and generous. Their speech is often helplessly poetic. Their celebrated conversations frequently are about religion, still a force in Irish life, and often touch on that other Irish trinity of interests: horses, politics, and personalities. But the Dubliners' humour is built on a long heritage of sorrow. After a thousand years of strife and treachery, their wit conceals an ingrained wariness.

Much of Dublin's flavour is that special taste of things at the source. The principal national seats of learning and the centres of financial and commercial power are located there, as is the central government, which controls so many aspects of Irish life. Dublin has long had the habit of authority over the rest of Ireland. It is said that 100 men run the country, and all—the leaders of the professions, the arts, high finance, industry, government, and sports—are to be found in Dublin.

HISTORY

Foundation and early development. From prehistoric times men have dwelt about Dublin Bay, and four of Ireland's five great roads converged near the spot called Baile Átha Cliath ("the town of the hurdle-ford"), the name stamped today on Dublin's postmark. Dublin appeared in Ptolemy's *Guide to Geography* (*c.* AD 140), and 151 years later "the people of Dublin," it was recorded, defeated a Leinster army. Yet, despite indications of habitation here 2,000 years ago, the first settlement for which there is historical proof was not Irish, but Norse.

Norse influences

The Vikings came in the 9th century (*c.* 831) and built upon the ridge above the river's south bank, on the spot where Dublin Castle rose 400 years later. The Norse invaders beat off most Irish attacks against their bridgehead until 1014, when they were defeated at the Battle of Clontarf on the north shore of the bay. They nevertheless reoccupied the town, and Norse Dublin survived and grew, although eventually the Norse kings were reduced to mere earls under Irish overlords. In 1167 they supported Roderic (Rory O'Connor) of Connaught (Connacht), claimant to the high kingship, in driving Dermot Mac Murrough, the hated king of Leinster, into exile. Dermot returned in 1170 with an army of Anglo-Normans from Wales, 60 miles away to the east across the Irish Sea, and retook Dublin. The following year the last of the Norse earls joined Roderic in a futile assault on Dublin. Alarmed lest his Norman vassals should claim Ireland for their own, King Henry II of England then hurried over with an army to affirm English sovereignty. This action was to be the key to the development of Dublin for the next 750 years, for it was to remain in English hands until 1922. For the bulk of this period, until the middle of the 17th century, Dublin remained a small walled medieval town, dominating only the "pale" —the thin strip of English settlement along Ireland's eastern seaboard. These were, nevertheless, troubled times, for, in those 500 years, three Irish uprisings in the city were suppressed, and a Scottish invasion and the ravages of the Black Death endured.

During the Reformation Dublin became Protestant, and, in the English Civil War, its Royalist defenders, after contemplating joining forces with an armed Irish Catholic confederacy, surrendered the city to Oliver Cromwell's English parliamentary army in 1649. By the end of the Cromwell era Dublin was a town of only 9,000 inhabitants. The turreted city wall with its eight gates was a shambles, the two cathedrals tottered, and the dilapidated castle was, as Cromwell himself put it, "the worst in Christendom."

The 18th-century ascendancy. Yet, in the 18th century, Dublin was to become the second city of the British Empire. This remarkable growth began at the end of the 17th century, when thousands of refugee Huguenot weavers from the European continent settled in Protestant Dublin after the revocation of the Edict of Nantes, in 1685, curtailed their privileges. Flemish weavers came in their wake, and soon the cloth trades were flourishing, with high-gabled houses surrounding Weavers' Square, and it was not long before the excellence of the weavers' work provoked export restrictions by the English. In the course of the 18th century, Dublin burst its walls, more bridges were built over the Liffey, and splendid new sections arose to the north and east. The streets—many of them 40 yards broad—were paved and were lighted by oil lamps, while beautifully proportioned brick town houses and an occasional palace built for a wealthy lord bordered the newly laid out squares and thoroughfares. The philanthropic institutions, for which Dublin later came to be admired, were also founded in this period, and a number of hospitals—including the first maternity hospital in the British Isles—were built. The new Parliament House was built on the mound of the Thingmote, the site of the Viking assembly, and behind the new quays on the north bank of the Liffey two great public buildings rose, the Custom House and the Four Courts, the work of the Dublin-born architect James Gandon. The steeples of new churches added to the changing skyline, and a modern port was constructed. Trinity College, founded in 1591, added some of its most beautiful buildings. The urban landscape that emerged is, in essence, that of the Dublin of today.

Georgian architecture

In the New Musick Hall, Handel conducted the first public performance of his *Messiah* in 1742, while such noted literary figures as Oliver Goldsmith, Richard Steele, Richard Brinsley Sheridan, and William Congreve also were active in Dublin during this period. For members of the Protestant Ascendancy, as the English establishment was called, Dublin was a gay, fashionable city of elegance and wit.

It was something less than that for Catholics, however. Toward the beginning of the 18th century the Penal Laws had disenfranchised the Catholics, leaving them without possessions, without professions, without representation, and, in many instances, without education.

The 19th and 20th centuries. The 1800 Act of Union between England and Ireland abolished the Irish Parliament and drastically reduced Dublin's status. With no governmental duties to compel their presence in Dublin, the leading figures of the Ascendancy returned to England, followed by the lesser English exploiters. The city fell into a decline from which it recovered only 150 years later. Dispossessed farmers crowded into the tenantless Georgian houses, reducing them to slums. Anyone who owed more than ten shillings could be imprisoned, and until the legislation was revised in 1864, five of Dublin's nine jails overflowed with debtors.

With the abatement of the Penal Laws, however, a Catholic middle class emerged, sending its sons to university, to law, and to medical school. The political dexterity of the Irish Catholic lawyer Daniel O'Connell won full citizenship for Irish Catholics when he achieved passage of the Emancipation Act by the English Parliament in 1829, and in 1841, after reforms in Dublin's municipal government, O'Connell became the first Catholic lord mayor of the city.

The railways came to Ireland in 1834 when a seven-mile link connected Dublin with the port of Kingston (Dún Laoghaire). A few light industries subsequently opened, and on the city's eastern fringes new houses went up, Victorian in date but Georgian in flavour. For the first time in 200 years Catholic churches and schools were built, and in 1851 the Catholic University of Ireland was founded on Saint Stephen's Green, with John Henry Newman as rector. Urban growth nevertheless remained slow, and Dublin continued to flourish modestly on the surface, but to fester beneath. The nationalist rebellions that broke out there in 1803, 1848, and 1867 were quickly suppressed. The legalist Home Rule movements, despite frequent reverses, did win a measure of home rule for Ireland in 1914; yet, with the outbreak of World War I in the same year, the English government

acquired other preoccupations, and home rule appeared to become a dead letter.

The condition of the poor remained wretched: a 1910 survey showed that 20,000 families were each living in one room; wages were very low, and a two-week survey of 22 public houses, or taverns, showed over 46,000 women, and 28,000 children, among the customers.

Cultural revival

By the middle of World War I the cultural renaissance known as the Irish Revival was well underway, with Dublin playing a leading role. This movement had begun with the establishment, in 1884, of the Gaelic Athletic Association for the revival of Irish games, and was broadened in 1893 by the foundation of the Gaelic League for the revival of Irish language and folklore. Early in the 20th century it received strong impetus with the opening of Dublin's famous Abbey Theatre for the revival of Irish drama, an enterprise associated with, among others, the great Irish poet William Butler Yeats and the playwright John Millington Synge. The net effect of these developments was to help crystallize a strong sense of national identity and a renewed pride in Irish heritage. The nascent Irish labour movement, fighting the Ascendancy, grew increasingly patriotic, even forming a small citizen army. The 1848 Fenian rebel movement awoke from half a century's hibernation, and the secret, revolutionary Irish Republican Brotherhood (IRB) was born. These trends were to prove of momentous significance to Dublin, for it was in Dublin, on Easter Monday, 1916, that leaders marched at the head of 1,000 men of the recently formed Irish Volunteers and the Citizen Army, proclaimed a republic, and occupied public buildings, which they held for a week against British troops and artillery. The bottom half of Sackville Street, Dublin's principal thoroughfare, was destroyed in the melee. Commerce and industry were halted, and a quarter of the city's population of 390,000 went on public relief. Finally defeated, the rebels were marched through the streets of Dublin to the jeers of the populace. But British martial law, execution of the leaders, and imprisonment of many of the survivors roused the Irish as the rebellion itself had not. Guerrilla warfare started throughout the nation in August 1919, continuing through two years of terror and counterterror, with the forces of both sides directed from Dublin. After a final truce was concluded in 1921, the British offered Ireland —except the northern province of Ulster—dominion status as a free state. Although the subsequent treaty was accepted by the revolutionary parliament, an antitreaty contingent of the republican army soon took possession of one of Dublin's great public buildings, the Four Courts. Eventually they were driven out by artillery, whose booming initiated 11 months of murderous civil war between the two Irish factions. Dublin again suffered heavily in the conflict that ensued, and the upper half of Sackville Street (later named O'Connell Street) was destroyed. The end of the civil war in 1922 did not mean the end of gunfire in the Dublin streets. Political assassinations and armed raids continued until 1927, and bitterness between protreaty and antitreaty partisans marked much of Dublin life.

Post-World War II developments

It was not until a decade after World War II that the government in Dublin launched new schemes intended to steer Ireland into the mainstream of the 20th century. Among these were detailed five-year plans for economic development; active pursuit of investment for new industry, especially from foreign sources; membership in more international political, economic, and cultural organizations; vigorous reorganization of national companies; and heavy investment in the tourist trade. All had great significance for the development of Dublin, and, consequently, these acts of social and economic renewal marked the full flowering of the city as an essentially Irish capital.

THE CONTEMPORARY CITY

Topography and institutions. The Norse, the Norman, and the Georgian, the three elements that comprise the architectural legacy of Dublin, all meet in Dublin Castle. In the first two decades of the 13th century the Normans obliterated the Viking stronghold and reared a château-fort. When the Georgians built the present red brick castle they left two towers of the old structure standing. The castle, the seat of British authority in Ireland until 1922, is now used for inauguration of the republic's presidents, who subsequently reside in the former-Viceregal Lodge in Phoenix Park.

Close to the castle, Sigtryggr Silkeskjess (Silkenbeard), Viking king of Dublin, built Christ Church Cathedral (*c.* 1030), which was replaced roughly 140 years later by a more magnificent Norman structure. By the 19th century the edifice was in ramshackle condition and required extensive repairs. At enormous cost, it was restored from 1870 to 1878. Christ Church is the cathedral for the diocese of Dublin and Glendalough, whereas its neighbour, St. Patrick's, is the national cathedral. Both are Protestant churches.

St. Patrick's, founded just outside the city walls, was originally a Danish church. The Normans rebuilt it in 1191, and, although it was enlarged and partially rebuilt over the centuries, it was in a state of collapse when Sir Patrick Guinness, the Irish brewery magnate, paid for its rebuilding, from 1846 to 1849.

The area between St. Patrick's and the Guinness Brewery on the Liffey is known as the Liberties, having been outside the city walls and under the sole jurisdiction of the archbishop. It was made up of the liberties of St. Patrick's and St. Sepulchre's, the latter the archiepiscopal palace, whose gateway is now the entrance to the Kevin Street police barracks. Since World War II large tracts of this district have been cleared for low-cost housing, and many of the district's inhabitants have been relocated in the new town of Ballymurn.

The Guinness Brewery, a distinctive Dublin institution, is the largest single private employer in Ireland, and its 4,000-man payroll contributes substantially to Dublin's total of 29 percent of all wages paid out in the country, and 45 percent of all income taxes taken in.

Barrels of export beer and stout, piled picturesquely on the quays, are important cargo for many of the 5,000 ships that clear Dublin's port each year. The basis of the modern port, the largest in the republic, was laid with remarkable prescience by the Ballast Office, founded in 1705. The office embanked the Liffey as it wound through the city harbour, constructing walls that reached more than five miles out into the bay. They also scoured a deep-water entrance through the roaring sand bars (the North and South Bull), and thus facilitated the reclamation of harbour-side wastelands.

Similar foresight enriched the work of city builders in the same period. Dublin's early private speculators had a sense of order and beauty as acute as their sense of profit. The city's streets were broad, its garden squares capacious. For their time, the 18th century, the houses were ultramodern—Georgian and neoclassical, in the manner of the great English architects Inigo Jones and Christopher Wren. The elegant yet simple facades were severe rectangles of red brick, lightened by the superb proportions of well-placed windows, usually trimmed in white. Standing in long terraces with no space between the buildings, the sweep of homes made a harmonious whole that still stands as a felicitous achievement of urban architecture. In the sections still maintained today, the street doors are painted in many different colours and embellished with burnished brass fittings; they are remnants of a heritage giving way, perhaps irreversibly, to concrete and steel.

Terraces and squares

In the southern half of the town, between Trinity College and St. Stephen's Green, Joshua Dawson, one of Dublin's leading citizens, built an impressive residence in 1705, selling it a decade later to the city of Dublin for the lord mayor's residence; it still serves this purpose. It was here that the first Irish republican parliament, the Dáil Éireann, met in 1919 to ratify the 1916 rebel declaration of the establishment of the republic.

Dawson's neighbours, the equally prominent Molesworths, followed his example and began building houses and entire streets. In 1745 the Earl of Kildare built, at the end of Molesworth Street, a palace that was renamed

The house of Parliament

Leinster House when he became duke of Leinster. Though now the seat of the Irish parliament, it is still called by its original name. This huge town-built country house is believed by some experts to be the prototype for the White House in Washington, D.C., which was designed by the Irish architect James Hoban in 1780. The Irish senate, the upper house, meets in the 18th-century salon, and the Dáil in the remodelled octagonal lecture theatre built by the Royal Dublin Society after its purchase of Leinster House in 1815. Twin Victorian buildings built on the same grounds in 1885 house the National Library and the National Museum.

While the first duke of Leinster was still in residence, Merrion Street was laid out along the eastern edge of the property. Immediately to the east, Merrion Square, another of the great Georgian brick squares, was begun in 1762, while Fitzwilliam Square, to the south, was built at the end of the 18th century.

Today the square is largely given over to institutions of higher study and diplomatic representation, though the fine houses remain generally intact. The oldest and largest of the city's squares is St. Stephen's Green, recorded in 1224 as common grazing land, and enclosed and bordered with houses in 1663, although the imposing mansions now surrounding it were built principally in the 18th century. By 1887 the parkland was run down and the Guinness family, whose former residence now houses the Department of External Affairs, paid for its rehabilitation.

From the western side of St. Stephen's Green to the river, and up the northern bank to Parnell Square, runs the city's north–south axis, Grafton Street, which has long been the street of Dublin's smart shops. It emerges onto College Green between Trinity College and the 1729 Parliament House, which has since become the headquarters of the Bank of Ireland. Further north O'Connell —formerly Carlisle—Bridge crosses the Liffey. Built as a narrow hump-backed bridge in 1794, it was renamed, flattened, and widened in 1880, and is popularly believed to be as broad as it is long. The river is also spanned by eight other road bridges, two railway bridges, and a pedestrians' crossing called Metal Bridge.

Along the quays are Dublin's finest monumental buildings, with the neoclassical masterpieces the Custom House (1781–91) to the east and the Four Courts (1786–1802) to the west. The former was set afire in 1921 by Republicans who wished to destroy British administrative records, and burned for five days. The latter was reduced by shell fire and mines at the outbreak of civil war in June 1922. Both have since been rebuilt, by government departments, with approximate authenticity but some loss of grace. O'Connell Street—at first called Drogheda and then Sackville Street—is Dublin's "downtown," an assemblage of shops, cinemas, and snack bars. The only building of distinction to survive the warfare that swept the street in 1916 and in 1923 is the General Post Office, headquarters of the 1916 rising. Badly damaged, it was reconstructed behind its surviving 1814 classical facade in 1929.

O'Connell Street was built in two installments. The upper half was created in 1744 by the Dublin banker Luke Gardiner, later Lord Mountjoy, and to this portion, forty years later, the Wide Street Commissioners joined a lower half that ran all the way down to the Liffey. At the top of the street Dr. Bartholemew Mosse constructed his Lying-In Hospital, adding over a period of 20 years a pleasure garden, assembly rooms, and the Rotunda, a concert and meeting hall, the profits from which helped support his hospital. The Rotunda is now a cinema, part of the assembly rooms served as the historic Gate Theater, and another part as a dance hall. The hospital is still devoted to obstetrics.

Behind the hospital Gardiner placed Rutland (now Parnell) Square in 1750. Many of the original Georgian houses are still intact. One, built for the Earl of Charlemont in 1762, now houses the Municipal Gallery of Modern Art. The same developer built nearby Mountjoy Square and the surrounding elegant streets, most of which have fallen into decay. On Marlborough Street,



The Four Courts on the River Liffey, Dublin, completed in 1802.
J. Allan Cash

east of O'Connell, the Catholic procathedral was built in 1816. The Ascendancy would not allow it to be built on the main thoroughfare. The Wide Street Commissioners, in addition to their other creations, circumscribed the whole of the new city with the North and South Circular roads. Describing the part of the North Circular Road he inhabited in the 1920s, the playwright Sean O'Casey called the houses "a long, lurching row of discontented incurables." Close to the South Circular Road, on Synge Street, is the birthplace of the dramatist George Bernard Shaw. North of these peripheral streets, the Grand Canal was constructed in 1756, and south of them, the Royal Canal, in 1789. Their banks were planted with elms, not for the arboreal pleasure that still endures, but to assure the supply of wooden water mains. Both came into the Liffey at the harbour entrance, and both connect with the Shannon River, though only the seldom used Grand is navigable today. The Grand Canal superseded a 13th-century weir near the Dublin Mountains as the city's source of drinking water and is still used by industry. In 1868, a 22-mile-long aqueduct from the Wicklow Hills was opened, and this now brings in soft, filtered, fluoridated drinking water.

Outlying improvements

In Phoenix Park Dublin has one of the world's great city parks, covering nearly three square miles on the north bank of the Liffey. Initially a royal deer park, it was laid out in the middle of the 18th century, and remains little changed today. Its zoo, celebrated for big-cat breeding (known locally as Ireland's lion industry), was founded in 1830.

Phoenix Park contains one of Dublin's three race courses. There is also a greyhound track at the opposite end of town, at Harold's Cross and Ringsend. Fourteen golf courses are dotted around the city. The sacred turf of the traditional Gaelic games, hurling and Irish football, is Croke Park, on the north bank of the Royal Canal. The whole of the bay front is an aquatic playground, while there are a dozen or more other spots officially designated as bathing places. Possibly the best known is five miles from the centre of town, at Sandymount. Above the rocks at Sandymount is the Martello Tower, celebrated in the opening pages of James Joyce's *Ulysses*. It is now a Joyce museum.

Suburbs

South along the coast from Sandymount are Dalkey, Bray, and Greystones—the latter, 16 miles from the city centre, marking the outer limits of the Dublin suburbs. The first seven miles to Dún Laoghaire, where the ferries from Britain land, are heavily built upon, but beyond are stretches of wild hillside and seafront. These shores harbour considerable numbers of retired folk, many of them English, many of them Protestant. They combine with a crescent of suburbs and former suburbs around the southern rim of Dublin—among them Terenure, Rathgar, Rathfarnham, and Ballsbridge—to complete Dublin's "Protestant belt." The Dublin area contains half of

the Presbyterian clergy in the republic, and most of the communicants of the (Protestant) Church of Ireland. Protestants prosper in Dublin, as in the country generally, and include a notably high proportion of business executives and professional men.

The suburb of Ballsbridge, in addition to becoming the legation quarter of Dublin, also houses the new steel-and-glass quarters of some government agencies. After the establishment of the Free State in 1922, the Royal Dublin Society, vacating Leinster House, moved there with large grounds for its horse, dog, and livestock shows. The Royal Dublin Society was founded in 1731 by progressive landlords to improve Irish agriculture and general culture. The National Library, Museum, and College of Art are outgrowths of the Society's activities.

Across the street is the home of the Irish Sweepstakes, established in 1930. Its profits have aided in the construction of 100 hospitals and 300 dispensaries, and annually contribute 12 percent of the amount spent on hospitals in the nation.

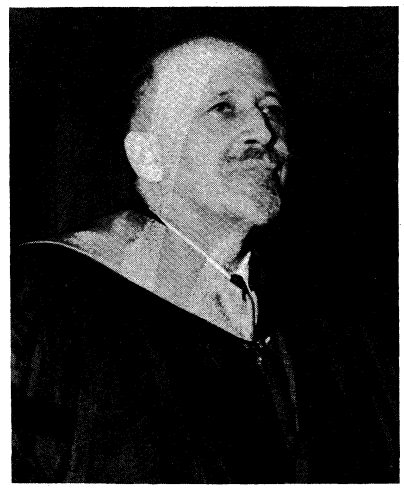
Administration. Dublin is a county borough, and its administrative history goes back to the first charter granted by Henry II of England in 1172. The subsequent turbulent fortunes of the city are mirrored by the numerous later charters—over 100 in all—preserved in the city archives. For parliamentary purposes, Dublin sends 24 members (from six constituencies) to the Dáil Éireann, and is itself governed by an elected corporation and a salaried manager. The functions of the lord mayor, who serves a one year term, are now largely ceremonial, a marked departure from the situation in earlier centuries. The corporation has the power to impose rates, raise loans, and make bylaws. The city manager, on the other hand, has control of the municipal officers. Political relations between the city and the national authorities became strained on several occasions during the 1960s and early 1970s, and the urban government was suspended on one occasion.

BIBLIOGRAPHY. The best introduction to Dublin is V.S. PRITCHETT, *Dublin: A Portrait* (1967), a study that seizes the essential character and brings out the flavour of the city; briefer and more general approaches, depending largely upon illustrations, are J.H. HARVEY, *Dublin* (1949); and DESMOND GUINNESS, *Portrait of Dublin* (1967); while *Dublin Phoenix* by OLIVIA ROBERTSON (1957), is an account illustrated by the author. On the history of the city, see D.A. CHART, *The Story of Dublin*, rev. ed. (1932). On special periods, the following should be consulted: G.A. LITTLE, *Dublin Before the Vikings* (1957); CHARLES HALIDAY, *The Scandinavian Kingdom of Dublin*, 2nd ed. (1884, reprinted 1969), a basic work illustrated by plates, facsimiles, maps (some coloured), and plans; J.P. MAHAFFY, *An Epoch in Irish History* (1903, reprinted 1970), concerned with the foundation and early fortunes of Trinity College, Dublin, during the years 1591–1660; MAURICE CRAIG, *Dublin 1660–1860: A Social and Architectural History*, new ed. (1960); C.E. MAXWELL, *Dublin Under the Georges: 1714–1830*, rev. ed. (1956); J.J. WEBB, *The Guilds of Dublin* (1929, reprinted 1970), and *Municipal Government in Ireland, Mediaeval and Modern* (1918), on aspects of economic and administrative history. On literary Dublin, P. HUTCHINS, *James Joyce's Dublin* (1950); and CYRIL PEARL, *Dublin in Bloomtime* (1969), on the city James Joyce knew, rely heavily on illustrations; R.M. KAIN, *Dublin in the Age of William Butler Yeats and James Joyce* (1962), is a more detailed study. GERARD FAY, *The Abbey Theatre: Cradle of Genius* (1958); and S. MCCANN (ed.), *The Story of the Abbey Theatre* (1967), are good accounts. A.J. HUMPHREYS, *New Dubliners* (1966), is a sociological study of urbanization and the Irish family. Two locally produced guidebooks are the *Official Guide: Dublin* (1955) and *Visitor's Guide to Dublin*, new ed. (1970).

(B.E.)

Du Bois, W.E.B.

W.E.B. Du Bois, America's first black sociologist, was also its most important Negro protest leader in the first half of the 20th century. In the early years of the century, he was the most influential critic of Booker T. Washington's philosophy of accommodation to white supremacy. As a founder (1905) of the Niagara Movement, an organization of black intellectuals agitating for Negro rights, and of the interracial National Association for the Advancement of Colored People (NAACP; 1909), Du Bois played a crucial role in developing the strategy and program that dominated black protest until the rise of the direct-action movement in the 1950s and 1960s. He was also a seminal influence in the Pan-African movement, which attempted to unite blacks all over the world in the struggle for their freedom.



Du Bois.
By courtesy of Atlanta University

Born in Great Barrington, Massachusetts, on February 23, 1868, William Edward Burghardt Du Bois graduated from Fisk University, a black institution, in 1888. He received a Ph.D. from Harvard University in 1895. His doctoral dissertation, *The Suppression of the African Slave-Trade to the United States of America, 1638–1870*, was published in 1896. Although Du Bois took an advanced degree in history, he was broadly trained in the social sciences; and at a time when sociologists were theorizing about race relations, he was conducting empirical inquiries of Negro conditions. For more than a decade he devoted himself to sociological investigations of blacks in America, producing 16 research monographs published between 1897 and 1914 at Atlanta University, where he was a professor, as well as *The Philadelphia Negro; A Social Study* (1899), the first case study of a black community in the United States. Though originally he had believed that social science could provide the knowledge to solve the race problem, he gradually came to the conclusion that in a climate of virulent racism, expressed in such evils as lynching, peonage, disfranchisement, Jim Crow segregation laws, and race riots, social change could be accomplished only through agitation and protest. In this view, he clashed with the most influential black leader of the period, Booker T. Washington, who, preaching a philosophy of accommodation, urged Negroes to accept discrimination for the time being and elevate themselves through hard work and economic gain, thus winning the respect of the whites. In 1903, in his famous book *The Souls of Black Folk*, Du Bois charged that Washington's strategy, rather than freeing the black man from oppression, would serve only to perpetuate it. This attack crystallized the opposition to Booker T. Washington among many Negro intellectuals, polarizing the leaders of the black community into two wings—the "conservative" supporters of Washington and his "radical" critics.

Dispute with Booker T. Washington

Two years later, in 1905, Du Bois took the lead in founding the Niagara Movement, which was dedicated chiefly to attacking the platform of Booker T. Washington. The small organization, which met annually until 1909, was seriously weakened by internal squabbles and Washington's opposition. But it was significant as an ideological forerunner and direct inspiration for the interracial NAACP founded in 1909. Du Bois played a prominent part in the creation of the NAACP and became the association's director of research and editor of its *Crisis* magazine. In this role he wielded an unequalled influence

among middle class blacks and progressive whites as the propagandist for the Negro protest from 1910 until 1934.

Both in the Niagara Movement and in the NAACP Du Bois acted mainly as an integrationist, but his thinking always exhibited, to varying degrees, separatist-nationalist tendencies. In *The Souls of Black Folk* he had expressed the characteristic dualism of black Americans:

> One ever feels his twoness—an American, a Negro; two souls, two thoughts, two unreconciled strivings; two warring ideals in one dark body, whose dogged strength alone keeps it from being torn asunder. . . . He simply wishes to make it possible for a man to be both a Negro and an American, without being cursed and spit upon by his fellows, without having the doors of Opportunity closed roughly in his face.

Du Bois's black nationalism took several forms—the most influential being his pioneering advocacy of Pan-Africanism, the belief that all people of African descent had common interests and should work together in the struggle for their freedom. Du Bois was a leader of the first Pan-African Conference in London in 1900 and the architect of four Pan-African congresses held between 1919 and 1927. Secondly, he articulated a cultural nationalism. As editor of the *Crisis* he encouraged the development of black literature and art and urged his readers to see "Beauty in Black." Thirdly, Du Bois's black nationalism is seen in his belief that Negroes should develop a separate "group economy" of producers-and-consumers cooperatives as a weapon for fighting economic discrimination and black poverty. This doctrine became especially important during the economic catastrophe of the 1930s and precipitated an ideological struggle within the NAACP.

Black nationalism

He resigned from the NAACP in 1934, charging that the organization was dedicated to the interests of the black bourgeoisie and ignored the problems of the masses. Du Bois's interest in cooperatives was not only part of his nationalism but it also developed out of his Marxist leanings. At the turn of the century, he had been an advocate of black capitalism and Negro support of Negro business, but by about 1905 he had been drawn toward Socialist doctrines. Although he joined the Socialist Party only briefly in 1912, he remained sympathetic with Marxist ideas throughout the rest of his life.

Upon leaving the NAACP he returned to Atlanta University, where he devoted the next ten years to teaching and scholarship. In 1940 he founded the magazine, *Phylon*, Atlanta University's "Review of Race and Culture." In 1945 he published the "Preparatory Volume" of a projected encyclopaedia of the Negro, for which he had been appointed editor in chief. He also produced two major books during this period. *Black Reconstruction: An Essay toward a History of the Part Which Black Folk Played in the Attempt to Reconstruct Democracy in America, 1860–1880* (1935) was an important Marxist interpretation of the Reconstruction era (the period following the Civil War during which the seceded Southern states were reorganized according to the wishes of Congress), and more significantly, it provided the first synthesis of existing knowledge on the role of black men in that critical period of American history. In 1940 appeared *Dusk of Dawn*, subtitled *An Essay Toward an Autobiography of a Race Concept*. In this brilliant book, Du Bois explained his role in both the African and Afro-American struggles for freedom, viewing his career as an ideological case study illuminating the varied and complex facets of the black–white conflict.

After giving up the editorship of the *Crisis* in 1934, Du Bois's influence as a race leader ended. Following a fruitful decade of teaching and publication at Atlanta University, he returned once more to a research position at the NAACP (1944–48). This brief connection ended in a second bitter quarrel, and thereafter Du Bois moved steadily leftward politically in his sympathies and writings. Identified with pro-Russian causes, he was indicted in 1951 as an unregistered agent for a foreign power. Although a federal judge directed his acquittal, Du Bois had become more disillusioned than ever about the United States. In 1961 he joined the Communist Party and, moving to Ghana, renounced his American citizenship over a year later. On August 27, 1963, at the age of 95, he died in Accra, Ghana.

Indictment

BIBLIOGRAPHY. There are two major biographies of Du Bois: FRANCIS L. BRODERICK, *W.E.B. Du Bois: Negro Leader in a Time of Crisis* (1959); and ELLIOTT M. RUDWICK, *W.E.B. Du Bois: Propagandist of the Negro Protest*, 2nd ed. (1968). See also *Dusk of Dawn* (1940), and *Autobiography of W.E.B. Du Bois* (1968), two autobiographical works; MEYER WEINBERG (ed.), *W.E.B. Du Bois: A Reader* (1970); and PHILIP S. FONER (ed.), *W.E.B. Du Bois Speaks* (1970), two anthologies of selected writings. Other relevant books are: AUGUST MEIER, *Negro Thought in America, 1880–1915* (1963); and CHARLES F. KELLOGG, *NAACP: A History of the National Association for the Advancement of Colored People*, vol. 1, *1909–1920* (1967).

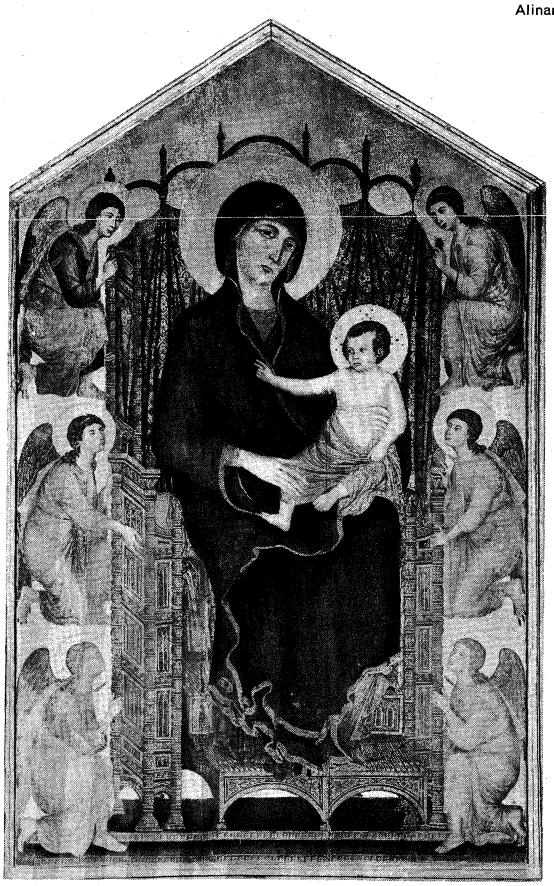
(El.R.)

Duccio di Buoninsegna

In the work of Duccio di Buoninsegna, one of the greatest Italian painters of the Middle Ages and the founder of the Sienese school, the formality of the ancient Byzantine tradition, strengthened by a clearer understanding of its evolution from classical roots, is fused with the new spirituality of the Gothic style, introducing a model blend of East and West. Although his formal style was quickly superseded by the radical innovations of Giotto, who was slightly younger than he, Duccio exerted immense influence on the painters of his own time and of the succeeding generations. The intimate, lyrical quality of his creations infused with soft religious fervour, his rich yet very delicate colours, and the exquisite musical quality of his lines are some of the elements that served to distinguish Sienese painting from Florentine and to establish Siena as one of the particularly important 14th-century schools of art not only in Italy but throughout all of Europe.

Beginnings. Documented information about Duccio's life and career is somewhat scarce. In large part his life must be reconstructed from the evidence of those works that can be attributed to him with certainty, from the evi-

Alinari



"Madonna Rucellai," tempera on wood by Duccio di Buoninsegna, 1285. In the Uffizi, Florence. 4.5 m × 2.9 m.

dence contained in his stylistic development, and from the learning his paintings reveal. Duccio was born in Siena a little after the middle of the 13th century. His father was from the town of Buoninsegna, near Siena, but at the time of Duccio's birth he lived in the town of Camporegio. He is first mentioned in 1278, when the treasurer of the commune of Siena commissioned him to decorate 12 strongboxes for documents. The following year he was given the task of decorating one of the wooden covers of the account books of the treasury. That Duccio was doing work more appropriate for an artisan than an artist must not lead one to assume that even at this time he was only a beginner. It is known that services of this type were requested, both in Siena and in Florence, of already established painters. Further, the fact that he was designated as "painter" and was working for himself demonstrate that he was a mature and independent artist by 1278. In 1280 Duccio was fined the large sum of 100 lire by the commune of Siena for some unrecorded misconduct. This was the first of a considerable number of fines that the artist incurred at various times and for various reasons, and they suggest that he was of a restless and rebellious temperament. He was fined more than once for nonpayment of debts; in 1295 he was penalized for refusing to pledge allegiance to the head of the *popolo* party; in 1302 for not appearing for military duty; and in the same year for what appears to have been practicing sorcery.

Early commissions

The "Madonna Rucellai." On April 15, 1285, the Compagnia dei Laudesi, or singers of praise, of the Virgin Mary at the church of Sta. Maria Novella in Florence, commissioned "Duccio di Buoninsegna, painter of Siena" to paint a great altarpiece that was to represent the Madonna and Child together with other figures. For the work he was to be paid 150 florins, but if the painting, which had to be "a most beautiful picture" and had to have a gold border, was not satisfactory, the artist would receive no reimbursement. Despite the fact that this employment contract, preserved in the State Archives of Florence, came to light in 1790 and was published in 1854, it was only in 1930 that it was indisputably determined that the document referred to the Madonna of Sta. Maria Novella, now called the "Madonna Rucellai." From the time of Giorgio Vasari, a minor Florentine Renaissance painter who was the earliest, and probably the most influential, biographer of early Italian artists, this altarpiece, which was the largest yet painted, was considered to be a masterpiece of the Florentine painter Cimabue. Vasari's attribution, whereas it was probably due in part to a desire not to deprive the Florentine school and its founder of credit for so brilliant a work, was accepted almost unanimously until the present century because of strong similarities to the work of Cimabue in the "Madonna Rucellai." Some recent critics, no longer able to deny that the work is by Duccio, have concluded that he was a pupil, and in all essentials of his art even an imitator, of Cimabue.

Influence of Cimabue on Duccio

The problem of the relative influence of Cimabue upon Duccio is critically very complex. The "Madonna Rucellai" shows affinities with the work of Cimabue in the type of the Virgin, in the serious and robust Child, and in the faces of the six adoring angels; nevertheless, it reveals strikingly new stylistic innovations in the softness of the angels set in midair, in the elegant and subtle lines, in the first feeling of French Gothic animated sweetness and spirituality, and in the light and shade modulation of the free-flowing, clear brush strokes.

There is no doubt that his knowledge of Cimabue's work was one of the components of Duccio's style at this time, but it was not the predominant, nor even the earliest influence; very probably Cimabue's influence was a late insertion into a personal style that had already evolved within the framework of the well-developed Sienese tradition. In the years between 1260 and 1280, largely due to the inspiration of its magnificent cathedral, Siena had emerged as one of the most vital centres of art in Italy. A remarkable succession of altarpieces by Sienese painters testifies to the simultaneous work of a number of artists, some of whom possessed quite distinct personalities. The variety of orientations of these painters shows that they did not work in conditions of provincial isolation but were sensitive to the diverse influences of the age, including Cimabue.

Other stylistic influences

Duccio certainly studied these painters and was influenced by them. Notably evident in his style are the influence of the older painter, Guido da Siena, with the serene dignity of his figures, permeated by lyrical tenderness and grace, in the now-fading stylized postures of the Byzantine tradition, and of the master of the "St. John the Baptist Altarpiece" in the Pinacoteca Nazionale of Siena, with his complex Byzantine iconography and his vivid, dense colouring. Duccio was able to draw from sources outside Siena as well: from the combination of linear stylization and Hellenistic types that characterized the illustrations of books imported from Constantinople and also from contemporary French Gothic miniatures, with their lively tone and lyrical, animated stylizations of clothing and gesture. Duccio may also have travelled to Florence in his early years, coming into contact with Cimabue, but such an explanation is not entirely necessary to account for the formation of his style. In fact, in Duccio's only certain work prior to the "Madonna Rucellai," echoes of Cimabue are even less apparent than in the Rucellai altarpiece. The conclusion that Duccio was nothing more than a follower of Cimabue at the time he painted the "Madonna Rucellai" is implausible and overlooks the originality, as well as the excellence, of the work. If, in fact, he was in 1285 entrusted with a work of such significance at Florence, his reputation must have already been established and have spread beyond the confines of his native Siena.

Later commissions. Traces of Duccio's association with Cimabue remain in the large round stained-glass window of the choir of the Siena Cathedral, for which Duccio made the designs. This work was commissioned between 1287 and 1288 and is the earliest known example of stained glass produced by an Italian.

Numerous documents attest to Duccio's action in Siena during the 20 years following the creation of the "Madonna Rucellai." He was by now the leading painter of the city and as such executed in 1302 an altarpiece, now lost, for the altar of the chapel of the Palazzo Pubblico, the city hall. During this period, some unsigned and undocumented altarpieces appeared, and some of these are certainly Duccio's work; the most significant of these is a small altarpiece representing the Virgin enthroned with angels and called "The Madonna of the Franciscans" because of the three monks kneeling at the foot of the throne. In this work a developed Gothic style appears in the curving outlines, which give an exquisite decorative effect.

The "Maestà." But the work in which the genius of Duccio unfolds in all its brilliant fullness and the one to which the painter owes his greatest fame is the "Maestà," the altarpiece for the main altar of the Cathedral of Siena. He was commissioned to do this work on October 9, 1308, for a payment of 3,000 gold florins, the highest figure paid to an artist up to that time. On June 9, 1311, the whole populace of Siena, headed by the clergy and civil administration of the city, gathered at the artist's workshop to receive the finished masterpiece. They carried it in solemn procession to the accompaniment of drums and trumpets to the cathedral. For three days alms were distributed to the poor and great feasts were held. Never before had the birth of a work of art been greeted with such public jubilation and never before had there been such immediate awareness that a work was truly a masterpiece and not just a reflection of the religious fervour of the people. Duccio himself was aware of the work's significance; he signed the throne of the Virgin with an invocation that was devout yet proud for the time: "Holy Mother of God, grant peace to Siena, and life to Duccio because he has painted you thus."

Form of the "Maestà"

The "Maestà" is in the form of a large horizontal rectangle, surmounted by pinnacles, and with a narrow horizontal panel, or predella, as its base. It is painted on both sides. The entire central rectangle of the front side is a single scene showing the Madonna and Child en-

throned in the middle of a heavenly court of saints and angels with the four patron saints of Siena kneeling at their feet. The back is subdivided into 26 compartments that illustrate the Passion of Christ. The front and back of the predella contain scenes of the infancy and the ministry of Jesus, and the pinnacles, crowning the entire work, represent events after the Resurrection. In all, there are 59 narrative scenes.

The rigorous symmetry with which the groups of adoring figures at the sides of the Virgin are arranged in the imposing scene of the central panel is inspired by compositions of the Byzantine tradition and gives evidence of Duccio's keen architectural sensibility by its power to draw attention to the "Maestà" as the true focal point of the cathedral's spatial and structural organization. Like elements of a living architecture, the 30 figures, through the slightest of gestures and turnings of the head, are intimately related, their positions repeated to give a feeling of intense lyrical contemplation. The consonance of feeling that arises from this contemplation gives the facial features of each a distinct, spiritual beauty, reminiscent, especially the faces of the angels, of the more idealistic creations of Hellenistic art. The Madonna, slightly larger than the other figures, seated on a magnificent and massive throne of polychrome marbles, inclines her head gently as if trying to hear the prayer of the faithful. Duccio thus succeeds in reconciling perfectly the Byzantine ideal of power and dignity with the underlying tenderness and mysticism of the Sienese spirit. The scenes in the predella, pinnacles, and back are filled with the Byzantine iconographic schemes from which Duccio finds it difficult to detach himself, and they are developed with a deeper concern for their narrative significance. The scenes are not, however, merely descriptions or chronicles. They include many touches from daily life, which provide a lyrical synthesis that harmonizes the character and gestures of the figures with their landscape and architectural surroundings.

Unity of Byzantine and Gothic in the "Maestà"

Last years. Only scanty bits of information are available about the few years that Duccio lived after the completion of the "Maestà." He had a prosperous workshop from which other works emerged, but they seem to have been executed in great part by students. His financial condition must have been quite sound because by 1304 he bought a vineyard in the neighbourhood of Siena. Nevertheless, in 1313 he was once again deep in debt. He died either at the end of 1318 or in the first half of 1319, survived by his wife Taviana and seven children. At least two of his children, Galgano and Giorgio, were painters, but nothing is known about their work or their merits. The identity of one of his direct followers is known, his nephew Segna di Buonaventura.

MAJOR WORKS

"Madonna Rucellai" (1285; Uffizi, Florence); "Triptych: The Virgin and Child with Saints" (*c.* 1300; National Gallery, London); "Madonna with Child and Angels" (*c.* 1300; Galleria Nazionale dell'Umbria, Perugia); "Maestà," altarpiece (1302; Palazzo Pubblico, Siena; now lost); "Maestà," altarpiece (1308–11; Museo dell'Opera del Duomo, Siena; panels from this altarpiece can be seen in the National Gallery, London; the Frick Collection, New York; and the National Gallery of Art, Washington, D.C.); "Madonna and Child of St. Cecilia a Crevole" (undated; Museo dell'Opera del Duomo, Siena); "Madonna Enthroned with Angels" (undated; Kunstmuseum, Bern); "The Madonna of the Franciscans" (undated; Pinacoteca Nazionale, Siena); "Madonna and Child" (undated; Stoclet Collection, Brussels); "St. Paul" (undated; Christian Museum [Keresztény Múzeum], Esztergom, Hungary); "The Annunciation" (undated; National Gallery, London); "Nativity with the Prophets Isaiah and Ezekiel" (undated; National Gallery of Art, Washington, D.C.); "The Presentation in the Temple" (undated; Museo dell'Opera del Duomo, Siena); "The Marriage at Cana" (undated; Museo dell'Opera del Duomo, Siena).

BIBLIOGRAPHY. C.H. WEIGELT, *Duccio di Buoninsegna* (1911), the first and the longest monograph on the artist (in German), prepared with great scientific exactitude, although some of the conclusions have been modified or superseded by succeeding studies; R.S. VAN MARLE, *The Development of the Italian Schools of Painting*, vol. 2 (1924), the chapter on Duccio constitutes the fullest treatment of the artist published in English, although it does not present any original ideas (also important for its treatment of the school of Duccio); E. CECCHI, *Trecentisti senesi* (1928), contains some very fine pages on Duccio written by an author who was an excellent art critic; P. BACCI, *Documenti e commenti per la storia dell'arte*, pp. 1–47 (1944), a chapter dedicated to the research in the archives on Duccio and his family; E. CARLI, *Vetrata duccesca* (1946), a study of the great stained-glass window in the choir of the Siena Cathedral, establishing the time it was executed and revealing the authorship of Duccio; *Duccio di Buoninsegna* (1961), a wise interpretation, in Italian, of the work of Duccio, including the entire corpus of his established works reproduced in colour; C. BRANDI, *Duccio* (1951), a thoroughly researched, critical study (in Italian) of the works of Duccio in the light of modern aesthetic theories, with an appendix containing accurate biographical data, chronological lists of documents, and philological notes; G. VIGNI, "Duccio di Buoninsegna," in the *Encyclopedia of World Art*, vol. 4, col. 503–509 (1961), an accurate and acute biographical and critical synthesis founded on the most recent research.

(E.Ca.)

Duchamp, Marcel

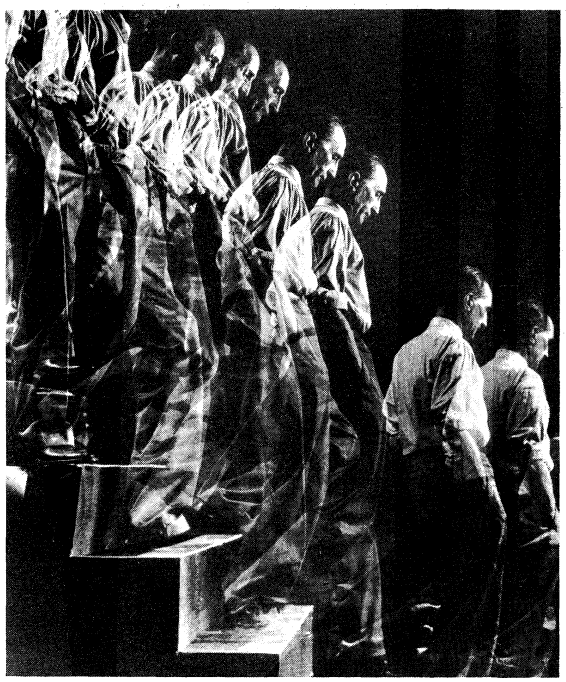
As artist and anti-artist, Marcel Duchamp is considered one of the leading spirits of 20th-century painting. With the exception of the "Nude Descending a Staircase, No. 2," however, a painting that created a sensation in New York in 1913, his works were ignored by the public for the greater part of his life. Until 1960 only such avant-garde groups as the Surrealists claimed that he was important, while to "official" art circles and sophisticated critics he appeared to be merely an eccentric and something of a failure. He was well over 70 when he emerged in the United States as the secret master whose entirely new attitude toward art and society, far from being negative or nihilistic, had led the way to "Pop art," "Op art," and many of the other movements embraced by younger artists everywhere. Not only did he change the visual arts but he also changed the mind of the artist.

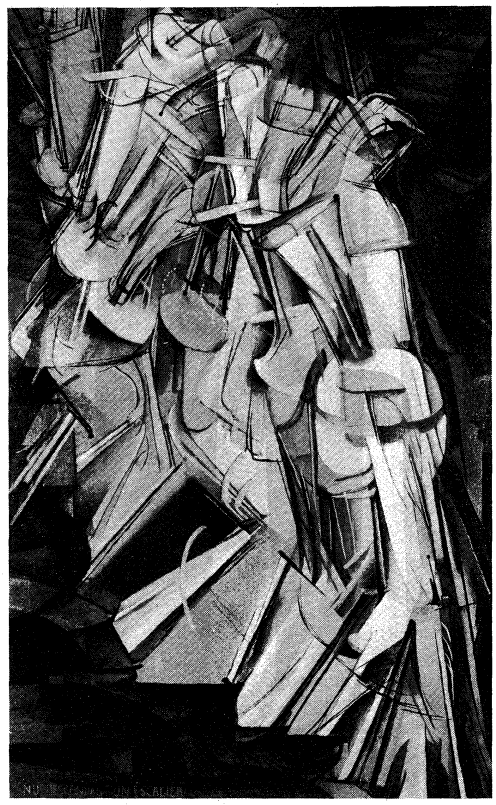
Early years. Marcel Duchamp was born at Blainville in Normandy, France, on July 28, 1887. Although his father was a notary, the family had an artistic tradition stemming from his grandfather, a shipping agent who practiced engraving seriously. Four of the six Duchamp children became artists. Gaston, born in 1875, was later known as Jacques Villon, and Raymond, born in 1876, called himself Duchamp-Villon. Marcel, the youngest of the boys, and his sister Suzanne (born in 1889) both kept the name Duchamp as artists.

An artistic heritage

Journey through styles. When Marcel arrived in Paris in October 1904, his two elder brothers were already in a position to help him. He had done some painting at home, and his "Portrait of Marcel Lefrançois" shows him already in possession of a style and of a technique. During the next few years, while drawing cartoons for comic magazines, Duchamp passed rapidly through the main contemporary trends in painting—Postimpressionism, the influence of Paul Cézanne, Fauvism, and finally Cubism. He was merely experimenting, seeing no virtue in making a habit of any one style. He was outside artistic tradition not only in shunning repetition but also in not attempting a prolific output or frequent exhibition of his work. In the Fauvist style Marcel painted some of his best early work three or four years after the Fauvist movement itself had died away. The "Portrait of the Artist's Father" is a notable example. Only in 1911 did he begin to paint in a manner that showed a trace of Cubism. He had then become a friend of the poet Guillaume Apollinaire, a strong supporter of Cubism and of everything avant-garde in the arts. Another of his close friends was Francis Picabia, himself a painter in the most orthodox style of Impressionism until 1909, when he felt the need of complete change. Duchamp shared with him the feeling that Cubism was too systematic, too static and "boring." They both passed directly from "semirealism" to a "nonobjective" expression of movement. There they met "Futurism" and "Abstractionism," which before they had known only by name.

The "Nude." To an exhibition in 1911, Duchamp sent a "Portrait" that comprised a series of five almost mono-





Duchamp descending a staircase (above), multiple exposure photograph by Eliot Elisofon, 1952, parallels (right) "Nude Descending a Staircase, No. 2," oil on canvas by Duchamp, 1912. In the Philadelphia Museum of Art. 1.47 m. × 88.9 cm.
By courtesy of (right) the Philadelphia Museum of Art, the Louise and Walter Arensberg Collection; photograph, (above) Eliot Elisofon

chromatic, superimposed silhouettes. In this juxtaposition of successive phases of the movement of a single body appears the idea for the "Nude Descending a Staircase, No. 2." The main difference between the two works is that in the earlier one the kangaroo-like silhouettes can be distinguished. In the "Nude," on the other hand, there is no nude at all but only a descending machine, a nonobjective and virtually cinematic effect that was entirely new in painting.

Critical reaction to the "Nude"

When the "Nude" was brought to the 28th Salon des Indépendants in February 1912, the committee, composed of friends of the Duchamp family, refused to hang the painting. These men were not reactionaries and were well accustomed to Cubism, yet they were unable to accept the novel vision. A year later at the Armory Show in New York, the painting again was singled out from among hundreds that were equally shocking to the public. Whatever it was that made the work so scandalous in Paris, and in New York so tremendous a success, prompted Duchamp to stop painting at the age of 25. A widely held belief is that Duchamp introduced in his work a dimension of irony, almost a mockery of painting itself, that was more than anyone could bear and that undermined his own belief in painting. The title alone was a joke that was resented. Even the Cubists did their best to flatter the eye, but Duchamp's only motive seemed to be provocation.

Farewell to art. In 1912, after the "Nude," Duchamp did a few more paintings. Some of these, notably "Le Passage de la Vierge à la Mariée" and "Mariée" (Philadelphia Museum of Art), both done in Munich, are among the finest works of the period. Again they were neither Cubist, nor Futurist, nor Abstract, but they expressed Duchamp's typical vision of the body perceived in its inmost impulses.

New values. There was no question that as a painter Duchamp was on a footing with the most gifted. What he lacked was faith in art itself, and he sought to replace aesthetic values in his new world with an aggressive intellectualism opposed to the so-called common-sense world. As early as 1913 he began studies for an utterly awkward piece. "The Large Glass, or The Bride Stripped Bare by Her Bachelors, Even." For it, he repudiated entirely what he called retinal art and adopted the geometrical methods of industrial design. It became like the blueprint of a machine, albeit a symbolic one, that embodied his ideas of man, woman, and love.

Like the "Nude," "The Large Glass" was to be unique among works of modern painting. Between 1913 and 1923, Duchamp worked almost exclusively on the preliminary studies and the actual painting of the picture itself. His farewell to painting was by no means a farewell to work.

The ready-made revolution

During this period a stroke of genius led him to a discovery of great importance in contemporary art, the so-called ready-made. In 1913 he produced the "Bicycle Wheel," which was simply an ordinary bicycle wheel. In 1914, "Pharmacy" consisted of a commercial print of a winter landscape, to which he added two small figures reminiscent of pharmacists' bottles. It was nearly 40 years before the ready-mades were seen as more than a derisive gesture against the excessive importance attached to works of art, before their positive values were understood. With the ready-mades, contemporary art became in itself a mixture of creation and criticism.

Conquest of America. When World War I broke out, Duchamp, who was exempt from military service, was living and working in almost complete isolation. He left France for the United States, where he had made friends through the Armory Show. When he landed in New York in June 1915, he was welcomed by reporters as a famous man. His warm reception in intellectual circles as well raised his spirits. The wealthy poet and collector Walter Arensberg arranged a studio for him in his own home, where the painter immediately set to work on "The Large Glass." He became the centre of the Arensberg group, enjoying a reputation that led to many offers from art galleries eager to handle the works of the painter of the "Nude." He refused them all, however, not wanting to start a full-time career as a painter. To support himself, he gave French lessons. He was then, and remained, an artist whose works would have been sought after but who was content to distribute them free among his friends or to sell them for intentionally small amounts. He helped Arensberg buy back as many of his works as could be found, including the "Nude." They became a feature of the Arensberg Collection, which was left to the Philadelphia Museum of Art.

Alliance with Dadaism. Besides "The Large Glass," on which he worked for eight more years until abandoning it in 1923, Duchamp did only a few more ready-mades. One, a urinal entitled "Fountain," he sent to the first exhibition of the Society of Independent Artists, in 1917. Although he was a founder-member of this society, he had signed the work "R. Mutt," and therefore it was refused. His ready-mades had anticipated by a few years the Dada movement, which Picabia introduced to New York in the magazine *291* (1917). As an echo of the movement, Duchamp helped Arensberg and H.P. Roché to publish *The Blind Man*, which had only two issues, and *Rongwrong*, which had only one. Later, with the painter Man Ray, he published a single issue of *New York Dada* in 1921.

In 1918 he sold "The Large Glass," which was still very much unfinished, to Walter Arensberg. With the money from this and another painting, his last, he spent nine months in Buenos Aires, where he heard of the armistice and of the deaths of his brother Raymond Duchamp-Villon and of Guillaume Apollinaire. In Paris in 1919, he stayed with Picabia and established contact with the first Dada group. This was the occasion of his most famous ready-made, a photograph of the "Mona Lisa" with a moustache and a goatee added. The act expressed the Dadaists' scorn for the art of the past, which in their eyes was part of the infamy of a civilization that had produced the horrors of the war just ended.

The mustachioed "Mona Lisa"

In February 1923, Duchamp stopped working on "The Large Glass," considering it definitely and permanently unfinished. As the years passed, art activity of any kind interested him less and less, but the cinema came to fulfill his pleasure in movement. His works to this point had been only potential machines, and it was time for him to create machines that were real, that worked and moved. The first ones were devoted to optics and led to a short film, *Anemic Cinema* (1926). With these and other products, including "optical phonograph records," he acted as a kind of amateur engineer. The modesty of his results, however, was a way by which he could ridicule the ambitions of industry. The rest of the time he was absorbed in chess playing, even taking part in international tournaments and publishing a treatise on the subject in 1932.

To Surrealism. Although Duchamp carefully avoided art circles, he remained in contact with the Surrealist group in Paris, composed of many of his former Dadaist friends. When in 1934 he published the *Green Box*, containing a series of documents related to "The Large Glass," the Surrealist poet André Breton perceived the importance of the painting and wrote the first comprehensive study of Duchamp, which appeared in the Paris magazine *Minotaure* in 1935. From that time on there was a closer association between the Surrealists and Duchamp, who helped Breton to organize all the Surrealist exhibitions from 1938 to 1959. Just before World War II he assembled his *Boîte-en-valise*, a suitcase containing 68 small-scale reproductions of his works. When the Nazis occupied France, he smuggled his material across the border in the course of several trips. Eventually he carried it to New York, where he joined a number of the Surrealists in exile, including Breton, Max Ernst, and Yves Tanguy. He was instrumental in organizing the Surrealist exhibition in New York in October and November 1942.

Unlike his co-exiles, he felt at home in America, where he had many friends. During the war, the exhibition of "The Large Glass" at the Museum of Modern Art, New York, helped to revive his reputation, and a special issue of the art magazine *View* was devoted to him in 1945. Two years later he was back in Paris assisting Breton with the Surrealist exhibition, but he returned to New York promptly and spent most of the remainder of his life there, becoming a U.S. citizen in 1955. After his marriage to Teeny Sattler in 1954, he lived more than ever in semiretirement, content with chess and with producing, as the spirit moved him, some strange and unexpected object.

This contemplative life was interrupted around 1960, when the rising generation of American artists realized that Duchamp had found answers for many of their problems. Suddenly tributes came to him from all over the world. Retrospective shows of his works were organized in America and Europe. Even more astonishing were the replicas of his ready-mades produced in limited editions with his permission, but the greatest surprise was still to come. After his death in Neuilly, near Paris, on October 2, 1968, his friends heard that he had worked secretly for his last 20 years on a major piece called "Étant donnés: 1. la chute d'eau, 2. le gaz d'éclairage" (Given: 1. the waterfall, 2. the illuminating gas"). At the Philadelphia Museum of Art, it offers through two small holes in a heavy wooden door a glimpse of Duchamp's enigma.

MAJOR WORKS

"Church at Blainville" (1902; Philadelphia Museum of Art); "Portrait of Marcel Lefrançois" (1904; Philadelphia Museum of Art); "Nude on Nude" (1909?; Arnold D. Fawcus Collection, Paris); "Portrait of the Artist's Father" (1910; Philadelphia Museum of Art); "Red Nude" (1910; Mary Sisler Collection, New York); "The Chess Players" (1910; Philadelphia Museum of Art); "Landscape" (1911; Museum of Modern Art, New York); "Sad Young Man in a Train" (1911; Peggy Guggenheim Collection, Venice); "Nude Descending a Staircase, No. 1" (1911; Philadelphia Museum of Art); "Portrait" (1911; Philadelphia Museum of Art); "Apropos of Little Sister" (1911; Solomon R. Guggenheim Museum, New York); "Nude Descending a Staircase, No. 2" (1912; Philadelphia Museum of Art); "Le Passage de la Vierge à la Mariée" (1912; Museum of Modern Art, New York); "The King and Queen Surrounded by Swift Nudes" (1912; Philadelphia Museum of Art); "Bride" (1912; Philadelphia Museum of Art); "Bicycle Wheel" (1913; replica, Museum of Modern Art, New York); "3 Standard Stoppages" (1913–14; Museum of Modern Art, New York); "Glider Containing a Water Mill in Neighbouring Metals" (1913–15; Philadelphia Museum of Art); "Bottle Dryer" (1914; replica, Moderna Museet, Stockholm); "Network of Stoppages" (1914; Museum of Modern Art, New York); "Nine Malic Moulds" (1914–15; Mrs. Marcel Duchamp Collection, New York); "In Advance of the Broken Arm" (1915; replica, Yale University Art Gallery); "With Hidden Noise" (1916; Philadelphia Museum of Art); "Nude Descending a Staircase, No. 3" (1916; Philadelphia Museum of Art); "Fountain" (1917; replica, Sidney Janis Collection, New York); "Tu m' " (1918; Yale University Art Gallery); "Nude Descending a Staircase, No. 4" (1918; Museum of the City of New York); "To be Looked at (from the Other Side of the Glass) with One Eye, Close to, for Almost an Hour" (1918; Museum of Modern Art, New York); "L.H.O.O.Q." (1919; Mary Sisler Collection, New York); "Why Not Sneeze Rose Sélavy?" (1921; Philadelphia Museum of Art); "The Large Glass" (1915–23; Philadelphia Museum of Art); "The Brawl at Austerlitz" (1921; W.N. Copley Collection, New York); "Rotary Demi-Sphere" (1925; Museum of Modern Art, New York); "Discs Inscribed with Puns" (1926; W.N. Copley Collection, New York); *The Green Box* (1934); *Rotoreliefs* (1935); "Door for Gradiva" (1937; replica, Dieter Keller Collection, Stuttgart, West Germany); *Boîte-en-valise* (1938–42; Mary Sisler Collection, New York); "In the Manner of Delvaux" (1942; V. and A. Schwarz Collection, Milan); "Pocket Chess Set" (1943; New York); "George Washington" (1943; André Breton Estate, Paris); "Given the Illuminating Gas and the Waterfall" (1948–49; Maria Martins Collection, Rio de Janeiro); "Female Fig Leaf" (1950; Mary Sisler Collection, New York); "Object-Dard" ("Dart-Object"; 1951; Mrs. Marcel Duchamp Collection, New York); "Wedge of Chastity" (1954; Mrs. Marcel Duchamp Collection, New York); "Self-Portrait in Profile" (1958; New York); "Cols alités" (1959; R. Lebel Collection, Paris); "With My Tongue in My Cheek" (1959; R. Lebel Collection, Paris); "Torture-morte" (1959; R. Lebel Collection, Paris); "Sculpture-morte" (1959; R. Lebel Collection, Paris); *Nine Etchings for the Large Glass and Related Works* by A. Schwarz (1967; A. Schwarz Collection, Milan); *Nine Etchings for the Large Glass and Related Works* by A. Schwarz (1968; A. Schwarz Collection, Milan); "Étant donnés: 1. la chute d'eau, 2. le gaz d'éclairage" (1946–66, assembled 1969; Philadelphia Museum of Art).

BIBLIOGRAPHY

Monographs and catalogues raisonnés: ROBERT LEBEL, *Marcel Duchamp* (1959, reprinted with an addenda to the catalogue raisonné and an updated bibliography, 1967), is the first major biography and critical study. The catalog includes a list of replicas, editions, working studies and documents, single copies or reconstructions, and variations on Duchamp's projects. ARTURO SCHWARZ (ed.), *The Complete*

Works of Marcel Duchamp (1969), is the most extensive catalogue raisonné and includes Duchamp's very last works.

Studies: GUILLAUME APOLLINAIRE, *Les Peintres cubistes* (1913; Eng. trans., *The Cubist Painters*, 2nd ed., 1949); CALVIN TOMKINS, *The World of Marcel Duchamp* (1966); PIERRE CABANNE, *Entretiens avec Marcel Duchamp* (1967); ARTURO SCHWARZ, *The Large Glass, and Related Works*, 2 vol. (1967–68); OCTAVIO PAZ, *Marcel Duchamp* (1968), in Spanish.

Major exhibition catalogs: Marcel Duchamp, Pasadena Art Museum: A Retrospective Exhibition, October 8 Through November 3, 1963 (1963), the first major exhibition of Duchamp's works; *Not Seen and/or Less Seen of/by Marcel Duchamp/Rrose Selavy, 1904–64. Mary Sisler Collection. Exhibition January 14–February 13, 1965* (1965); *The Almost Complete Works of Marcel Duchamp: Catalogue of an Exhibition at the Tate Gallery 18 June–31 July 1966* (1966); ROUEN, MUSEE DES BEAUX-ARTS, *Les Duchamps* (1967), and PARIS, MUSEE NATIONAL D'ART MODERNE, *Marcel Duchamp* (1967), the first comprehensive exhibition of Duchamp's works in France.

(R.Le.)

Dulles, John Foster

John Foster Dulles, generally regarded as the most powerful and controversial secretary of state in the history of the United States, became one of the most redoubtable protagonists of the Cold War, the power struggle that arose between the United States and the Soviet Union after World War II. A deeply committed anti-Communist, Dulles, serving as the 53rd secretary of state under Pres. Dwight D. Eisenhower from January 1953 to April 1959, extended the anti-Communist alliance system begun in Europe with the North Atlantic Treaty Organization (NATO) by forming a series of regional pacts with pro-United States Asian governments.

By courtesy of U.S. Information Agency; photograph, Harris and Ewing



Dulles.

Dulles was born in Washington, D.C., on Feb. 25, 1888, one of five children of Allen Macy and Edith (Foster) Dulles. His maternal grandfather was John Watson Foster, who served as secretary of state under Pres. Benjamin Harrison. Robert Lansing, Dulles' uncle by marriage, was secretary of state in the cabinet of Pres. Woodrow Wilson.

Dulles was educated in the public schools of Watertown, N.Y., where his father served as a Presbyterian minister. A brilliant student, he attended Princeton and George Washington universities and the Sorbonne, and in 1911 entered the New York law firm of Sullivan and Cromwell, specializing in international law. By 1927, he was head of the firm.

But Dulles, who never lost sight of his goal of becoming secretary of state, actually started his diplomatic career in 1907 when, aged 19, he accompanied his grandfather John Foster, then a private citizen representing China, to the second international peace conference at The Hague. At 30, Dulles was named by Pres. Woodrow Wilson as legal counsel to the American delegation to the Versailles Peace Conference, at the end of World War I, and afterward he served as a member of the war reparations commission.

In World War II, Dulles helped prepare the United Nations charter at Dumbarton Oaks and in 1945 served as a senior adviser at the San Francisco United Nations conference. When it became apparent that a peace treaty with Japan acceptable to the United States could not be concluded with the participation of the Soviet Union, Pres. Harry Truman and his secretary of state, Dean Acheson, decided not to call a peace conference to negotiate the treaty. Instead, they assigned to Dulles the difficult task of personally negotiating and concluding the treaty. Dulles travelled to the capitals of many of the nations involved, and in 1951 the previously agreed to treaty was signed in San Francisco by Japan and 48 other nations.

Secretary of state

Emboldened by his formidable achievements, Dulles viewed his appointment as secretary of state by President Eisenhower, in January 1953, as a mandate to originate foreign policy, which is normally regarded as the domain of the President. "The State Department," Dulles once told an aide, "can only keep control of foreign policy as long as we have ideas." A man bent on realizing his ideas, he was an assiduous planner; and once he enjoyed President Eisenhower's complete confidence, policy planning flourished during his administration.

Dulles, fully aware that NATO would be effective only for the defense of Western Europe, leaving the Near East, the Far East, and the Pacific islands unprotected, was eager to fill these gaps. He initiated the Manila conference in 1954, which resulted in the Southeast Asia Treaty Organization (SEATO) pact that united eight nations either located in Southeast Asia or with interests there in a neutral defense pact. This treaty was followed in 1955 by the Baghdad Pact, later renamed the Central Treaty Organization, uniting the so-called northern tier countries of the Middle East—Turkey, Iraq, Iran, and Pakistan—in a defense organization.

In Europe, Dulles was instrumental in putting into final form the Austrian State Treaty (1955), restoring Austria's pre-1938 frontiers and forbidding a future union between Germany and Austria, and the Trieste agreement (1954), providing for partition of the free territory between Italy and Yugoslavia.

Three factors determined Dulles' foreign policy: his profound detestation of Communism, which was in part based on his deep religious faith; his powerful personality, which often insisted on leading rather than following public opinion; and his strong belief, as an international lawyer, in the value of treaties. Of these three, passionate hostility to Communism was the leitmotif of his policy. Wherever he went, he carried with him Joseph Stalin's *Problems of Leninism* and impressed upon his aides the need to study it as a blueprint for conquest similar to Adolf Hitler's *Mein Kampf*. He seemed to derive personal satisfaction from pushing the U.S.S.R. to the brink. In fact, in 1956 he wrote in a magazine article that, "if you are scared to go to the brink, you are lost." Once, during the Austrian State Treaty negotiations, he refused to compromise on some minor points, even though the Austrians themselves pleaded with him to do so for fear the U.S.S.R. would walk out. Dulles stood his ground, and the U.S.S.R. yielded.

Policy of brinkmanship

But Dulles could be equally intransigent with the allies of the United States. His insistence upon the establishment of the European Defense Community (EDC) threatened to polarize the free world, when in 1953 he announced that failure to ratify EDC by France would result in an "agonizing reappraisal" of the United States' relations with France. That expression, and Dulles' announcement in a Paris speech that the United States would react with "massive nuclear retaliation" to any Soviet aggression, found a permanent place in the vocabulary of American foreign policy. It can also be argued that Dulles' brusque rejection in July 1956 of Pres. Gamal Abdel Nasser's request for aid in building Egypt's Aswān Dam was the beginning of the end of the influence that the United States had exerted in the Mid-

dle East. In a complete reversal of his former pro-Egyptian policy, Dulles claimed that the president of Egypt was "nothing but a tin-horn Hitler." Although Dulles later conceded that his refusal could have been more subtle, he never wavered in his belief that Nasser, who had already purchased arms from the Soviet bloc, was bound to turn decisively against the United States because he felt that he had the Soviet Union on his side.

Dulles' detractors in the United States and abroad viewed him as harsh, inflexible, and a tactician, rather than an architect of international diplomacy. But President Eisenhower ignored all criticisms. He said of his secretary of state, "He is one of the truly great men of our time." Whatever their opinion of the man and his policies, many leading statesmen of the non-Communist nations have credited his firmness with having checkmated Communist Cold War strategy. Seriously ill with cancer, Dulles resigned his Cabinet position on April 15, 1959. Shortly before he died, at Washington, D.C., the following May 24, he was awarded his country's highest civil decoration, the Medal of Freedom.

BIBLIOGRAPHY. JOHN R. BEAL, *John Foster Dulles*, rev. ed. (1959), a formal biography written by an admirer, a former State Department correspondent for *Time* magazine; EDWARD WEINTAL and CHARLES BARTLETT, *Facing the Brink* (1967), a critical and intimate account of crisis diplomacy and management by Presidents Eisenhower, Kennedy, and Johnson and their Secretaries of State; RICHARD GOOLD-ADAMS, *John Foster Dulles: A Reappraisal* (1962), a critical biography concluding that upon re-evaluation Dulles' concepts and policies will rank higher than his methods used in implementing them; ELEANOR LANSING DULLES, *John Foster Dulles: The Last Year* (1963), a friendly biography by his sister, replete with anecdotal material.

(E.W.)

Duncan, Isadora

Among the first of her profession to raise the interpretive dance from artificiality to the status of a creative art, Isadora Duncan was acclaimed by the foremost musicians, artists, and writers of her day. Although an inspiration to the intelligentsia, with whom she remained a cult, she was often an object of attack by the less broad-minded. Her ideas were too much in advance of their time, and she flouted conventions too flamboyantly to be regarded by the wider public as anything but an advocate of "free love."

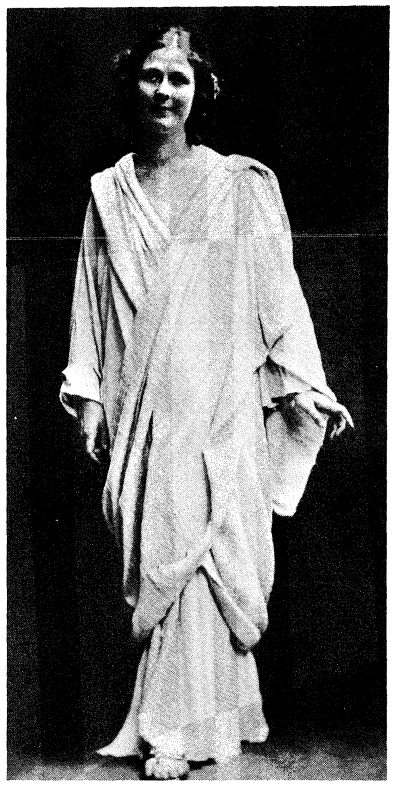
She was born in San Francisco on May 27, 1878, one of three children brought up in genteel poverty by their mother, a music teacher. As a child she rejected the rigidity of the classic ballet and based her dancing on more natural rhythms and movements, an approach she later used consciously in her interpretations of the works of the great composers—particularly Brahms, Wagner, and Beethoven. Her earliest public appearances, in Chicago and New York, met with little success, and at the age of 21 she left America to seek recognition abroad. With her meagre savings she sailed on a cattle boat for England. There her crusading spirit was soon rewarded. At the British Museum her study on the art of ancient Greece confirmed the one-time use of those dance rhythms that hitherto instinct alone had caused her to practice and upon a revival of which her method was largely founded. Through the patronage of the celebrated actress Mrs. Patrick Campbell, she was invited to appear at the private receptions of London's leading hostesses, where her dancing, distinguished by a complete freedom of movement, enraptured those who were familiar only with the conventional forms of the ballet, then in a period of decay. It was not long before the phenomenon of a young woman dancing barefoot, as scantily clad as a woodland nymph, crowded theatres and concert halls throughout Europe. Her controversial first visit to Russia, in 1905, brought the dance to the attention of the art critic Sergey Diaghilev, who as impresario was soon to lead a resurgence of ballet throughout western Europe. At one time or another she founded dance schools in Germany, Russia, and the United States, but none of these survived.

Success in England

Her private life, quite as much as her art, kept her name in the headlines, a situation she deplored yet unwittingly encouraged by her constant defiance of existing taboos. The father of her first child, Deirdre, was the stage designer Gordon Craig, who shared her abhorrence of marriage; the father of her second child, Patrick, was Paris Singer, the heir to a sewing machine fortune and a prominent art patron. With him she enjoyed a love-hate relationship for some years until they finally parted; but not before a tragedy occurred, from which she never really recovered. In 1913 the car in which her two children and their nurse were riding in Paris rolled into the Seine and all three were drowned. In an effort to sublimate her grief she was about to open another school when the advent of World War I put an end to her plans. Her subsequent tours in South America, Germany, and France were less successful than before, but in 1920 she believed the chance had come to realize a lifelong ambition: she was invited to establish a school of her own in Moscow. To someone with her revolutionary temperament, Russia seemed the land of promise: "I go there to meet my destiny," she declared. In fact she met Sergey Aleksandrovich Yesenin, a poet 17 years younger than she, whose work had won him a considerable reputation. She married him in 1922, sacrificing her scruples against marriage in order to take him with her on a tour of America. She could not have chosen a worse time for their arrival. Fear of the "Red Menace" was at its height, and she and her husband were unjustly labelled as Bolshevik agents. When, at the Symphony Hall, Boston, she introduced him at one of her concerts, members of the audience shouted abuse at her. She harangued them from the stage and the evening ended in uproar. Leaving her native country once more, she told reporters: "Good-bye America, I shall never see you again!" She never did. There followed an unhappy period with Yesenin in Europe, where his increasing mental instability turned him against her. He returned alone to the Soviet Union and, in 1925, committed suicide.

Last American tour

During the last years of her life she was a somewhat pathetic figure, living precariously in Nice on the French



Isadora Duncan, c. 1919.

By courtesy of the Dance Collection (Irma Duncan Collection), The New York Public Library at Lincoln Center, Astor, Lenox and Tilden Foundations

Riviera, where she met with a fatal accident on September 14, 1927. While riding in a car, her scarf became entangled in the rear wheel and she was thrown onto the pavement and strangled.

André Maurois' comment on the French novelist, George Sand, applies equally to Isadora Duncan: "Though not technically a chaste woman, she always looked upon an alliance as a marriage. She gave herself freely, entrusted herself wholly. She behaved like an honourable man." Certainly her place as a great innovator in dance is secure: her repudiation of artificial technical restrictions and reliance on the grace of natural movement helped to liberate the dance from its dependence on rigid formulas and on displays of brilliant but empty technical virtuosity, paving the way for the later acceptance of modern dance as it was developed by Mary Wigman, Martha Graham, and others.

BIBLIOGRAPHY. The most comprehensive account of the dancer's life and career is *Isadora: A Revolutionary in Art and Love* by A.R. MACDOUGALL (1960); an excellent profile of Isadora is "The Gloves of Isadora" by R.E. JONES, *Theatre Arts Magazine*, 31:17–22 (1947). The most complete bibliography to date is provided by E.C. WAGENKNECHT in *Seven Daughters of the Theatre* (1964), and the essay contained in it is of particular interest, as are the essays in *Twelve Against the Gods* by W. BOLITHO (1929); and *Here Lives the Heart* by M. DE ACOSTA (1960). Their authors' personal relationships with Isadora Duncan are interestingly related in: *Index to the Story of My Days: Some Memoirs, 1872–1907* by GORDON CRAIG (1957); and *Duncan Dancer: An Autobiography* by IRMA DUNCAN (1966). An exclusive account of Isadora Duncan's South American tour is *An Amazing Journey* by M. DUMESNIL (1932); an account of her years in Russia during the early 1920s is I.I. SCHNEIDER, *Isadora Duncan: The Russian Years*, trans. by D. MAGARSHACK (1969). Studies of the dancer's technique useful for students of the dance are *The Merry-Go-Round* by C. VAN VECHTEN (1918); and *The Technique of Isadora Duncan* by IRMA DUNCAN (1937). Important also for her statement of attitudes and beliefs is ISADORA DUNCAN, *My Life* (1927; reissued as *Isadora*, 1968).

(S.St.)

Duns Scotus, John

John Duns Scotus, called Doctor Subtilis, was a Scholastic philosopher and theologian who pioneered the classical defense of the doctrine that Mary, the mother of Jesus, was conceived without original sin (the Immaculate Conception) and argued that the Incarnation was not dependent on the fact that man had sinned, that will is superior to intellect and love to knowledge, and that the essence of heaven consists in beatific love rather than the vision of God. Though his claim that universal concepts are based on a "common nature" in individuals made him a prime target for the Nominalism of William of Ockham in the 14th-century controversy between Realists and Nominalists concerning the question whether general types are figments of the mind or are real, it later deeply influenced Charles Sanders Peirce, an American philosopher, who considered Duns Scotus the greatest speculative mind of the Middle Ages as well as one of the "profoundest metaphysicians that ever lived." His strong defense of the papacy against the "divine right" of kings made him unpopular with the English Reformers of the 16th century for whom "dunce" (a Dunsman) became a word of obloquy, yet his theory of intuitive cognition suggested to John Calvin, the Genevan Reformer, how God may be "experienced."

Sources on Duns Scotus' life

Early life and career. As Ernest Renan, a 19th-century French historian and philosopher, noted, there is perhaps no other great medieval thinker whose life is as little known as that of Duns Scotus. Yet patient research during the past half-century has unearthed a number of facts. Early 14th-century manuscripts, for instance, state explicitly that John Duns was a Scot, from Duns, who belonged to the English province of Friars Minor (the order founded by Francis of Assisi), that "he flourished at Cambridge, Oxford, and Paris and died in Cologne." In selecting 1966 for an International Congress to commemorate the seventh centenary of his birth and in erecting a cairn near the Pavilion Lodge of the Duns Castle in Berwickshire, scholars honoured not only a centuries-old tradition as to where he was born but also when.

Though accounts of his early schooling and entry into the Franciscan Order are unreliable, Duns Scotus' own remark that "a 13 year old today is better versed in sacred matters than a 20 year old adult in the primitive Church" does suggest an early interest in religion and entrance into the order. As a novice, he would have learned not only of St. Francis' personal love for Christ in the Eucharist, his reverence for the priesthood, and his loyalty to "the Lord Pope"—themes given special emphasis in Duns Scotus' own theology—but also of St. Bonaventure's interpretation of the Franciscan ideal, namely a striving for God through learning that will culminate in a mystical union of love. In his early *Lectura Oxoniensis*, Duns Scotus insisted that theology is not a speculative but a practical science of God and that man's ultimate goal is union with the divine Trinity through love. Though this union is known only by divine revelation, philosophy can prove the existence of an infinite being and herein lies its merit and service to theology. Duns Scotus' own intellectual journey to God is to be found in his prayerful *Tractatus de primo principio* (Eng. trans., *A Treatise on God As First Principle*, 1966), perhaps his last work.

Study in theology

Jurisdictionally, the Scots belonged to the Franciscan province of England, whose principal house of studies was at Oxford where Duns Scotus apparently spent 13 years (1288–1301) preparing for inception as master of theology. There is no record of where he took the eight years of preliminary philosophical training (four for a bachelor's and four for the master's degrees) required to enter such a program.

After studying theology for almost four years, John Duns was ordained priest by Oliver Sutton, bishop of Lincoln (the diocese to which Oxford belonged). Records show the event took place at St. Andrew's Church in Northampton on March 17, 1291. In view of the minimum age requirements for the priesthood, this suggests that Duns Scotus must have been born no later than March 1266, certainly not in 1274 or 1275 as earlier historians maintained.

Duns Scotus would have spent the last four years of the 13-year program as bachelor of theology, devoting the first year to preparing lectures on Peter Lombard's *Sentences*—the textbook of theology in the medieval universities—and the second to delivering them. A bachelor's role at this stage was not to give a literal explanation of this work but rather to pose and solve questions of his own on topics that paralleled subject "distinctions" in Lombard. Consequently the questions John discussed in his *Lectura Oxoniensis* ranged over the whole field of theology. When he had finished, he began to revise and enlarge them with a view to publication. Such a revised version was called an *ordinatio* in contrast to his original notes (*lectura*) or a student report (*reportatio*) of the actual lecture. If such a report was corrected by the lecturer himself, it became a *reportatio examinata*. From a date mentioned in the prologue, it is clear that in 1300 Duns Scotus was already at work on his monumental Oxford commentary on the *Sentences*, known as the *Ordinatio* or *Opus Oxoniense*.

Statutes of the university required that the third year be devoted to lectures on the Bible and in the final year, the bachelor *formatus*, as he was called, had to take part in public disputations under different masters including his own. In John's case this last year can be dated rather precisely, for his name occurs among the 22 Oxford Franciscans, including the two masters of theology, Adam of Howden and Philip of Bridlington, who were presented to Bishop Dalderby on July 26, 1300, for faculties, or the proper permissions to hear confessions of the great crowds that thronged to the Franciscans' church in the city. Because the friars had but one chair of theology and the list of trained bachelors waiting to incept was long, regent masters were replaced annually. Adam was the 28th and Philip the 29th Oxford master, so that Philip's year of regency was just beginning. It must have coincided with Duns Scotus' final and 13th year because an extant disputation of Bridlington as master indicates

John Duns was the bachelor respondent. This means that by June of 1301, he had completed all the requirements for the mastership in theology, yet in view of the long line ahead of him there was little hope of incepting as master at Oxford for perhaps a decade to come.

Years at the University of Paris. When the turn came for the English province to provide a talented candidate for the Franciscan chair of theology at the more prestigious University of Paris, Duns Scotus was appointed. One *reportatio* of his Paris lectures indicates that he began commenting on the *Sentences* there in the autumn of 1302 and continued to June 1303. Before the term ended, however, the university was affected by the long smoldering feud between King Philip IV the Fair and Pope Boniface VIII. The issue was taxation of church property to support the King's wars with England. When Boniface excommunicated him, the monarch retaliated by calling for a general church council to depose the Pope. He won over the French clergy and the university. On June 24, 1303, a great antipapal demonstration took place. Friars paraded in the Paris streets. Berthold of Saint-Denis, bishop of Orleans and former chancellor of the university, together with two Dominicans and two Franciscans, addressed the meeting. On the following day royal commissioners examined each member of the Franciscan house to determine whether he was with or against the King. Some 70 friars, mostly French, sided with Philip, while the rest (some 80 odd) remained loyal to the Pope, among them John Duns Scotus and Master Gonsalvus Hispanus. The penalty was exile from France within three days. Boniface countered with a bull of August 15 suspending the university's right to give degrees in theology or canon and civil law. As a result of his harassment and imprisonment by the King's minister, however, Boniface died in October and was succeeded by Pope Benedict XI. In the interests of peace, Benedict lifted the ban against the university in April 1304, and shortly afterwards the King facilitated the return of students.

Problems with King Philip IV the Fair

Where Duns Scotus spent the exile is unclear. Possibly John's Cambridge lectures stem from this period, although they may have been given during the academic year of 1301–02 before coming to Paris. At any rate, Duns Scotus was back before the summer of 1304, for he was the bachelor respondent in the *disputatio in aula* (public disputation) when his predecessor, Giles of Ligny, was promoted to master. On November 18 of that same year, Gonsalvus, who had been elected minister general of the Franciscan order at the Pentecost chapter, or meeting, assigned as Giles' successor "Friar John Scotus, of whose laudable life, excellent knowledge, and most subtle ability as well as his other remarkable qualities I am fully informed, partly from long experience, partly from report which has spread everywhere."

Literary activity

The period following Duns Scotus' inception as master in 1305 was one of great literary activity. Aided by a staff of associates and secretaries, he set to work to complete his *Ordinatio* begun at Oxford, using not only the Oxford and Cambridge lectures but also those of Paris. A search of manuscripts reveals a magisterial dispute Duns Scotus conducted with the Dominican master, Guillaume Pierre Godin, against the thesis that matter is the principle of individuation (the metaphysical principle that makes an individual thing different from other things of the same species), but so far no questions publically disputed *ordinarie—i.e.*, in regular turn with the other regent masters—have turned up. There is strong evidence, however, that some questions of this sort existed but were eventually incorporated into the *Ordinatio*. Duns Scotus did conduct one solemn quodlibetal disputation, so-called because the master accepted questions on any topic (*de quodlibet*) and from any bachelor or master present (*a quodlibet*). The 21 questions Duns Scotus treated were later revised, enlarged, and organized under two main topics, God and creatures. Though less extensive in scope than the *Ordinatio*, these *Quaestiones quodlibetales* are scarcely less important because they represent his most mature thinking. Indeed, Duns Scotus' fame depends principally on these two major works.

The short but important *Tractatus de primo principio*, a compendium of what reason can prove about God, draws heavily upon the *Ordinatio*. The remaining authentic works seem to represent questions discussed privately for the benefit of the Franciscan student philosophers or theologians. They include, in addition to the *Collationes* (from both Oxford and Paris), the *Quaestiones in Metaphysicam Aristotelis* and a series of logical questions occasioned by the Neoplatonist Porphyry's *Isagoge* and Aristotle's *De praedicamentis*, *De interpretatione*, and *De sophisticis elenchis*. These works certainly postdate the Oxford *Lectura* and may even belong to the Parisian period. Antonius Andreus, an early follower who studied under Duns Scotus at Paris, expressly says his own commentaries on Porphyry and *De praedicamentis* are culled from statements of Duns Scotus *sedentis super cathedram magistralem* ("sitting on the master's chair").

Final period at Cologne. In 1307 Duns Scotus was appointed professor at Cologne. Some have suggested that Gonsalvus sent Scotus to Cologne for his own safety. His controversial claim that Mary need never have contracted original sin seemed to conflict with the doctrine of Christ's universal redemption. Duns Scotus' effort was to show that the perfect mediation would be preventative, not merely curative. Though his brilliant defense of the Immaculate Conception marked the turning point in the history of the doctrine, it was immediately challenged by secular and Dominican colleagues. When the question arose in a solemn quodlibetal disputation, the secular master Jean de Pouilly, for example, declared the Scotist thesis not only improbable, but even heretical. Should anyone be so presumptuous as to assert it, he argued impassionedly, one should proceed against him "not with arguments but otherwise." At a time when Philip IV the Fair had initiated heresy trials against the wealthy Knights Templars, Jean's words have an ominous ring. There seems to have been something hasty about Duns Scotus' departure in any case. Writing a century later, the Scotist William of Vaurouillon referred to the traditional account that John received the Minister General's letter while walking with his students and set out at once for Cologne taking little or nothing with him. Duns Scotus lectured at Cologne until his death, on November 8, 1308. His body at present lies in the nave of the Franciscan church near the Cologne Cathedral and in many places he is venerated as blessed.

Departure from Paris

Whatever the reason for his abrupt departure from Paris, Duns Scotus certainly left his *Ordinatio* and *Quodlibet* unfinished. Eager pupils completed the works, substituting materials from *reportationes examinatae* for the questions Duns Scotus left undictated. The critical Vatican edition begun in 1950 is aimed, among other things, at reconstructing the *Ordinatio* as Duns Scotus left it, with all his corrigenda, or corrections.

Despite their imperfect form, Duns Scotus' works were widely circulated. Even today they are found in hundreds of manuscripts, and from 1472 onward went through more than 30 different editions. During the 16th to 18th centuries among Catholic theologians his following rivalled that of Thomas Aquinas and in the 17th century outnumbered that of all the other schools combined.

BIBLIOGRAPHY. ODULF SCHAFER, *Bibliographia de vita, operibus et doctrina Iohannis Duns Scoti, saec. XIX-XX* (1955), continued in "Resenha Abreviada da Bibliografia Escotistica Mais Recente (1954–1966)," *Revista Portuguesa de Filosofia*, 23:338–363 (1967), are exhaustive bibliographies prepared for the Scotistic Commission. *De doctrina Iohannis Duns Scoti*, 4 vol. (1968), contains 125 papers given at the 2nd International Scotistic Congress 1966 on the philosophy, theology, and influence of Scotus; *Deus et Homo ad mentem I. Duns Scoti* (1972), is a similar collection from the 3rd International Scotistic Congress 1970. A.B. WOLTER (trans.), *Duns Scotus, Philosophical Writings* (1962), includes selections from the *Ordinatio;* and *A Treatise on God As First Principle* (1966), the *De Primo Principio* and selections from the *Lectura*. On the theology of Scotus, see the article by C. BALIC in the *New Catholic Encyclopedia*, vol. 4 (1967). On his philosophy, the following are particularly recommended: A.B. WOLTER in the *Encyclopedia of Philosophy*, vol. 2 (1967), a concise general introduction; EFREM BETTONI, *Duns Scotus: The Basic Principles of His Philosophy* (1961), a more detailed

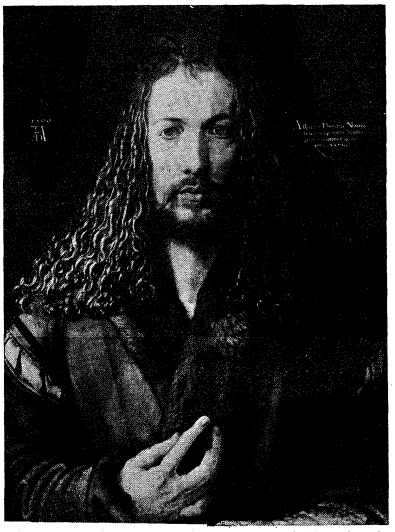
analysis with a good bibliography; J.K. RYAN and B.M. BONANSEA (eds.), *John Duns Scotus, 1265–1965* (1965), a series of essays; F.C. COPLESTON, *A History of Philosophy*, 2nd ed., vol. 2 (1950); and D.E. SHARP, *Franciscan Philosophy at Oxford in the Thirteenth Century* (1930).

(A.B.W.)

Dürer, Albrecht

Albrecht Dürer was the foremost German painter, engraver, and woodcut and decorative designer of the Renaissance, as well as the author of a number of theoretical works. His art has been accepted as a model by painters and engravers of succeeding ages, but his significance does not rest on that fact alone. The versatility of his creativity and his ability to reflect on his own individuality and on his activity and to set down methodically on paper the result of his reflections have also contributed to his importance. He was, like Leonardo, an intellectual artist who was capable of taking part in the great contemporary intellectual upheavals brought about by the Reformation and the humanist movement and who was personally associated with various humanists and members of the circles around Martin Luther.

By courtesy of the Alte Pinakothek, Munich



Dürer, self-portrait, oil on wood panel, 1500. In the Alte Pinakothek, Munich. 65 cm × 48 cm.

The influence that Dürer exerted and still exerts arises mainly from the fact that, as a true Renaissance man who felt himself to be a distinct, important "personality," he left behind a large amount of self-revealing material, including painted and sketched self-portraits, letters, a family chronicle (1524), a fragmentarily preserved *Gedenkbuch* (journal) for the years 1502–03 and 1514, a diary from his trip to the Netherlands (1520–21), the account of an apocalyptic vision (1525), and his poetry.

Education and early career. Dürer was born in Nürnberg, Germany, on May 21, 1471. He was the second son of the goldsmith Albrecht Dürer the Elder (1427–1502), who had left Hungary to settle in Nürnberg in 1455, and of Barbara Holper, who was born in Nürnberg. Dürer began his training as a draughtsman in the goldsmith's workshop of his father. His precocious skill is evidenced by a remarkable self-portrait done in 1484, when he was 13 years old (Albertina, Vienna), and by a "Madonna with Musical Angels," a pen drawing done in 1485 (Kupferstichkabinett, Staatliche Museen Preussischer Kulturbesitz, Berlin), which is already a finished work of art in the late Gothic style. In 1486, Dürer's father arranged for his apprenticeship to the painter and woodcut illustrator Michael Wohlgemuth (1434–1519), a skillful, somewhat prosaic Nürnberg artist, whom he portrayed in 1516 as a pious 82-year-old man (panel in Germanisches Nationalmuseum, Nürnberg). After three years in Wohlgemuth's workshop, he left for a period of travel. In 1490 Dürer completed his earliest known painting, a portrait of his father (Uffizi, Florence) that heralds the familiar characteristic style of the mature master.

Precociousness of Dürer's first works

Dürer's years as a journeyman from 1490 to 1494 probably took the young artist first to the Netherlands and then in 1492 to Colmar in Alsace. There he found that Martin Schongauer, the leading German graphic artist of the time, was no longer alive; several of Dürer's drawings from the time immediately preceding his travels, however, bear signs that he had already seriously studied the work of Schongauer. In the same year, 1492, Dürer went to Basel, Switzerland; there he completed his first authenticated woodcut, a picture of "St. Jerome Curing the Lion" (Kupferstichkabinett, Kunstmuseum, Basel), which seems somewhat ungainly and archaic in comparison with the 1490 portrait of the elder Dürer. During 1493 or 1494 Dürer was in Strasbourg for a short time, returning again to Basel to design several book illustrations. A few surviving works have been ascribed by later research to this series, the *Comedies of Terence*, drawings on wood blocks: "Der Ritter vom Turn" ("The Knight of Turn"), dated 1493, and "Das Narrenschyff" ("Ship of Fools"), dated 1494, both woodcuts. In these works can be seen certain stylistic affinities with Dürer's authenticated and, in general, more mature drawings from the years 1493–95. An early masterpiece from this same period is a self-portrait with a thistle painted on parchment in 1493 (Louvre, Paris). Several small religious paintings (one example is in Staatliche Kunsthalle, Karlsruhe) also belong to this period of Dürer's travels.

First journey to Italy. At the end of May 1494, Dürer returned to Nürnberg, where on July 7, 1494, he married Agnes Frey, the daughter of a merchant. His wife was his only survivor, for they had no children. In the autumn of 1494 Dürer seems to have undertaken his first journey to Italy, where he remained until the spring of 1495. A number of bold watercolours dealing with subjects from the Alps of the southern Tirol are generally linked with this journey; they are among Dürer's most beautiful creations. Depicting segments of reality cleverly chosen for their compositional values, they are painted with broad strokes, in places roughly sketched in, with an amazing harmonization of detail. Dürer used predominantly unmixed, cool, sombre colours, which, despite his failure to contrast light and dark adequately, still suggest depth and atmosphere.

Landscape watercolours

After Dürer's return to Nürnberg, he began to associate frequently with a boyhood friend, the humanist Willibald Pirkheimer (1470–1530), with whom he eventually formed a lifelong friendship. Pirkheimer, who had studied at the great universities of Padua and Pavia, was a leading German humanist scholar. His knowledge of the history and languages of classical times, as well as his extensive library, probably contributed decisively to Dürer's intellectual growth.

The trip to Italy had a strong effect on Dürer; direct and indirect echoes of Italian art are apparent in most of his drawings, paintings, and graphics of the decade between 1495 and 1505, the date of his second trip. Between 1494 and 1496 especially, Dürer produced a large number of drawings based mainly on engravings from northern Italy, copies of which he may have seen before his journey there. Some of these drawings are taken from engravings by the Venetian Andrea Mantegna (Winckler 59, 60; 1494; Albertina, Vienna), an artist greatly preoccupied with classical themes and with precise linear articulation of the human figure. Twenty of the drawings follow the so-called tarocchi ("tarot cards") engravings from about 1460 based on drawings by an unknown Ferrarese artist (W 122–141; British Museum, London; Louvre, Paris; Museum Boymans-van Beuningen, Rotterdam). Dürer's teacher, Wohlgemuth, had also copied these works. The "Dying Orpheus" of one of the tarocchi engravers inspired Dürer to a free re-creation (drawing of 1494 in Kunsthalle, Hamburg, etching Bartsch 73). While in Venice and perhaps also before he went to Italy, Dürer saw engravings by masters from central Italy. Of these artists, the Florentine Antonio Pollaiuolo influenced the young Dürer most. Pollaiuolo, with his sinuous, energetic line studies of the movement of the human body, inspired

a generation of Florentine artists to make significant advances in this area, and Dürer also learned much from him. He copied several of Pollaiuolo's works, including his "Rape of the Sabine Women" in a drawing of 1495 (W 82; Musée Leon Bonnat, Bayonne, France).

Early paintings in the Italian spirit

Dürer's secular, allegorical, and frequently self-enamoured paintings of this period are often either adaptations of Italian models or entirely independent creations that breathe the free spirit of the new age. Dürer adapted the figure of Hercules from Pollaiuolo's "The Rape of Deianira" (Yale University Art Gallery, New Haven, Connecticut) for a painting of "Hercules and the Birds of Stymphalis" (Germanisches Nationalmuseum, Nürnberg). A purely mythological painting in the Renaissance tradition, its subject makes it exceptional among Dürer's pictures. The centre panel from the "Dresden Altarpiece," which Dürer painted perhaps around 1498, is stylistically similar to the "Hercules" and betrays influences of Mantegna. In most of Dürer's free adaptations the additional influence of the more lyrical, older painter Giovanni Bellini (*c.* 1430–1516), with whom Dürer had become acquainted in Venice, can be seen.

The most striking painting illustrating Dürer's growth toward the Renaissance spirit is a self-portrait, painted in 1498 (Prado, Madrid). Here Dürer sought to convey, in the representation of his own person, the aristocratic ideal of the Renaissance. He liked the way he looked as a handsome, fashionably attired young man, confronting life rather conceitedly. In place of the conventional, neutral, monochromatic background, he depicts an interior, with a window opening on the right. Through the window can be seen a tiny landscape of mountains and a distant sea, a detail that is distinctly reminiscent of contemporary Venetian and Florentine paintings. The focus on the man in the interior distinguishes his world from the vast perspective of the distant scene, another world to which the artist feels himself linked.

Italian influences in Dürer's graphics

Italian influences were slower to take hold in Dürer's graphics than in his drawings and paintings. Strong late Gothic elements dominate the visionary woodcuts of his *Apocalypse* series (the Revelation of St. John), published in 1498; all the woodcuts display emphatic expression, rich emotion, and crowded, frequently overcrowded, composition. The same tradition influences the earliest woodcuts of Dürer's *Great Passion* series, also from about 1498. Nevertheless, the fact that Dürer was adopting a more modern conception, a conception inspired by Classicism and humanism, during the long period in which these series were taking form, is indicative of his basically Italian orientation. The woodcuts "Samson and the Lion" (B 2; *c.* 1497) and "Hercules Conquering Cacus" (B 127), and many prints from the woodcut series *The Life of the Virgin* (especially B 77, 79–83, 88, 91, 93–95; *c.* 1500–10) have a distinct Italian flavour. Many of Dürer's copper engravings are in the same Italian mode. Certain elements of late Gothic tradition do occasionally appear in these engravings, but this may well be because their reproduction was addressed to a broad audience that was, as yet, mistrustful of new ideas in art. Some examples that may be cited are "Fortune" (B 78; *c.* 1496), "The Four Witches" (B 75; 1497), "The Temptation of the Idler" ("The Dream of the Doctor"; B 76; *c.* 1497–98), "The Sea Monster" (B 71; *c.* 1498), "Hercules" (B 73; *c.* 1498), "Adam and Eve" (B 1; 1504), "Apollo and Diana" (B 68; *c.* 1505), "Musical Satyr and Nymph with Baby" (B 69; 1505), and "The Small Horse" (B 96; 1505), and "The Large Horse" (B 97; 1505). Dürer's influence in the field of graphics later influenced the art of the Italian Renaissance that had originally inspired his creative efforts.

Consolidation of his artistic aims. Dürer's style continued to vacillate between Gothic and Italian Renaissance art until about 1500. Then his restless striving finally found definite direction. He seems clearly to be on firm ground in the penetrating half-length portraits of Oswolt Krel (Alte Pinakothek, Munich), in the portraits of three members of the aristocratic Tucher family of Nürnberg (Kunstsammlungen, Weimar, West Germany, and Staatliche Kunstsammlungen, Kassel, West Germany)—all dated 1499—and in the "Portrait of a Young Man" of 1500 (Alte Pinakothek, Munich). In the same year Dürer painted another self-portrait (Alte Pinakothek, Munich), a flattering Christ-like portrayal.

Period of stylistic consolidation

During this period of consolidation in Dürer's style, the Italian elements of his art were strengthened by his contact with Jacopo de' Barbari, a minor Venetian painter and graphic artist who was seeking a geometric solution to the rendering of human proportions and whose acquaintance Dürer had made in Venice in 1494–95. Later, Dürer encountered him often while de' Barbari was in the service of the Holy Roman emperor Maximilian I in Nürnberg in 1500 and during the Venetian's stay in Wittenberg between 1503 and 1505; it is perhaps due to his influence that Dürer began, around 1500, to grapple with the problem of human proportions in true Renaissance fashion. Initially, the most concentrated result of his efforts was the great engraving "Adam and Eve" (B 1; 1504) in which he sought to bring the mystery of human beauty to an intellectually calculated ideal form. In this he followed not only Jacopo de' Barbari but also the greatest of the Italian artist-theoreticians, particularly the 15th-century Umbrian painter Piero della Francesca and the Florentine Leonardo da Vinci. In all aspects Dürer's art was becoming strongly classical. One of his most significant classical endeavours is his painting "Altar of the Three Kings" (1504), which was executed with the help of pupils. The Italian character of this altar composition is immediately evident because it is made up of five separate pictures. Dürer's intellect and imagination went beyond direct dependence on Italian art, however, for he had by this time freely grasped and internalized the whole of it. From this maturity of style comes the bold, natural, relaxed conception of the centre panel, "The Adoration of the Magi" (Uffizi, Florence), and the ingenious and unconventional realism of the side panels, one of which depicts the "Drummer and Piper" and the other "Job and His Wife" (Wallraf-Richartz-Museum, Cologne).

Second journey to Italy. In the autumn of 1505, Dürer made a second journey to Italy, where he remained until the winter of 1507. Once again he spent most of his time in Venice. Of the Venetian artists, Dürer now most admired Giovanni Bellini, the leading master of Venetian early Renaissance painting, who, in his later works, completed the transition to the High Renaissance. Dürer attested to this admiration in a letter to Pirkheimer of November 7, 1506; and his pictures of men and women from this Venetian period reflect the sweet, soft portrait types especially favoured by Bellini. One of Dürer's most impressive small paintings of this period, a compressed half-length composition of the "Young Jesus with the Doctors" of 1506, harks back to Bellini's free adaptation of Mantegna's "Presentation in the Temple" (Staatliche Museen Preussischer Kulturbesitz, Berlin). Dürer's work is a virtuoso performance that shows mastery and close attention to detail. That Dürer himself so considered it is shown by the inscription on the scrap of paper out of the book held by the old man in the foreground, which reads, *Opus quinque dierum* ("the work of five days"). Dürer thus must have executed this painstaking display of artistry, which required detailed drawings, in no more than five days. Of greater artistic merit than this quickly executed work, however, are the half-length portraits of young men and women painted between 1505 and 1507, which seem to be entirely in the style of Bellini. In these paintings there is a flexibility of the subject, combined with a warmth and liveliness of expression and a genuinely artistic technique, that Dürer otherwise only rarely attained. The paintings include a half-length "Portrait of a Venetian Woman" of 1505 (Kunsthistorisches Museum, Vienna), a "Portrait of a Young Woman," tentatively dated around 1506 (Staatliche Museen Preussischer Kulturbesitz, Berlin), a half-length portrait of a cheerful young girl with a fashionable little hat, dated 1507, in the same museum, and that of a spirited young man, also done in 1507, on the reverse side of which is a dreadful allegorical representation of "Avarice" (Kunsthistorisches Museum, Vienna). In 1506, in Venice, Dürer also completed his great altarpiece "The Feast of the Rose Gar-

Admiration for Bellini

lands" for the funeral chapel of the Germans in the church of S. Bartholomew, Venice. This is a bold extension of the "Madonna with Donors" type of picture, an altarpiece composition typically showing portraits of the financial benefactors of the church in panels appended to a central scene of the Virgin and Child; "The Feast of the Rose Garlands" is one panel showing, as well as the Virgin and Child, the recognizable portraits of over a dozen contemporaries, including such notables as Pope Julius II, the emperor Maximilian I, Cardinal Domenico Grimani, and a portrait of Dürer himself.

Sojourn in Bologna

Later in 1506 Dürer went to Bologna "for the sake of art, in a hidden perspective, which someone wants to teach me," ostensibly one Agostino dalle Prospettive. In the course of this brief trip, estimated to be eight or ten days, it is likely that Dürer also took the opportunity to familiarize himself with the art circle of the nearby city of Emilia-Romagna, which was largely independent of Venetian influence. There the works of older, more austere masters like Francesco del Cossa (1436–78) and Ercole de' Roberti (*c.* 1456–96) probably influenced him more strongly than those of the sentimental younger painters Francesco Francia (1450–1517/18) and Lorenzo Costa (died 1535). He later returned to Venice for a final three months, where, as he complained, he was "the object of much envy." He also met with great favour, however, particularly from the aged Bellini, of whom he wrote, "[Bellini] has praised me before many noblemen. He wants something of mine very much and has himself come to me and asked that I make him something, and wants to pay well." The extent to which Dürer viewed Italy as his artistic and personal home is revealed by the frequently quoted words in his last letter from Venice (dated October 1506) to Pirkheimer, anticipating his imminent return to Germany: "O, how cold I will be away from the sun; here I am a gentleman, at home a parasite."

Development after the second Italian trip. By February 1507 at the latest, Dürer was back in Nürnberg, where two years later he acquired a fairly impressive house across from the gateway of the zoological gardens. (The "Dürer-Haus" is still standing today, an old-fashioned, four-story house with a high sloping roof.) It is clear that the artistic impressions gained from his Italian trips continued to influence Dürer to employ classical principles in creating largely original compositions. Among the paintings belonging to the period after his second return from Italy, the "Martyrdom of the Ten Thousand," painted in 1508 for Frederick the Wise, elector of Saxony, and the "Adoration of the Trinity," painted for M. Landauer for the Zwölfbrüderhauses in Nürnberg, a painting showing tiers of saints and angels in Italian High Renaissance fashion (1511), are both crowd scenes whose failure is in varying ways due to the intractability of the design. The problem of crowd composition is solved more successfully in the "Assumption of the Virgin" (known only through a copy) from "The Heller Altarpiece," which was commissioned by Jakob Heller for the Dominikanerkirche, in Frankfurt am Main (1508–09). The drawings for this picture recall Mantegna and betray Dürer's striving for classical perfection of form through sweeping lines of firmly modelled and simple drapery. Even greater simplicity and grandeur characterize the diptych of "Adam and Eve" (1507; Prado, Madrid, with a variant in Uffizi, Florence), in which the two figures stand calmly in relaxed classical poses against dark, almost bare, backgrounds.

The "Passion" series

Between 1507 and 1513 Dürer completed a "Passion" series in copperplate engravings and, between 1509 and 1511, the *Small Passion* in woodcuts, both of which are characterized by their tendency toward spaciousness and serenity. Dürer's four "Passion" series, one in copper engraving, two in woodcut, and the so-called "Green Passion" (Albertina, Vienna), which consists of drawings on greenish-tinted paper, are small graphic "stations of the cross" for private home use. In 1511, the book editions of *The Great Passion*, the *Small Passion*, and *The Life of the Virgin* appeared, as did the second edition of the *Apocalypse*, to which had been added a title illustration. During 1513 and 1514 Dürer created the greatest of his copperplate engravings: the "Knight, Death and Devil," "St. Jerome in His Study," and "Melencolia I,"—all of approximately the same size, varying from 24.2 by 19.1 centimetres (9.5 by 7.5 inches) to 24.8 by 19.1 centimetres (9.8 by 7.5 inches). The extensive, complex, and often contradictory literature concerning these three engravings deals largely with their enigmatic, allusive, iconographic details. Although repeatedly contested, it probably must be accepted that the engravings were intended to be interpreted together. There is general agreement, however, that Dürer, in these three so-called master engravings, wished to raise his artistic intensity to the highest level, which he succeeded in doing. Finished form and richness of conception and mood merge into a whole of classical perfection. To the same period belongs Dürer's most expressive portrait drawing—that of his mother (Staatliche Museen Preussischer Kulturbesitz, Kupferstichkabinett, Berlin).

Service to Maximilian I. While in Nürnberg, from February 4 to April 21, 1512, the emperor Maximilian I enlisted Dürer—by now recognized as the greatest painter of the city—into his service, and Dürer continued to work mainly for the Emperor until 1519. He collaborated with several of the greatest German artists of the day on a set of marginal drawings for the Emperor's prayer book (Staatliche Museen Preussischer Kulturbesitz, Kupferstichkabinett; Staatliche Graphische Sammlung, Munich; Bibliothèque Municipale, Besançon). He also did a number of etchings in iron (between 1515 and 1518) that proclaim his complete mastery of the medium and his freedom of imagination. In contrast to these pleasing improvisations are the monumental woodcuts, overloaded with panegyrics, made for Maximilian; "The Emperor Maximilian I's Triumphal Arch" (1515; overall size 3.409 by 2.922 metres [11.184 by 9.587 feet]), a complicated Roman triumphal arch with three passageways and a dome atop the centre passage, done in the Northern Renaissance style; and "The Emperor Maximilian I's Triumphal Chariot" (1522; total length 2.318 metres [7.605 feet]). In these somewhat stupendous, ornate woodcuts, Dürer had to strain to adapt his creative imagination to his client's mentality, which was foreign to him.

Woodcuts for Maximilian

Besides a number of formal show pieces—a painting entitled "Lucretia" (1518; Alte Pinakothek, Munich), and two portraits of the Emperor (1519; Germanisches Nationalmuseum, Nürnberg, and Kunsthistorisches Museum, Vienna)—during this decade Dürer also did a number of more informal paintings of considerably greater charm. He also travelled. In the fall of 1517 he stayed in Bamberg. In the summer of 1518 he went to Augsburg in the company of two friends who were to represent Nürnberg in the Imperial Diet, or German legislative assembly; the Diet was to deliberate on the actions of Martin Luther, who had in the previous year posted his 95 Theses denouncing the sale of papal indulgences on the door of the Wittenberg Castle Church. Dürer later became a devoted follower of Luther. In May or June of the following year he spent some time in Switzerland with his friend Pirkheimer. Dürer had achieved an international reputation by 1515, when he exchanged works with the illustrious High Renaissance painter Raphael.

Final journey to the Netherlands. In July 1520 Dürer embarked with his wife on a journey through the Netherlands. He travelled through the countryside and visited Antwerp, Brussels, Aachen, and Mechelen. In Aachen, at the October 23 coronation of the emperor Charles V, successor to Maximilian I (who had died January 12, 1519), Dürer met the mystical and dramatic Matthias Grünewald, who stood second only to Dürer in contemporary German art, and presented him with several etchings. Dürer returned to Antwerp by way of Nijmegen and Cologne, remaining there until the summer of 1521. He had maintained close relations with the leaders of the Netherlands school of painting—Lucas van Leyden, Quentin Massys, Jan Provost, and Bernard van Orley. In December 1520 Dürer visited Zeeland and in April 1521 travelled to Bruges and Ghent, where he saw the works of the 15th-century Flemish masters Jan and

Meeting with Matthias Grünewald

Hubert van Eyck, Rogier van der Weyden, and Hugo van der Goes, as well as the Michelangelo Madonna. Dürer's sketchbook of the Netherlands journey contains immensely detailed and realistic drawings. While in the Netherlands, he also began numerous other sketches, most of which are executed with an extreme precision that betrays Dürer's increasing tendency toward overexactitude. Some paintings that were created either during the journey or about the same time seem spiritually akin to the Netherlands school; for example, the portrait of Anna Selbdritt (Metropolitan Museum of Art, New York), a half-length picture of St. Jerome (1521, Museu Nacional de Arte Antiga, Lisbon), and the small portrait of Bernhard von Resten, previously Bernard van Orley (Gemäldegalerie, Dresden).

Final works. On July 12, 1521, the travellers were back in Nürnberg and, from this time on, Dürer's health began to decline. He devoted his remaining years mostly to theoretical and scientific writings and illustrations, although several well-known character portraits and some important portrait engravings and woodcuts also date from this period. One of Dürer's greatest paintings, the so-called "Four Apostles" (St. John, St. Peter, St. Paul, and St. Mark), was done in 1526. With this work Dürer reached his final and certainly his highest level of achievement as a painter. His constant delight in his own virtuosity no longer stifled the ideal of a spaciousness that is simple, yet deeply expressive.

Dürer's writings

Dürer's writings encompass widely varied subject matter. They consist of autobiographical writings, letters, poems, and three treatises: *Underweyssung der Messung, mit dem Zirckel un Richtscheyt, in Linien, ebnen unnd Gantzen Corporen, Durch Albrecht Dürer* (1525; a "Treatise on Mensuration with the Compasses and Ruler in Lines, Planes, and Whole Bodies"), which was intended as a guide for artists; *Herinn sind Begriffen vier Bucher von Menschlicher Proportion Durch Albrecht Dürer* (1525; published posthumously in Nürnberg in 1528; "Treatise on Human Proportions"); and a treatise on fortification (1527), which appeared under the title *Etliche Underricht, zu Befestigung der Stett, Schloss, und Flecken* ("Instruction on the Fortification of Cities, Castles, and Towns"). The woodcut illustrations for the treatise on mensuration are typical of Dürer's final style, while in those for the treatise on proportion the classical aesthetic ideal takes on mannerist features. The illustrations for the treatise on fortification are distinguished by a naïve realism that detracts from their artistic merit. The language in the treatises and in his personal writings is simple, powerful, and picturesque, less abstract than might be anticipated in view of Dürer's frequently theoretical approach to art.

Influence and significance. On April 6, 1528, Dürer died and was buried in the churchyard of Johanniskirchhof in Nürnberg. That he was one of his country's most influential artists is manifest in the impressive number of pupils and imitators that he had. Even Dutch and Italian artists did not disdain to imitate Dürer's graphics occasionally; probably the most gifted of his Dutch and Italian imitators was Jacopo Pontormo (1494–1557), one of the creators of Florentine Mannerism. In the period around 1600, when artists frequently sought models in the art works of Dürer's time, there was a virtual "Dürer Renaissance."

The extent to which Dürer was internationally celebrated is apparent in the literary testimony of the Florentine artist Giorgio Vasari (1511–74), in whose *Lives of the Most Eminent Italian Architects, Painters and Sculptors*, the importance of Albrecht Dürer, the "truly great painter and creator of the most beautiful copper engravings," is repeatedly stressed. Dürer himself recorded with satisfaction the numerous honours accorded to him, in his letters from Venice and especially in his diary of the journey to the Netherlands (1520–21):

> And on Sunday, . . . the painters invited me, along with my wife and servant, to their chambers, where they had all sorts of things, a silver service, other costly appurtenances, and overly expensive food . . . And when I was ushered to the table, the people stood off on both sides, as one behaves toward a great gentleman. There were also among them very fine people of rank, who all presented themselves to me with deep bows, in the most gracious way.

Like most famous Italian artists, Dürer probably felt himself to be an "artist-prince," and his self-portraits seem incontestably to show a man sure of his own genius.

MAJOR WORKS

Paintings

RELIGIOUS PAINTINGS: "Mater Dolorosa" (polyptych; 1496–97; Alte Pinakothek, Munich); "Dresden Altarpiece," (commissioned, 1496; Gemäldegalerie, Dresden, East Germany); "Lamentation of Christ" (*c.* 1500; Alte Pinakothek, Munich); "Salvator Mundi" (1502; Metropolitan Museum of Art, New York); "The Paumgärtner Altarpiece" (1502–04; Alte Pinakothek, Munich); "The Adoration of the Magi" (1504; Uffizi, Florence); "The Ober St. Veit Altarpiece" (1505–06; Ober St. Veit, Archbishop's Palace, Austria); "The Feast of the Rose Garlands" (1506; National Museum, Prague); "Young Jesus with the Doctors" (1506; Thyssen Bornemisza Collection, Castagnola, Switzerland); "Adam and Eve" (1507; Prado, Madrid); "Martyrdom of the Ten Thousand" (1508; Kunsthistorisches Museum, Vienna); "Adoration of the Trinity" (1511; Kunsthistorisches Museum, Vienna); "St. Jerome" (1521; Museu Nacional de Arte Antiga, Lisbon); "Four Apostles" (1526; Alte Pinakothek, Munich).

PORTRAITS: "Self-Portrait" (1493; Louvre, Paris); "Frederick the Wise" (1496; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Self-Portrait" (1498; Prado, Madrid); "Self-Portrait" (1500; Alte Pinakothek, Munich); "Portrait of a Young Man" (1500; Alte Pinakothek, Munich); "Portrait of a Young Man" (1506; Hampton Court, Middlesex); "Portrait of Charlemagne" (1512–13; Germanisches Nationalmuseum, Nürnberg); "Lorenz Sterck" (1521; Isabella Stewart Gardner Museum, Boston).

Graphic works

ENGRAVINGS: "Adam and Eve" (1504); "Agony in the Garden" (1508); "Man of Sorrows" (1509); *Small Passion* (series, 1509–11); "Ecce Homo" (1512); "Knight, Death and Devil" (1513); "St. Jerome in His Study" (1513); "Melencholia I" (1514); "Erasmus of Rotterdam" (1526).

WOODCUT CYCLE: *Apocalypse* (1498; title page, 1511).

BIBLIOGRAPHY. The Bartsch numbers given throughout the article refer to ADAM VON BARTSCH, *Le Peintre graveur*, 21 vol. (1803–21) and the Winckler numbers to FRIEDRICH WINCKLER, *Die Zeichnungen Albrecht Dürers*, 4 vol. (1936–39). See also the bibliography of M.J. FRIEDLANDER's article on Dürer in the Thieme-Becker, *Allgemeines Lexikon der bildenden Künstler*, vol. 10, pp. 63–70 (1914). The most important publications since then are: ERWIN PANOFSKY, *Dürers Kunsttheorie* (1915); GUSTAV PAULI, "Die Dürer-Literatur der letzten drei Jahre," *Repertorium für Kunstwissenschaft*, 41:1–34 (1919); M.J. FRIEDLANDER, *Albrecht Dürer* (1921), in German; ERWIN PANOFSKY, *Dürers Stellung zur Antike* (1922); CAMPBELL DODGSON, *Albrecht Dürer* (1926); PIERRE DU COLOMBIER, *Albrecht Dürer* (1927), in French; HANS TIETZE and ERICA TIETZE-CONRAT, *Kritisches Verzeichnis der Werke Albrecht Dürers*, 2 vol. (1928–38); JOSEPH MEDER, *Dürer-Katalog* (1932); EMIL WALDMANN, *Albrecht Dürer* (1933), in German; *Drawings and Water-Colours*, selected and with an introduction by E. SCHILLING (1949); FRIEDRICH WINKLER, *Dürer und die Illustrationen zum Narrenschiff* (1951); H.T. MUSPER, *Albrecht Dürer* (1952), in German; *Watercolours by Albrecht Dürer*, selected and with an introduction by A.M. CETTO (1954); ERWIN PANOFSKY, *The Life and Art of Albrecht Dürer*, 4th ed. (1955); FRIEDRICH WINKLER, *Albrecht Dürer, Leben und Werk* (1957); and the catalog of the exhibition "Albrecht Dürer 1474–1971" in the Germanisches Nationalmuseum, Nürnberg-Munich (1971).

(Eb.R.)

Duricrusts

The term duricrust (Latin *durus*, "hard") was first applied in Australia to layered materials at or near the Earth's surface, such as laterites, bauxites, and quartzites. These crusts are not of themselves landforms but represent the chemical alteration of the upper parts of plains and other features of low relief. In a sense, they are soils (*q.v.*) of an extreme type. In general, duricrust connotes a surface or near-surface, hardened accumulation of silica (SiO_2), alumina (Al_2O_3), and iron oxide (Fe_2O_3), in varying proportions. Admixtures of other substances commonly are present and duricrusts may be enriched with oxides of manganese or titanium within restricted areas. Thus, siliceous, ferruginous, and aluminous crusts constitute duricrusts proper but the term also is applied, by extension,

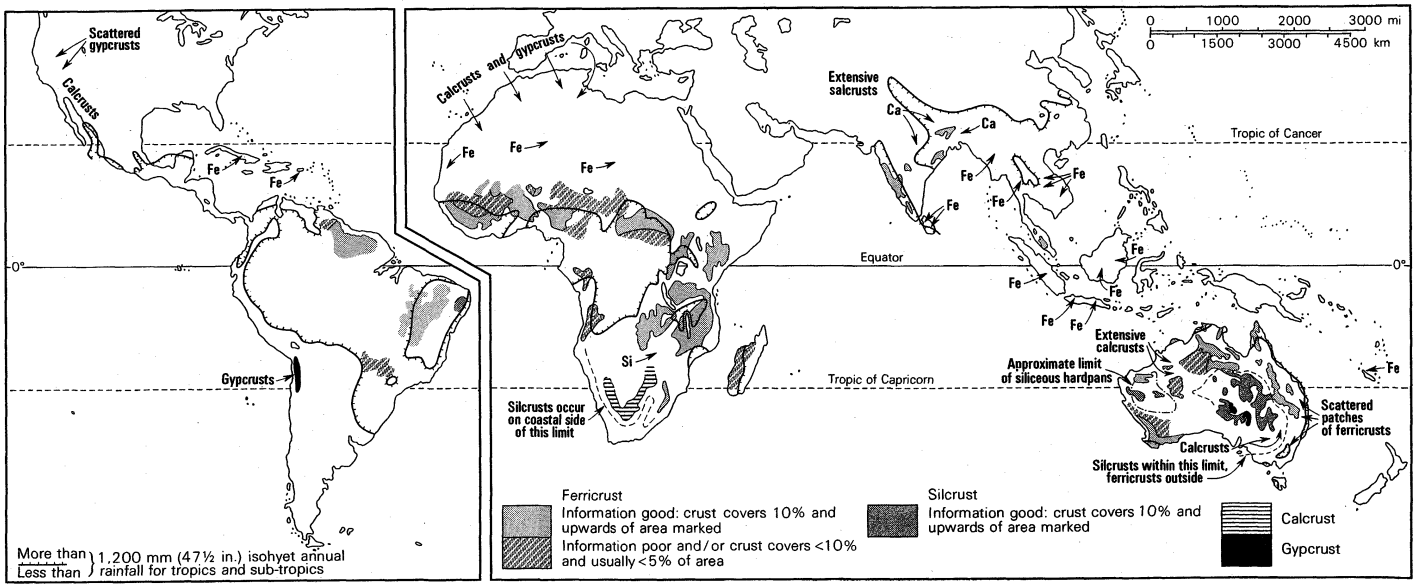


Figure 1: World distribution of duricrusts.

to encrusted layers of calcium carbonate, gypsum, and salt.

Classification and general distribution

Two partial classifications use compound names ending in -crete, to indicate the kind of cementation, or in -crust, to indicate the basic chemical content (Table 1, first two columns). Both classifications are defective, although the working distinction between silcrusts and ferricrusts is useful. A more serviceable classification (Table 1, third column) adapts and extends the nomenclature developed by soil scientists in Africa. The type boundaries that fall within duricrusts proper must be considered transitional.

Where best preserved, duricrusts proper are the most continuous and extensive types of those listed in Table 1. Representing the end-products of weathering (*q.v.*), denudation, and soil formation, they occur mainly on erosional platforms such as pediments (*q.v.*) or as cappings and residuals on stream divides. The crusts usually form parts of deep-weathering profiles that may be as thick as 120 metres (400 feet). Alternatively, they occur at the bases of cliffs and scarps, in river terraces, or on valley bottoms, usually near to and lower than residual cappings. Except at the wasting or developing edges of crusts, the thickness ranges from about 0.5 m (1.5 ft) to at least 12 m (40 ft). This contrasts with the platelike weathering rinds as thick as 15 centimetres (6 inches) that are often associated with cavernous (alveolar) weathering, particularly in arid areas.

Duricrusts are concentrated in intertropical to subtropical areas, with notable extratropical extensions, especially in South Africa and Australia (Figure 1). They normally are absent from equatorial rain forests. Many are fossil crusts, in the sense that they relate to past climatic, biologic, and geomorphic environments and are not forming under present conditions in these areas.

Other types of crust are associated with subhumid to arid climates, although not necessarily with the arid climatic zones of today. Crusts of calcium carbonate (calccrusts) and calcium sulfate (gypcrusts), up to 4 m (about 13 ft) thick, occur in basins of inland drainage, where they form initially as evaporites (*q.v.*). Alternatively, calccrusts form as surface to subsurface soil horizons, or zones, at and near the extreme end of the calcium-rich soil range. Gypsum-rich horizons, common in many desert and semidesert soils, seem not resistant enough to erosion to become crusts. Calcsilitic crusts, which result from the silification of calccrusts or other surficial limestones, have been little studied in this context. Salcrusts (salt crusts) form in depressions along desert coasts or wherever saline groundwater emerges, but unless they crystallize into rock salt these crusts also lack resistance to erosion and can be ephemeral.

This article treats the physical and chemical characteristics of the several kinds of duricrusts, their present distribution, and the factors that are involved in their formation. For additional information on the latter topic, see SOILS; WEATHERING; GEOCHEMICAL EQUILIBRIA AT LOW TEMPERATURES AND PRESSURES; and CLIMATIC CHANGE. See also DESERTS; PEDIMENTS; and PLAYAS, PANS, AND SALINE FLATS for relevant information on duricrust environments.

Table 1: Classification of Duricrusts

classification by			chemistry (not exclusive)
cement	content (i)	content (ii)	
	types of crust		
Duricrusts proper			
Silcrete	silcrust	silitic	SiO_2
		siallitic	SiO_2, $Al_2O_3/Al_2O_3 \cdot 2H_2O$
Silcrete/ferricrete	silcrust/ferricrust	fersilitic	Fe_2O_3, SiO_2
Ferricrete	ferricrust	fersiallitic*	$Fe_2O_3 \pm FeOOH$, SiO_2, $Al_2O_3 \cdot 2H_2O \pm AlOOH$
		ferrallitic	Fe_2O_3, $FeOOH$, $Al_2O_3 \cdot 2H_2O$, $AlOOH$
		ferritic	Fe_2O_3, $FeOOH$
		fermangitic	Fe_2O_3, MnO_2
		tiallitic	TiO_2, $Al_2O_3/Al_2O_3 \cdot 2H_2O$
		allitic	$Al_2O_3 \cdot 2H_2O$, $AlOOH$
Calcrete	calccrust	calcitic	$CaCO_3$ (calcite)
		calcsilitic	$CaCO_3$, SiO_2 (calcite + chalcedonic silica)
Gypcrete	gypcrust	gypsitic	$CaSO_4 \cdot 2H_2O$ (gypsum)
Salcrete	salcrust	halitic	NaCl (usually impure; rock salt)

*Characteristic minerals in the fersiallitic range: SiO_2, quartz + chalcedonic silica; Al_2O_3, amorphous, to $Al_2O_3 \cdot 2H_2O$, gibbsite, and $AlOOH$, boehmite; Fe_2O_3, hematite, to $FeOOH$, geothite, and $Fe_2O_3 \cdot 2H_2O$, limonite (where unhydrated); TiO_2, rutile/anatase; MnO_2, pyrolusite/psilomelane.

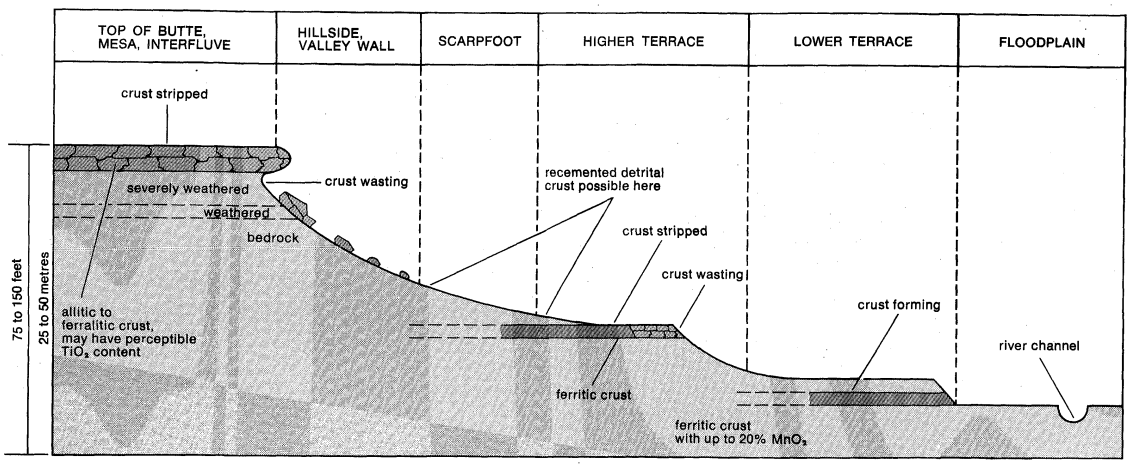


Figure 2: Relationships of duricrusts and weathering profiles in West Africa.

PHYSICAL CHARACTERISTICS AND CHEMICAL COMPOSITION

Ferricrusts and silcrusts. The size, shape, and arrangement of materials in ferricrusts (iron-rich crusts) and silcrusts (silica-rich crusts) ranges from nodular, through slaggy, to blocky or massive. Bedrock structures can be entirely destroyed or, as in some silcrusts, perfectly preserved. Silt-sized material is rare to absent. Duricrusts observed in the field exhibit a range of colours, but silica-rich crusts often are yellowish or yellowish-red, the several iron-rich crusts (Table 1) are more definitely reddish or reddish-brown, and allitic crusts (alumina-rich) are usually iron-stained and exhibit reddish hues on their exterior surfaces.

The name laterite (properly, lateritic crust) is widely applied to ferricrusts and even to silcrusts. This usage is unfortunate in the opinion of some, but it dates back to 1807, when F. Buchanan described a zone below the upper part of a deep-weathering profile (thick zone of weathered material above bedrock) as laterite. Ferricrusts result from the hardening *in situ*, however caused, of subsurface oxides of iron and aluminum; this material is termed plinthite in the soils classification used by the U.S. Department of Agriculture. The hydrated forms of these oxides, the minerals goethite and gibbsite, are often present, but others, such as limonite and boehmite, are less common or less abundant (see Table 1). Silica content, which consists of residual quartz and combined silica, usually is less than 20 percent and can fall below 1 percent (Table 2). Fe_2O_3 and Al_2O_3 content, after ignition loss, can exceed 80 percent; combined water is usually between 5 percent and 25 percent. Local enrichment with MnO_2 or TiO_2 can amount to 20 percent. Interstices in ferricrusts, if not barren, contain mainly kaolinitic to halloysitic clays (see CLAY MINERALS).

Table 2: Composition of Average Rocks and of Selected Duricrusts*
(percentage)

	SiO_2	Al_2O_3	Fe_2O_3
Average granite	70	14.5	1.5
Average basalt	50	16	5.5
Average sandstone	78.5	5	1
Average shale	60	15.5	4
Ferritic crust, Guinea	3	11	84.5
Allitic crust, Bihar, India	2.5	79	2.5
Silitic duricrust and deep-weathering profile, Queensland, Australia			
crust	96	2	1
mottled zone	88.5	10	1
pallid zone	76	21	2

*Values after ignition, somewhat rounded.

Figure 2 illustrates characteristic interrelationships of allitic (aluminum-rich), ferritic (iron-rich), and fermangitic (iron-manganese) crusts in West Africa. The deep-weathering profile beneath the highly leached, residual allitic cap consists of kaolinitic clay, at least partly blotched and reddened, often with a sandy horizon just above bedrock. Crusts at lower levels (so-called detrital and groundwater laterites) form from detritus and solutes supplied mainly by the eroding cap. In eastern Africa and in Australia the profiles are strongly differentiated into a light-coloured kaolinitic zone at the bottom, a mottled zone in the middle, and a so-called indurated (hardened) zone at the top. This subdivision is imprecise, however, because the mottled zone can be partly or wholly indurated.

Weathering profiles in West Africa and Australia

Figure 3 summarizes the transition from ferricrusts to silcrusts in the eastern half of Australia. Silcrusts contain as much as 99 percent silica in the form of sand and pebbles, and chalcedony and other kinds of silica. The crusts weather into bouldery blocks and angular fragments and supply a common type of gibber material from their receding edges. Thin hematite skins on quartz grains in silcrust can produce misleading red colours. Deep-weathering profiles topped by silcrust may be highly siliceous throughout; Al_2O_3 content rises only as the light-coloured zone becomes increasingly kaolinized (Table 2).

In Australia, India, Africa, and South America, the main expanses of duricrust mantle pediments and plains in varying states of dissection, although some crusts occur in valleys in terrain of high relief. Allitic crusts yield commercial bauxite. Detrital and valley-floor duricrusts occur in all these countries, chiefly adjacent to the margins of residual caps. These crusts include economic reserves of manganese ore in western Africa and silicified terrace gravels in southern Australia. Possible combinations of terrain, weathering, erosion and dissection of duricrusts and continued or renewed duricrust formation are highly complex. Additionally, some duricrusts now lie buried beneath continental (nonmarine) sediments.

Limits of modern formation

Rough limits to present-day ferricrust formation are the 500- to 700-millimetre (20- to 27.5-inch) isohyet (contour of equal rainfall values), below which iron is not readily mobilized, and the 1,200-mm (47.5-in.) isohyet (shown in Figure 1), above which dehydration is unusual. High mean annual temperatures, on the order of 20° to 25° C (68° to 77° F), also are necessary. Duricrusts that occur beyond the indicated limits are generally fossil (related to former climatic regimes), and many within these limits also are fossil. Ages determined by stratigraphic or radiometric methods are as great as 50,000,000 years in western Africa and more than 23,000,000 years in Australia, but duricrust formation is occurring today in some places. Phenomena related to fossil crusts include the Tertiary deep weathering of the southern Piedmont area of the United States, of massifs in western and northwestern Europe, and the Tertiary formation of residual bauxite at latitude 65° N in Siberia.

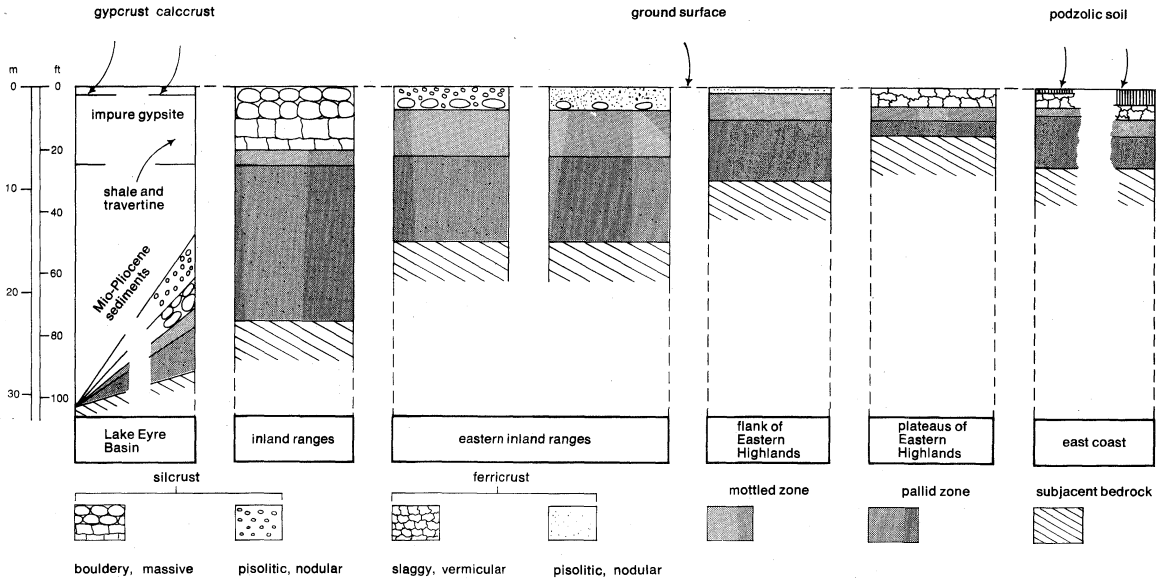


Figure 3: Structures of duricrusts in vertical sections in eastern Australia.

Calccrusts and gypcrusts. Calccrusts formed in basins of inland drainage are largely calcitic, and calcite content often exceeds 85 percent. They frequently display a travertine-like structure but range in origin from shoreline deposits as with the Lake Bonneville tufas, to depression-floor sheets, as in western and central Australia. Calccrusts formed as soil horizons are especially well known from the southern High Plains and southwestern U.S. Varying in structure from massive and well cemented to loose and friable, and ranging in thickness to about 2 m (7 ft), they are most resistant when they contain appreciable silica.

Gypcrusts (gypsum-rich crusts) typically contain more than 95 percent gypsum and vary from massive crystalline gypcrete to loose, powdery deposits. Initially an evaporite, loose gypsum may be transported by the wind and redeposited to encrust sand dunes and rockcut pediments, as in northern Africa.

Calccrusts and gypcrusts are widely scattered through the world's deserts, where some record a former pluvial period (moist period of the Pleistocene Epoch); Figure 1 can only indicate selected occurrences. Salcrusts (salt crusts) in flat-bottomed depressions typify the sabkha (saline flats) of northern Africa and Arabia and corresponding features in other countries. Parts of the Punjab are affected by saline crusting on low-angled alluvial fans.

FACTORS INVOLVED IN THE FORMATION OF DURICRUSTS

Leaching and end products of weathering. The formation of ferricrust requires the selective removal of the mobile (easily leached) soil constituents sodium, potassium, calcium, and usually magnesium, and also the removal of silica other than residual quartz. Conversely, Fe_2O_3, Al_2O_3 or its hydrates, and sometimes TiO_2 become concentrated. In addition, the oxide-rich soil horizon must eventually dry out and the overlying material be stripped off to reveal the indurated crust.

Selective removal of soil constituents

Mobilization and leaching of Na, K, Ca, and Mg offers no problem, and nonquartzose silica is recognized as more mobile than TiO_2, Fe_2O_3, or Al_2O_3. Although the rate of silica removal can be matched in other humid areas, silica in solution provides about half the dissolved load of tropical rivers generally. For some catchments, a denudation rate of 4 to 5 cm (1.6–2 in.) per 1,000 years is indicated. Silica solubility rises with increasing alkalinity or pH; the mineral can readily be transported as a hydrous oxide gel or in allophane gel (an amorphous or unstructured state) and forms a significant bulk fraction of many tropical plants. In acid environments (pH less than 7) silica undergoes imprecisely known processes to combine with aluminum oxides and hydroxides, and with hydroxyls to form kaolinitic clays, but severe and repeated wetting can reseparate these into Al_2O_3 and $Si(OH)_4$. Fe_2O_3 is insoluble except under highly acid conditions (below about pH = 4), and Al_2O_3 is soluble in the same range as well as at pH = 10. Varying combinations of relative mobility are therefore possible.

The formation of crusts involves great loss of weathered material. A generalized example from the tropical weathering of a nepheline–syenite (intrusive igneous rock) shows a reduction of silica (SiO_2) from 55 percent in the fresh rock to 5 percent in the duricrust, but an increase of alumina (Al_2O_3) from 1 percent to 45 percent, of iron oxide (Fe_2O_3) from 5 percent to 23 percent, and of combined water from 1 percent to 25 percent (Table 2).

The circulation of nutrients between plants and soil in tropical forests involves excess uptake by the plants, and this in turn promotes deep weathering. Within the deep-weathering profile, silt-size material is broken down or

By courtesy of George H. Dury

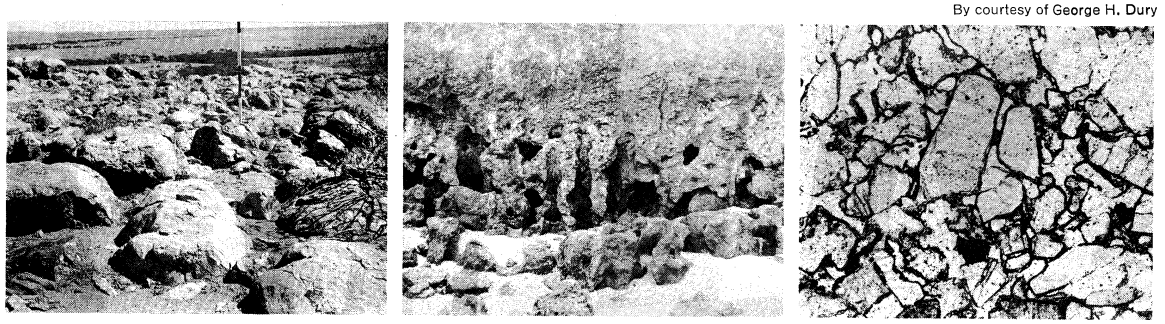


Figure 4: (Left) Bouldery silcrust on summit of residual; vertical divisions on picket are one foot (25 centimetres). (Centre) Coastal section of indurated plinthite, part of a ferricrust profile. (Right) Quartz and hematite skins; width of field is two millimetres.

leached out. Clay minerals tend to be dispersed and moved downward, especially where high rainfall and vigorous plant growth lower the electrolyte concentration. The remaining oxides tend to aggregate into forms in which spheroidal microstructures are common.

Mechanisms that are capable of promoting dehydration and hardening of ferricrusts, whether before, during, or after stripping of the overlying soil, include the destruction of forest and lowering of the water table, both of which can occur in several ways. Aside from clearance by man, for example, forest destruction may be caused by climatic change and downcutting by fluvial processes.

Silcrust formation requires the selective concentration of silica, a fact that has led some writers to consider silcrusts as the lower parts of ferricrust profiles. The distributional contrast between silcrusts and ferricrusts is clear, however, and the transition between the types is well documented (Figures 1 and 3). Silcrusts often, but not invariably, result from the silicification of sandstones and quartzitic conglomerates. They occur in areas that are currently drier than those with ferricrusts, but the fossil nature of many, plus the deep-weathering profiles to which they usually belong, presumably indicate humid climates at the time of formation and inhibit direct reference to existing controls. Like ferricrusts, silcrusts are usually taken to have originated below the ground surface, possibly under a layer of erodible, fine material.

Mobilization, migration, and concentration of ions.

Ion-exchange processes

Soil-formation processes of selective concentration of oxides of iron and aluminum, and in some circumstances of silica, include ion exchange as a most important factor. Although not yet completely understood, this involves the exchange of ions held by negative charges with other ions in the electrolyte (soil solution). Ion exchange is influenced by the fit of ions into a mineral structure. Relevant processes include hydration (adsorption of water), hydroxylation (adsorption of H^+ and OH^- ions), oxidation (combination of oxygen, with loss of electrons to weathering agents), and reduction (depletion of combined oxygen). Ion exchange is controlled by the cation exchange capacity (CEC) expressed as the amount of exchangeable cations in milliequivalents per 100 grams clay at pH 7. Low CEC values are typical of kaolinitic clays and of actual or potential duricrusts.

Soil water will separate into oppositely charged ions, H^+ and OH^-, and the CO_2 of the atmosphere and soil will yield HCO_3^- and free H^+ ions in solution. These products promote displacement of some metal cations, especially those in mineral silicates, largely by H^+ ions that combine with OH^- in removable solutes. The H^+ ions are small and highly charged in relation to their size and can readily enter many crystal lattices; OH^- ions neutralize the small charges of Na^+, K^+, and the larger charges of Ca^{++} and Mg^{++}. Positive charges in soil particles are partly related to hydrous oxides of iron, aluminum, and manganese. Negative charges, increasing with falling pH, are neutralized by positive ions, among which $Al(OH)_2^+$ is one of the more significant. Negatively charged colloidal SiO_2 and colloidal Al_2O_3 and positively charged Fe_2O_3 probably interact at high concentrations of H^+ ions to form clay minerals. Among these, the most stable are the one-to-one layer silicates of the kaolin family, in which each silicon–oxygen sheet is condensed with one aluminum hydroxide sheet.

Role of organisms and organic substances

At least part of the ion-exchange process involves organisms and organic substances. Chelating agents, complex amino acids, and allied compounds inactivate ions of aluminum and iron and hold them firmly in lattice structures. The ions then behave as if they were not present, except when acidity markedly decreases and they are redeposited. Manganese and silicon can be similarly treated. The combined processes of solution and eluviation (soluviation) and of chelation and eluviation (cheluviation) appear to act powerfully in the formation of oxide-rich plinthite prior to duricrust formation. In the mobilization and fixing of iron, as in the general production of organic acids, bacteria also play a part. Some act to form soluble iron, others oxidize soluble ferrous iron ($Fe(OH)_2$) to insoluble ferric iron (Fe_2O_3); and soil micro-organisms, including bacteria, are specifically involved in the production of a number of prominent chelating agents.

Rock type, terrain, and water table fluctuations. Duricrusts occur on a wide range of igneous, metamorphic, and sedimentary rocks, including granites, basalts and gabbros, arenites, and argillites. There is only the roughest of tendencies for duricrust chemistry to be controlled by bedrock chemistry, even in similar climates, although nepheline–syenites characteristically weather into allitic (aluminum-rich) crusts, basic igneous rocks into ferritic (iron-rich) to tiallitic (titanium–aluminum) crusts, and arenites and argillites, in some areas, into silitic (silica-rich) crusts. Ferritic crusts are more highly indurated, more variable in structure, and less strongly hydrated than aluminous crusts. Although structures in silitic crusts vary from pea-sized nodules to blocky and massive, with natural, subsurface erosion pipes at lower levels, these crusts are not hydrous.

Profile drainage is influential; ready leaching and alkaline to neutral conditions favour removal of silica and concentration of aluminum and also of titanium if available. Nearness to the water table promotes concentration of iron, whereas poor site drainage and acidity possibly favour accumulation of silica. As previously stated, however, known distributions suggest geographic contrasts between ferricrust and silcrust formation, rather than lithological control, which appears to be effective only in transitional belts.

Terrain requirements for duricrust formation include gentle slopes or situations where groundwater can supply oxides of iron and manganese or both of these. Well-preserved fossil crusts on pediments or plains with maximum slopes of 8° to 10° (and average slopes of 2° or less) suggest feeble lateral movement of groundwater and relative enrichment of crusts by leaching. This contrasts with the active translocation responsible for the absolute enrichment of crusts at the base of scarps and on valley floors. Also indicative of groundwater action are the light-coloured and mottled zones of many deep-weathering profiles; the former are regarded as the result of kaolinization in a reducing (de-ionizing) environment, and the latter from seasonal fluctuation of the groundwater level. Incapacity of these zones to supply the iron content of numerous crusts confirms relative enrichment.

Effects of climate and time. Calcrusts, gypcrusts, and salcrusts are referable to dry climates, but duricrusts proper, at least in present and late Holocene occurrences, are referable to humid tropical climates, probably with seasonal dryness, coincident wet and warm seasons, and soil temperatures in the average range 25° to 30° C (about 75° to 85° F). Under these conditions, 50 percent or more of the original rock volume can be lost during weathering, but the preservation of structures in some profiles indicates downward thickening rather than overall diminution.

A span of 30 to 50 years will convert a drying ferallitic clay to a ferallitic duricrust; but extrapolation from

By courtesy of George H. Dury

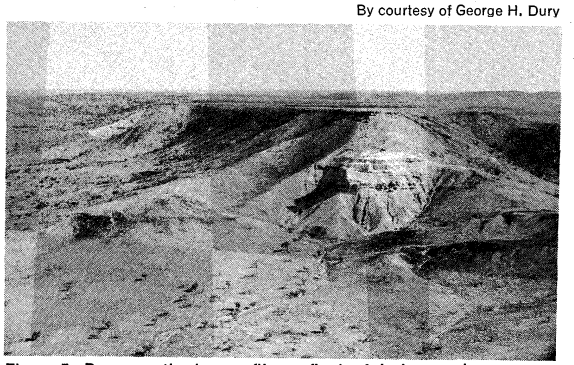


Figure 5: Deep-weathering profile on flank of duricrusted residual, New South Wales, Australia. Vertical distance between top of residual and base of pallid zone is about 130 feet (40 metres).

known values suggests that up to 15,000,000 years may be required to form really deep-weathering profiles. Such time spans seem to be well within the range of duration of humid tropical forests in the Tertiary, however.

Duricrusts as indicators of climatic change

Climatic change presumably is responsible for the presence of duricrusts in equatorial areas that now receive more than 1,200 mm (about 47 in.) mean annual precipitation. The former northward extension of aridity in Africa, with Kalahari sand extending 1,600–3,000 km (1,000–1,900 mi) beyond its present limit, is well documented. Similarly, former climates of the current humid tropical type are probably responsible for fossil crusts outside the tropics and for relict Tertiary deep weathering. Such climates seem explicable in terms of reduced pole-to-Equator temperature gradients. Increased temperatures are indicated by oxygen-isotope analysis, and increased humidities by floral records, including those of Tertiary soft coals. These climates are not explicable by continental drift (*q.v.*), the influence of which, in northward-moving Australia, they more than offset.

Although dehydration and hardening of duricrusts are often called irreversible, this is not true over the long term. Apart from disaggregation of eroding caps, residual ferricrusts can be attacked by renewed soil-formation processes, which remobilize iron and produce red-yellow soils called "lateritic podzolics" in older classifications.

Remaining problems. Further elucidation of ionic and micro-organic action in the formation of tropical weathering profiles are essentially soil-science problems. Geomorphic and stratigraphic problems relating to duricrusts include the refinement of the world distribution map (Figure 1) and the establishment of a well-based history of deep weathering and crust formation. Lower limits of the requisite temperature and precipitation need to be fixed, with particular reference to the reconstruction of Tertiary climates, for which it seems necessary to postulate a general pattern of atmospheric circulation unlike that of today. Exact analogues of duricrusts proper have still to be demonstrated in the geological column, although numerous writers relate the formation of red beds to processes like those currently responsible for red (hematitic) tropical weathering. Silicified arenites (silica-rich sandstones) of terrestrial origin are also known. The contrast between silcrust formation and ferricrust formation has yet to be explained in terms other than those of parent rock type. There also is room for study of the relationship of the silcrust formation associated with deep weathering and thorough silicification, which affects bedrock on steep slopes and high residuals above the general silcrusted pediplain at some Australian sites.

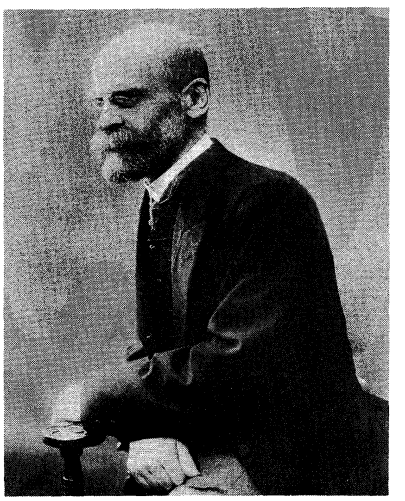
BIBLIOGRAPHY. G.W. LAMPLUGH, "Calcrete," *Geol. Mag.*, 9:575 (1902), "The Geology of the Zambezi Basin Around the Batoka Gorge, Rhodesia," *Q. Jl. Geol. Soc. Lond.*, 63: 162–216 (1907); and W.G. WOOLNOUGH, "The Duricrust of Australia," *J. Proc. R. Soc. N.S.W.*, 61:24–53 (1927), the above works introduce names ending in "-crete" and the term duricrust; G.H. DURY, "Rational Descriptive Classification of Duricrusts," *Earth Sci. J.*, 3:77–86 (1969), treats the taxonomy of duricrusts in terms of iron, aluminum, and silica content; R. MAIGNIEN, *Review of Research on Laterites* (1966), a general summary, especially useful for references to French and Belgian work on African pedology; J.A. PRESCOTT and R.L. PENDLETON, *Laterite and Lateritic Soils* (1952), a brief but effective review, especially on ferricrusts; A.M.J. DE SWARDT, "Lateritisation and Landscape Development in Parts of Equatorial Africa," *Z. Geomorph.*, 8:313–333 (1964), penetrating synthesis of general conclusions about ferricrusting; T. LANGFORD-SMITH and G.H. DURY, "Distribution, Character, and Attitude of the Duricrust in Northwest of New South Wales and Adjacent Areas of Queensland," *Am. J. Sci.*, 263:170–190 (1965); and J.A. MABBUTT, "The Weathered Land Surface in Central Australia," *Z. Geomorph.*, 9:82–114 (1965), two papers that discuss silcrusting and list numerous references; R. COQUE, *La Tunisie présaharienne* (1962), deals with gypsitic crusts, including wind-laid examples; UNITED STATES DEPARTMENT OF AGRICULTURE, *Soil Classification, Comprehensive System, 7th Approximation* (1960), a taxonomic work in advanced pedology; A.E.M. NAIRN (ed.), *Descriptive Palaeoclimatology* (1961), assembles varied evidence for paleoclimatic change, entries dealing with the Tertiary are especially relevant; N.M. STRAKHOV, *Principles of Lithogenesis*, vol. 1 (1967; orig. pub. in Russian, 1960–62), a partly paleoclimatic interpretation of the sedimentary record, conservative in its views of continental drift but highly informative on the occurrence of evaporites in the geological column.

(G.H.D.)

Durkheim, Émile

Émile Durkheim was the first social scientist to subject specific phenomena of everyday life to close sociological study and to formulate a vigorous methodology. His influence on his discipline has probably not been matched in the 20th century. A modest and a generous man, he paid tribute to the 18th-century French political philosopher Montesquieu as the precursor of his sociological approach to law and government studies, and to Henri de Saint-Simon and to Auguste Comte as his more direct predecessors in the investigation of society.

By courtesy of Presses Universitaires de France



Durkheim.

Durkheim was born in April of 1858, in the small town of Épinal, in the Vosges region of eastern France. His family was Jewish and of very modest means. It was taken for granted that he would study to become a rabbi, like his father. The death of his father before Durkheim was 20 burdened him with heavy responsibilities. That, and the atmosphere of France's eastern provinces bordering on Germany, during an intense period of national rivalry, may have contributed to making Durkheim a severely disciplined young man. As early as his late teens Durkheim became convinced that effort and even sorrow are more conducive to the spiritual progress of the individual than pleasure or joy. His outstanding success at school designated him clearly as a candidate to the renowned École Normale Supérieure in Paris—the most prestigious teachers' college in France. While preparing for the École Normale at the lycée Louis le Grand, Durkheim took his board at the Institution Jauffret in the Latin Quarter, where he became acquainted with another gifted young man from the provinces, Jean Jaurès, later to lead the French Socialist Party and at that time inclined like Durkheim toward philosophy and the moral and social reform of his countrymen. Durkheim passed the stiff competitive examination for the École Normale one year after Jaurès, in 1879. It is clear that his religious faith had vanished by then. His thought had become altogether secular but with a strong bent toward moral reform. Like a number of French philosophical minds during the Third Republic, he looked to science and in particular to social science and to profound educational reform as the means to eschew the perils of anarchistic individualism or "anomie," as he was to call this condition in which norms for conduct were either absent, weak, or conflicting.

Childhood and education

He enjoyed the intellectual atmosphere of the École Normale—the discussion of metaphysical and political issues pursued with eagerness and animated by the utopian dreams of young men destined to be among the

leaders of their country. He soon enjoyed the respect of his peers and of his teachers, but he was impatient with the excessive stress then laid in French higher education on elegant rhetoric and surface polish. His teachers of philosophy struck him as too fond of generalities and of monotonous worship of the past. Fretting at the conventionality of formal examinations, he passed the last competitive examination in 1882, but without the brilliance that his friends had predicted for him. He then accepted a series of provincial assignments as a teacher of philosophy at the state secondary schools of Sens, Saint-Quentin, and Troyes between 1882 and 1887. In 1885–86 he took a year's leave of absence to pursue research in Germany, where he was impressed by Wilhelm Wundt, pioneer experimental psychologist. In 1887 he was appointed as lecturer at the University of Bordeaux, where he subsequently became professor and taught social philosophy until 1902.

Durkheim was familiar with several foreign languages and reviewed volumes in German, English, and Italian at length in the learned journal *L'Année Sociologique*, which he founded in 1896. But it has been noted, at times with disapproval and amazement, by non-French social scientists, that he travelled little and that, like many French scholars as well as the noted British anthropologist Sir James Frazer, he never undertook any fieldwork. The vast information he studied on the tribes of Australia or of New Guinea or on the Eskimos was all collected by other anthropologists, travellers, or missionaries. This was not, in Durkheim's case, due to provincialism or lack of attention to the concrete. He did not resemble the French philosopher Auguste Comte in making venturesome and dogmatic generalizations and disregarding empirical observation. He did, however, maintain that concrete observation in remote parts of the world does not always lead to illuminating views on the past or even on the present. To him facts had no meaning for the intellect unless they were grouped into types and laws. He claimed repeatedly that it is from a construction erected on the inner nature of the real that knowledge of concrete reality is obtained, a knowledge not perceived by observation of the facts from the outside. He thus constructed concepts such as that of the sacred or of totemism, exactly in the same way that Karl Marx developed the concept of class.

Attitude toward fieldwork

In truth, Durkheim's vital interest did not lie in the study for its own sake of so-called primitive tribes, but rather in the light such a study might throw on the present. The outward events of his life as an intellectual and as a scholar may appear undramatic. Still, much of what he thought and wrote stemmed from the events that he witnessed in his formative years, in the 1870s and 1880s, and in the earnest concern he took in them. The Second Empire, which collapsed in the French defeat of 1870 at the hands of Germany, had seemed an era of levity and dissipation to the earnest young man that Durkheim was. France, with the support of many of her liberal and intellectual elements, had plunged headlong into a war for which she was unprepared; her leaders proved incapable. The left-wing Commune that took over Paris after the French defeat in 1871 led to senseless destruction, which appeared to Durkheim's generation, in retrospect, as evidence of the alienation of the working classes from capitalist society. The bloody repression that followed the Commune was taken as evidence of the ruthlessness of capitalism and of the selfishness of the frightened bourgeoisie. Later, the crisis of 1886 over Georges Boulanger, minister of war who demanded a centralist government to execute a policy of revenge against Germany, was one of several events that testified to the resurgence of nationalism, soon to be accompanied by anti-Semitism. Such major French thinkers of the older generation as Ernest Renan and Hippolyte Taine interrupted their historical and philosophical works, after 1871, to analyze those evils and to offer remedies. Durkheim was one of several young philosophers and scholars, fresh from their École Normale training, who became convinced that progress was not the necessary consequence of the development of science and technology; that it could not be represented by an ascending curve, justifying complacent optimism. He perceived around him an absence of moral norms or "anomie," an excess of individualism bordering on anarchy; material prosperity set free greed and passions that threatened the equilibrium of society.

These sources of Durkheim's sociological reflections, never remote from moral philosophy, were first expressed in his very important doctoral thesis, *De la division du travail social* (1893; *The Division of Labour in Society*) and in *Le Suicide* (1897; *Suicide*). In his view ethical and social structures were being endangered by the advent of technology and mechanization. The division of labour rendered workmen both more alien to one another and more dependent upon one another, since none of them any longer built the whole product by himself. Suicide appeared to be less frequent where the individual was closely integrated with his culture; thus the apparently purely individual decision to renounce life could be explained through social forces. These early volumes, and the one in which he formulated with scientific rigour the rules of his sociological method, *Les Règles de la méthode sociologique* (1895; *The Rules of Sociological Method*), brought Durkheim fame and influence. But the new science of society frightened timid souls and conservative philosophers, and he had to endure many attacks. The Dreyfus affair—resulting from the false charge against a Jewish officer, Alfred Dreyfus, of spying for the Germans—erupted in the last years of the century, and the slurs or outright insults aimed at Jews that accompanied it opened Durkheim's eyes to the latent hatred and passionate feuds hitherto half concealed under the varnish of civilization. He took an active part in the campaign to exonerate Dreyfus. He was not elected to the Institut de France, although his stature as a thinker suggests that he should have been named to that prestigious, learned society. He was, however, appointed to the University of Paris in 1902 and made a full professor there in 1906.

Effect of Dreyfus affair

More and more, the sociologist's thought became concerned with education and religion as the two most potent means of reforming man or of molding the new institutions required by the deep structural changes in society. His colleagues admired Durkheim's zeal in behalf of educational reform. His efforts included participating in numerous committees to prepare new curriculums and methods; working to enliven the teaching of philosophy, which too long had dwelt on generalities; and attempting to teach teachers how to teach. A series of courses that he had given at Bordeaux on the subject of *L'Évolution pédagogique en France* ("Pedagogical Evolution in France") was published posthumously in 1938; it remains one of the best informed and most impartial books on French education. The other important work of Durkheim's latter years dealt with the totemic system in Australia and bore the title of *Les Formes élémentaires de la vie religieuse* (1915; *The Elementary Forms of the Religious Life*). The author, despite his own agnosticism, evinced a sympathetic understanding of religion in all its stages. French conservatives, who in the years preceding World War I turned against the Sorbonne, which they charged was unduly swayed by the prestige of German scholarship, railed at Durkheim who, they thought, was influenced by the German urge to systematize, making a fetish of society and a religion of sociology. In fact, Durkheim did not make an idol of sociology as did the positivists schooled by Comte, nor was he a "functionalist" who explained every social phenomena by its usefulness in maintaining the existence and equilibrium of a social organism. He did, however, endeavour to formulate a positive social science that might direct men's behaviour toward greater solidarity.

The outbreak of World War I came as a cruel blow to him. For many years he had expended too much energy on teaching, on writing, on outlining plans for reform, on ceaselessly feeding the enthusiasm of his disciples, and eventually his heart had been affected. His gaunt and nervous appearance filled his colleagues with foreboding. The whole of French sociology, then in full bloom thanks to him, seemed to be his responsibility. The breaking

Death and legacy

point came when his only son was killed in 1916, while fighting on the Balkan front. The father stoically attempted to hide his sorrow; but the loss, coming on top of insults by nationalists who denounced a professor of "apparently German extraction" who taught a "foreign" discipline at the Sorbonne, was too much for him. He died on November 15, 1917.

He left behind him a brilliant school of researchers. He had never been a tyrannical master; he had encouraged his disciples to go farther than himself and to contradict him if need be. His nephew, Marcel Mauss, who held the chair of sociology at the Collège de France, was less systematic than Durkheim and paid greater attention to symbolism as an unconscious activity of the mind. Claude Lévi-Strauss, who has since occupied the same chair of sociology and resembles Durkheim in the way he combines reasoning with intensity of feeling, also offered objections and corrections to Durkheim's views. With Durkheim sociology had become in France a seminal discipline that broadened and transformed the study of law, of economics, of Chinese institutions, of linguistics, of ethnology, of art history, and of history.

BIBLIOGRAPHY. HARRY ALPERT, *Émile Durkheim and his Sociology* (1939, reprinted 1961), offers, in its first section, a good account of Durkheim's life. The article by TALCOTT PARSONS in *The International Encyclopedia of the Social Sciences*, 4:311–20 (1968), is clear and succinct. A collective volume, *Émile Durkheim, 1858–1917: A Collection of Essays with Translations and a Bibliography*, ed. by K.H. WOLFF (1960; issued in paperback as *Essays on Sociology and Philosophy, with Appraisals of Durkheim's Life and Thought*, 1964), contains a full chapter on Durkheim's life. The two best volumes in French are those of A.P. LA FONTAINE, *La Philosophie d'Émile Durkheim* (1926); and of JEAN DUVIGNAUD, *Durkheim: sa vie, son oeuvre* (1965).

(H.M.P.)

Duse, Eleonora

The most fluent and expressive actress of her day, Eleonora Duse created afresh every role she played and was different in each of them. Her gift was in marked contrast to the talented contemporary star of the French theatre, Sarah Bernhardt, a great technician who always strove to project her own personality from the stage, whatever character she might be playing. Duse found her great interpretative roles in the heroines of the Norwegian playwright Henrik Ibsen, but she could also lend her soul to the puppet-like creations of such dramatists as Victorien Sardou and the younger Dumas, bringing to life on stage the characters they themselves had failed to animate.

Eleonora Duse was born in a railway coach on October 3, 1858, near Vigevano, Italy. Most of her family were actors who played in the same touring troupe, and Eleonora made her first stage appearance at the age of four in a dramatization of Victor Hugo's *Les Misérables*. By the age of 14, when she played Juliet at Verona, her talents were already being recognized by critics; but after her family died she moved from one company to another, without a great deal of success, until her appearance at Naples in 1878. This marked the turning point of her career. Her performances there as Electra and Ophelia were admired, but it was her characterization of the title role in Émile Zola's *Thérèse Raquin* that won greatest acclaim, with audiences and critics united in the opinion that a woman's anguish had never before been played with such truth.

In 1882 Duse took an opportunity to watch Bernhardt perform. The French actress's success in modern roles gave Duse the idea also of appearing in plays by contemporary French dramatists (for she had discovered that Italian audiences were bored by the stale pieces that formed the traditional repertory), and so for three years she acted in a number of plays by the younger Dumas. The first of these was Lionette in *La Princesse de Bagdad*, a play from which Parisians had recoiled when a French actress played the lead. Duse, however, scored a triumph and followed it up with Cesarine in *La Femme de Claude;* in 1884 she created the title role of Dumas's latest play, *Denise*, and also the part of Santuzza in Giovanni Verga's



Eleonora Duse.
By courtesy of the Library of Congress, Washington, D.C.

Cavalleria rusticana. With Cesare Rossi, a prominent actor-manager, she toured South America in 1885, but after her return to Italy she formed her own company, the Drammatica Compagnia della Città di Roma, and with it toured Austria, Germany, England, France, Russia, Egypt, Belgium, and Portugal, making visits also to the United States in 1893, 1896, and 1902. Idolized wherever she went, she felt most appreciated as an artist in the city of Vienna.

Love affair with Gabriele D'Annunzio

In 1894 she met and fell in love with a rising young poet, Gabriele D'Annunzio, and he wrote for her a number of plays. Her belief in his talent was boundless, and his cult of beauty added another dimension to her acting. A radiant glow is said to have emanated from her whenever she played one of his characters. D'Annunzio told the story of their love in his novel *Il fuoco* (1900; Eng. trans., *The Flame of Life*).

Aside from D'Annunzio's plays, Duse found an inexhaustible source of self-expression in the dramas of Ibsen. She never tired of playing Nora in *A Doll's House*, Rebecca West in *Rosmersholm*, Ella Rentheim in *John Gabriel Borkman*, and, above all, Ellida in *The Lady from the Sea*. To the title role in *Hedda Gabler* she brought a demonic quality, a touch of the fantastic—deeply troubling to Ibsen when he saw her perform it—as though she had gone beyond the frontiers of realism.

Style of acting

The British playwright George Bernard Shaw was one of the many critics fascinated by Duse's ability to produce an illusion "of being infinite in variety of beautiful pose and motion." He confessed that "in an apparent million of changes and inflexions" he had never seen her at an "awkward angle" in defiance of the "natural gravitation toward the finest grace" (*Dramatic Opinions and Essays*, 1907). She had a thousand faces; her physical command, range, and choice of gesture were superb; she had a different way of walking for each part; she even mastered the colour of her face, which was unspoiled by makeup, and Shaw's professional curiosity was aroused when he pondered the question of whether Duse's blush could always be summoned at her will. Yet the total effect was of more than "naturalistic" acting: Duse acted not only the reality, she also commented on the characters she played

—she "knew" far more about Nora, for instance, than Ibsen's heroine could possibly have known about herself. One of her critics wrote that Duse played what was between the lines; she played the transitions. A tremor of her lips could reveal exactly what went on in her mind; and, where the character's inner life was lacking, because the dramatist had failed his task, she supplied motivation herself. To watch her was to read a psychological novel.

In 1909 Eleonora Duse quit the stage, mainly for reasons of health. Financial losses incurred during World War I, however, obliged her to emerge from retirement in 1921. Her acting powers were undiminished, but her health was still not good and interfered with her late career. In 1923 she appeared in London for six matinee performances, then played three nights in Vienna, before she embarked upon her last tour of the United States. This began with an appearance on the vast stage of the Metropolitan Opera House in New York City and ended in Pittsburgh where the actress died on April 21, 1924. Her body was taken back to Italy and, in compliance with her request, she was buried there in the small cemetery of Asolo.

BIBLIOGRAPHY. JEANNE BORDEUX, *Eleonora Duse: The Story of Her Life* (1925), the first biography in English—a simple, occasionally naïve account of the actress's tribulations and triumphs; EDOUARD SCHNEIDER, *Eleonora Duse, souvenirs, notes et documents* (1925), contains new material related to the last years of Duse's life, based upon conversations with the actress; ARTHUR SYMONS, *Eleonora Duse* (1926), rhapsodic impressions of the actress at various stations of her career; E.A. RHEINHARDT, *Das Leben der Eleonora Duse* (1928; Eng. trans., *The Life of Eleonora Duse*, 1930), a biography that reads like a novel and is basically uncritical—no attempt is made to analyze her acting style; O.R. SIGNORELLI, *La Duse* (1938), chronologically arranged anecdotes; BERTITA HARDING, *Age Cannot Wither: The Story of Duse and D'Annunzio* (1947), a narrative concentrating on the celebrated love affair; EVA LE GALLIENNE, *The Mystic in the Theatre: Eleonora Duse* (1966), a fellow artist's attempt to evoke the ephemeral fascination of another; the only study that pays attention to Duse's technique and the spiritual aspects of her art.

(A.M.N.)

Dvina River, Northern

One of the largest rivers and certainly the most important waterway of the northern part of the European portion of the Soviet Union, the Northern Dvina (Severnaya Dvina) flows slowly and majestically through a cold and inhospitable terrain. It is formed by the junction of the Sukhona and Yug rivers (349 miles [562 kilometres] and 357 miles [574 kilometres] long, respectively) at the city of Velikiy Ustyug, which lies in the Vologodskaya *oblast* (region) of the Russian Soviet Federated Socialist Republic just north of 61° north latitude. The river flows in a generally northwestern direction through the Vologodskaya and Arkhangelskaya *oblasti* and enters the Dvina inlet of the White Sea below the city of Arkhangelsk, just south of 65° north latitude, 140 miles from the Arctic Circle. The river's length is 462 miles (744 kilometres), and it drains a basin that, at 138,000 square miles (357,000 square kilometres), is larger than the whole of Poland. The fall of the river along its course is only about 160 feet; the flow is calm and, especially near the river mouth, very slow. (For related information see RUSSIAN SOVIET FEDERATED SOCIALIST REPUBLIC.)

Source of the river

The course of the river. Until its confluence with the tributary Vychegda River, the Northern Dvina is also occasionally called the Little Northern Dvina (Malaya Severnaya Dvina) with the remainder of its course known as the Greater Northern Dvina (Bolshaya Severnaya Dvina). A four-part division of the whole consists of an initial section downstream to the city of Kotlas; a portion downstream to the confluence of the Vaga; from the Vaga to the Pinega; and, finally, downstream to the mouth.

In the first section, the river flows northeast in the middle of a wide valley, with high banks alternating on the left and right. The width of the river is 1,000–1,600 feet and the depth from seven to 16 feet. In the second sector the amount of water in the river more than doubles with the influx of the Vychegda, the wide valley has been strewn with alluvial deposits, and the many-braided river bed—in which there are many shoals and sandbars—is contained between high banks. The well-established main channels include the Novinsky, Peschansky, Tikhy, and Syamovsky, and the strong spring floods constantly promote the formation of new channels. Near the river, extensive meadowlands, often marshy and containing dried-up river channels, are vulnerable to the swollen river waters. The river in this portion is 1,300–2,600 feet wide and from ten to 26 feet deep.

In the Vaga–Pinega section the river valley narrows considerably, and the river itself runs between steep limestone banks. Its width is now 2,000–3,300 feet and the depth 20–40 feet; it contains islands and shoals and occasionally spills over into narrow flood meadows. In its final section, the average width of the river valley is from three to four miles, but this conceals variations from 0.6 miles below the confluence with the Pinega to 11 miles at the village of Kholmogory, the home of the world-famous 18th-century scientist and poet M.V. Lomonosov. The river is single branched only at the beginning of this sector and at Arkhangelsk, but its basic bed averages 3,000–3,900 feet in width, and water depth is 23–39 feet; the floodplain is mainly on the left bank. At its mouth, the Northern Dvina Delta has an area of 425 square miles and is laced with a multitude of channels and branches, among which the Nikolsky, Korabeliny (both from Arkhangelsk to the sea), Kuznechikha, Maymaksy, and Murmansky are the most important.

The basin. The landscape drained by the Northern Dvina Basin is formed of low, undulating plains, falling away to the White Sea. It is bounded on the east by the low Timanski Ridge (Timanskiy Kryazh; where the Vychegda and its tributaries have their source) and the Severnye Uvaly Hills, which form the watershed with the Volga Basin to the south. The northern and central portions of the basin have a thick coating of coniferous forests, while mixed forests, with conifers predominant, are found to the south. In all, more than half the basin is forest clad. The soils of the region are mainly leached in type, reflecting their forest origins. Beyond the left bank of the river there are many low-lying bogs and lakes (including the large Lake Kubena [Ozero Kubenskoye]), which are often the source of tributary rivers. The Yemtsa (left-bank) and Pinega (right-bank) tributary basins are carved into karst scenery, with the limestone rocks of the areas permeated by water-dissolved holes and crevasses, springs, and disappearing streams; the erratic Sheleksa River is, in fact, a tributary of several other rivers.

The climate of the southern portion of the basin is moderately continental, becoming more extreme in the north. The average January temperatures range from 10° F (−12° C) to −4° F (−18° C), and the July readings are 50° F (10° C) to 64° F (18° C). Annual precipitation decreases from about 20 inches (500 millimetres) in the south to nearly half that figure in the north.

The banks of the Northern Dvina are largely covered by forests and shrubs, although in the scattered peopled sections they adjoin partly plowed areas crisscrossed by drainage canals. The floodplain is generally meadow-like in character and mostly marsh ridden. The mouth of the river is rich in fish, with about 30 species regularly caught.

Water flow. The Northern Dvina is primarily fed by melting snow, with widespread fluctuations in the water level occasioned by the building up of ice dams and ensuing spring floods. Highest levels occur in May, the lowest in late summer and autumn. The river-mouth sections are also influenced by the White Sea tides. The flow of water in the river shows a remarkable variation; although the average amount passing a point in the middle sections, for example, is about 70,600 cubic feet per second, the range is from a minimum of 2,300 cubic feet per second to a maximum of 700,000 cubic feet per second in flood time. Average figures for the upper and lower reaches (which also conceal variations) are 26,500 and

Water-flow variations

120,000 cubic feet per second, respectively. Most of the annual river flow (some 55–70 percent) is concentrated in the April–June period, as compared with only 6–8 percent the first three months of the year. In its upper course, the Northern Dvina begins to freeze around the end of November, becoming ice-free again by the end of April; the lower course is frozen for a slightly longer period. The spring witnesses frequent ice jams and floods along the whole river, particularly at Kotlas, around the Pinega confluence, and at Arkhangelsk.

The human imprint. The Northern Dvina Basin has undergone extensive improvement work in recent decades, most of it directed at facilitating the drainage of marsh, forest, and floodplain. Navigable sections of the river network of the basin have been improved by the straightening of portions of many smaller rivers, by the construction of timber dams, and by the formation of small reservoirs. The Northern Dvina itself is the main waterway of the northern European section of the Soviet Union and is linked—via the Sukhona River, Lake Kubena, the Porozovitsa River, and Northern Dvina Canal—with the important Volga–Baltic waterway and also, through the Pinega River, with the Kuloy and Mezen rivers. Navigation is possible for an average of 174 days each year in the upper course and 168 days in the lower. The main dock facilities are at Velikiy Ustyug, Kotlas, Krasnoborsk, Verkhnyaya Toyma, Yemetsk, and the famous northern port of Arkhangelsk. Prospects for the construction of a number of medium-sized hydroelectric stations in the upper and middle portions of the Northern Dvina attest to its emerging regional importance. (A.M.Ga.)

Dvina River, Western

The Western Dvina, or Zapadnaya Dvina in Russian (Zapadnaja Dvina in the transliteration system of the Soviet Akademiya Nauk) and Daugava in Lettish, is a major river in the European part of the Union of Soviet Socialist Republics, with a total length of 632 miles (1,020 kilometres). Its source is the small Lake Dvinets on the western slope of the Valdai Hills not far from the source of the Volga, in Kalinin *oblast* of the Russian Soviet Federated Socialist Republic. It flows southward for about 150 miles, then southwest and west for 185 miles through the Belorussian Soviet Socialist Republic; at Beshenkovichi it turns to the northwest and flows through the Latvian S.S.R., where it is known as the Daugava, discharging into the Gulf of Riga on the Baltic Sea just north of the city of Riga.

The river basin

Most of the river basin is between 300 and 700 feet above sea level—a rolling plain with many swamps and forests. The river drains an area of 34,000 square miles (88,000 square kilometres), but its tributaries are all small; they include the Mezha, Kasplya, Ulla, and Disna entering from the left; and the Toropa, Drissa, Aiviekste, with its tributary the Pededze, and Ogre entering from the right. The basin has more than 5,000 lakes, most of them quite small; among the larger are lakes Rezna and Lubana, in Latvia; Zhizhitsa, in the upper reaches of the river; Osveya and Drisvyaty, in the middle part of the basin on the border of Belorussia and Latvia; and Lukoml, in the southernmost part.

The basin has a humid climate with warm summers and mild winters. Average July temperatures are 61°–64° F (16°–18° C). January temperatures at Riga average about 23° F (−5° C) and at Velizh in the upper basin about 18° F (−8° C). Precipitation ranges from 28 to 30 inches in the middle part of the basin to 32 to 38 inches near the sea.

The character and size of the Western Dvina change gradually from its upper reaches to its mouth. In the upper reaches it flows on loamy, and then on morainic (glacially deposited) soils. Above the city of Vitebsk, it enters an area of sand and dolomite interspersed with seams of clay and marl. Throughout its middle course and much of its lower, the river flows along a valley with high, steep slopes cut with eroded ravines. It winds slowly, with few forks, until it approaches the mouth, where it divides into numerous branches that enter the gulf. Just above the city of Riga is the largest island of the Western Dvina, Dalen Island. Near the Latvian city of Daugavpils, the river is 650–1,000 feet wide, and it broadens to more than 2,000 feet in the lower reaches. Throughout its course there are rapids, rocks, and shallows, notably in the middle part around Disna and Kraslava and above Daugavpils. The current is fast and the slope steep.

The Western Dvina draws much of its water from melting snow, and, consequently, like other rivers of the eastern European plains, it has high spring floodwaters. It also floods after heavy rains. In spring the water level rises by 20 to 35 feet or more at various places. Its annual discharge is about 5.2 cubic miles. The average discharge is 301 cubic yards per second at Vitebsk, increasing to 406 at Polotsk, 615 near Daugavpils, and 890 at the mouth. The icebound period begins in the upper reaches in late November or early December and somewhat later in the middle part of the course. Thawing begins near the mouth of the river about the end of March, and in the upper reaches the water is open by about the middle of April.

Development of the Western Dvina

The Western Dvina has been an important water route since early times. Connected in its upper reaches by easy portages to the basins of the Dnepr, Volga, and Volkhov, it formed part of the great trade route from the Baltic region to Byzantium and to the Arabic east. At the beginning of the 19th century, it was joined by canals through its tributary the Ulla to the Berezina and thus to the Dnepr, but this system was never much used except for rafting timber. Through another tributary, the Drissa, it is connected with Lake Sebezha, and a small canal unites the Western Dvina with the Gavya River.

The river was first studied intensively in 1701, when, by command of Peter I, a survey was made from its source to the city of Polotsk. In 1790–91 a detailed atlas of the Western Dvina from Vitebsk to Riga was published.

The abundance of rapids and, in recent decades, the presence of dams have restricted navigation on the river to separate stretches: in the upper part, from Velizh to the mouth of the Ulla; in the middle and lower parts, from Kraslava to Livana and the dams of the Plyavinsky Hydroelectric Station; and in the lower part, from Marushka to Riga. The main items carried are lumber, construction materials, and grain. Seagoing vessels navigate the mouth of the river as far as Riga, nine miles from the sea.

Two hydroelectric stations have been built on the Western Dvina at Kegum (70,000 kilowatts) and Plyavinsky (825,000 kilowatts). Plans in the early 1970s called for a series of other stations with a total capacity of more than 1,000,000 kilowatts, the construction of additional irrigation systems, and various other measures to develop the river basin. (A.P.D.)

Dvořák, Antonín

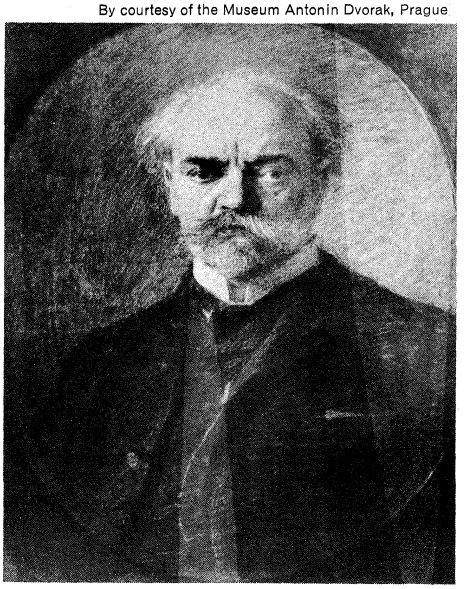
The first Bohemian composer to achieve worldwide recognition, Antonín Dvořák became one of the leading figures in the movements that injected traditional native folk materials, both musical and thematic, into the general framework of musical Romanticism that was characteristic of the 19th century. His fellow Bohemian Bedřich Smetana, 17 years his senior, already had laid the foundations of the Bohemian nationalist movement in music, while in Russia such composers as Aleksandr Borodin, Modest Mussorgsky, and Nikolay Rimsky-Korsakov were following a similar trend from the stores of that nation's past. His *Symphony No. 9 (From the New World;* 1893) remains his best known work partly, no doubt, because it was thought to be based on Negro spirituals and other influences gained during Dvořák's years in the United States. Although this may be true to some extent, the music is also characteristically Bohemian in its themes, possibly revealing the composer's nostalgia for his homeland.

Musical kinship

Dvořák was born in Nelahozeves, a Bohemian (now Czechoslovakian) village on the Vltava River north of

Prague, on September 8, 1841. He came to know music early, in and about his father's inn, and as a youngster became an accomplished violinist contributing to the amateur music making that accompanied the dances of the local couples. In 1857, a perceptive music teacher, understanding that young Antonín had gone beyond his own modest abilities to teach him, persuaded the elder Dvořák to enroll his son in an organ school in Prague. Later, without his father's financial assistance, Dvořák completed a two-year course and played the viola in various inns and with theatre bands, augmenting his small salary with a few private pupils.

By courtesy of the Museum Antonin Dvorak, Prague



Dvořák, portrait by Max Švabinský, c. 1900. In the Muzeum Antonína Dvořáka, Prague.

The 1860s were trying years for Dvořák, who was hard pressed for both time and the means, even paper and a piano, to compose. In later years he said he had little recollection of what he wrote in those days, but around 1864 two symphonies, an opera, chamber music, and numerous songs lay unheard in his desk. The varied works of this period show, however, that his earlier leanings toward Beethoven and Shubert were becoming increasingly tinged with the influence of Wagner and Liszt. In November 1873, at a time when a few successful concerts of his works had begun to make his name well-known in Prague, he married Anna Čermáková and began an unusually happy family life.

Years of acclaim

In 1875 Dvořák was awarded a state grant by the Austrian government, and this award brought him into contact with Brahms, with whom he formed a close and fruitful friendship. Brahms not only gave him valuable technical advice but also found him an influential publisher in Fritz Simrock, and it was with his firm's publication of the *Moravian Duets* (composed 1876) for soprano and contralto and the *Slavonic Dances* (1878) for piano duet that Dvořák first attracted worldwide attention to himself and to his country's music. The admiration of the leading critics, instrumentalists, and conductors of the day continued to spread his fame abroad, which led naturally to even greater triumphs in his own country. In 1884 he made the first of ten visits to England, where the success of his works, especially his choral works, was a source of constant pride to him, although only the *Stabat Mater* (1877) and *Te Deum* (1892) continue to hold a position among the finer works of their kind. In 1890 he enjoyed a personal triumph in Moscow, where two concerts were arranged for him by his friend Tchaikovsky. The following year he was made an honorary doctor of music of the University of Cambridge.

Dvořák accepted the post of director of the newly established National Conservatory of Music in New York in 1892, and, during his years in the United States, he travelled as far west as Iowa. Though he found much to interest and stimulate him in the New World environment, he soon came to miss his own country, and he returned to Bohemia in 1895. The final years of his life saw the composition of several string quartets and symphonic poems and his last three operas. He died in Prague on May 1, 1904.

The appeal of Dvořák's music lies chiefly in its teeming melodic invention and heart-warming simplicity. He essayed all of the musical genres and left (several) works that are regarded as classics in all of them; only his operas have failed to enjoy lasting success.

MAJOR WORKS

Orchestral works

SYMPHONIES: (revised numbers)—*No. 5 in F Major*, op. 76 (1875, rev. 1887; originally *No. 3*, op. 24); *No. 6 in D Major*, op. 60 (1880; originally *No. 1*); *No. 7 in D Minor*, op. 70 (1885; originally *No. 2*); *No. 8 in G Major*, op. 88 (1889; originally *No. 4*); *No. 9 in E Minor* (*From the New World*), op. 95 (1893; originally *No. 5*).

CONCERTOS: *Piano Concerto in G Minor*, op. 33 (1876); *Violin Concerto in A Minor*, op. 53 (1880); *Cello Concerto in B Minor*, op. 104 (1895).

SYMPHONIC POEMS: *Vodník*, op. 107 (*The Water-Goblin*, 1896); *Polednice*, op. 108 (*The Noonday Witch*, 1896); *Zlatý kolovrat*, op. 109 (*The Golden Spinning-Wheel*, 1896); *Holoubek*, op. 110 (*The Wood Dove*, 1896); *Píseň bohatýrská*, op. 111 (*Hero's Song*, 1897).

OVERTURES: *Domov můj*, op. 62 (*My Home*, 1881); *V přírodě*, op. 91 (*In Nature's Realm*, 1891); *Carnival*, op. 92 (1891); *Othello*, op. 93 (1892); opp. 91, 92, 93, originally a cycle of overtures entitled *Příroda, Život a Láska* (*Nature, Life and Love*).

RHAPSODIES AND DANCES: *Rhapsody in A Minor* (1874); *Three Slavonic Rhapsodies*, op. 45 (1878); *Slavonic Dances*, opp. 46, 72 (1878, 1886; orchestral version of piano duets).

MISCELLANEOUS: *Serenade in D Minor*, for two oboes, two clarinets, two bassoons, double bassoon, three horns, cello, and double bass, op. 44 (1878).

Chamber music

Thirteen string quartets; three piano trios; two piano quartets; quintets for piano and string quartet, for string quartet and viola, string quartet and double bass; sextet for string quartet and viola and cello; *Terzetto* for two violins and viola, op. 74 (1887); *Dumky Trio* for piano, violin, and cello, op. 90 (1891); pieces for violin and piano.

PIANO MUSIC: Waltzes, mazurkas, impromptus; *Slavonic Dances* for piano duet; *Eight Humoresques*, op. 101 (1894); *Ze Šumavy*, op. 68 (*From the Bohemian Woods*, 1884), for piano duet.

Vocal music

OPERAS: *Král a uhlíř*, op. 14 (*King and Collier*, first performed 1874); *Tvrdé palice*, op. 17 (*The Pig-headed Peasants*, 1881); *Šelma sedlák*, op. 37 (*The Peasant a Rogue*, 1878); *Dimitrij*, op. 64 (1882); *Jakobín*, op. 84 (*The Jacobin*, 1889); *Čert a Káča*, op. 112 (*The Devil and Kate*, 1899); *Rusalka*, op. 114 (1901); *Armida*, op. 115 (1904).

CHORAL WORKS: *Hymnus*, patriotic hymn for chorus and orchestra, op. 30 (*The Heirs of the White Mountain*, 1872; rev. 1880); *Stabat Mater*, for solo voices, chorus, and orchestra, op. 58 (1876–77); *The 149th Psalm* for male chorus and orchestra, op. 52 (1879), and mixed choir, op. 79 (1887); *V přírodě*, op. 63 (*Amid Nature*, 1882), five choral pieces for mixed voices; *Svatební košile*, op. 69 (*The Spectre's Bride*, 1884), dramatic cantata for solo voices, chorus, and orchestra; *Hymna československý rolnictva*, op. 28 (*Hymn of the Czech Peasants*, 1885), for chorus and orchestra; *Svatá Ludmila*, op. 71 (*St. Ludmilla*, 1886), oratorio for solo voices, chorus, and orchestra; *Mass in D Major*, for solo voices, chorus, and orchestra, op. 86 (1887; orchestral version, 1892); *Requiem Mass* for solo voices, chorus, and orchestra, op. 89 (1890); *Te Deum*, for soprano and bass solo, chorus, and orchestra, op. 103 (1892); *The American Flag*, cantata for alto, tenor, and bass solo, chorus, and orchestra, op. 102 (1893).

SONGS: *Moravské dvojzpěvy*, op. 32 (*Moravian Duets*, 1876), for soprano and contralto.

BIBLIOGRAPHY. The definitive work is OTAKAR SOUREK, *Život a dílo Antonína Dvořáka*, 2nd–3rd ed. rev., 4 vol. (1954–57), which has been published only in Czech. Books in English include Sourek's *Antonín Dvořák: His Life and Works* (1954); ALEC ROBERTSON'S informative and scholarly *Dvořák* (1945); JOHN CLAPHAM, *Antonín Dvořák: Musician and Craftsman* (1966), an admirable and erudite study of the

composer's working methods; and GERVASE HUGHES, *Dvořák: His Life and Music* (1967), in which Dvořák's compositions are discussed within a biographical framework and in relation to his sociological background. A comprehensive international bibliography complete up to 1959 is incorporated with JARMIL BURGHAUSER, *Antonín Dvořák: Thematic Catalogue* (1960).

(D.M.L.-J.)

Dyes and Dyeing

Most modern dyes are coloured substances synthesized from certain chemical compounds called benzenoid hydrocarbons, obtained from either coal tar or petroleum. Their most important use is in dyeing textiles. Wool or cotton fabrics may have dyes applied at any stage of manufacture from fibre to finished garment, depending on the nature of the textile, its subsequent processing, and the requirements in use. Dyeing, as the term is generally understood, occurs only when the dye is in solution, usually in an aqueous medium. Some dyes whose solubility is small are called disperse dyes, because they are dispersed rather than dissolved in the water. Some colorants, called pigments, are completely insoluble; they are melted together with synthetic resins to impart colour before the synthetic fibres are extruded. Synthetic textiles such as nylon are coloured in this way. (The term colorant is used to describe both dyes and pigments.) At the present time, dyeing from nonaqueous solvents, called simply solvent dyeing, shows considerable promise as a low-cost process that conserves water and reduces pollution.

Very few dyes of natural (animal or vegetable) origin are presently used by the commercial dyer; but before the mid-19th century they were the only colorants available. Their use over the centuries, especially from the Middle Ages on, laid the foundations of the art, and later the science, of the modern dyeing industry, which is of considerable complexity. Over 7,000 synthetic organic colorants are presently in commercial use, differing in fastness and other properties, and requiring different methods of application. Most of the dyestuffs are used by the textile industry, but the leather, paper, food, and cosmetic industries are also important users. Synthetic organic pigments are used in the manufacture of printing inks, paints, and automobile finishes, and for incorporation in plastics and rubber. Since each industry requires different properties in the dyes and pigments it employs, the large dye-making concerns usually provide technical service and advice to their customer industries on the selection and use of products. Despite the fact that colorants form only a small proportion of the cost of a finished article (less than 1.5 percent of the cost of a woman's dress and less than one percent of that of a woollen overcoat), the dye-making industry remains of considerable economic importance.

HISTORY

Dyes of natural origin. The discovery of red ochre in very ancient burial sites indicates that the use of colour for aesthetic and other purposes is at least 15,000 years old. The art of dyeing cloth is believed to have been known in China as long ago as 3000 BC. On the evidence of cultural and religious records, both madder, a dye prepared from the root of the madder herb, and indigo, a blue dye obtained from indigo plants, were used in India from around 2500 BC. At relatively the same time, dyeing of cloth in yellows, reds, and greens was also practiced in Egypt, where the earliest known textiles, linen mummy wrappings dating from 2000 BC, have been found dyed blue with either indigo or woad, a dye prepared from the leaves of the woad plant, which contains indigo as an essential constituent. The safflower, an Old World herb also called dyer's thistle, was used by the Egyptians around 2000 BC to obtain yellow and red shades; by 1450 BC ways of dyeing fine fabrics, principally linen, in a full range of hues had been gradually evolved.

Dyer's thistle

Wool, the chief material of the Babylonian and Assyrian textile industries that arose from the 15th century BC, was generally dyed in the raw state before spinning. Linen, in contrast, was dyed as thread before being woven into cloth. Homer mentions dyed textiles, and purple-dyed leather.

The art of purple dyeing, though not invented by the Phoenicians, was brought to a high standard by them. The dye they employed, Tyrian purple, was of excellent fastness even by modern standards and was extracted at great cost from the purpura shellfish found in Mediterranean countries. Pliny the Elder writes of the dyeing recipes used in his day and refers to the Egyptians' use of mordants (*i.e.*, chemicals that fix dyes in the material by forming insoluble compounds). He also mentions the skillful dyers of Gaul, who used vegetable dyes that vied with the finest purple.

Few records of the arts and crafts have survived from the early Middle Ages. Dyers' guilds were in existence in many places in the high Middle Ages. Dyes and dyed fabrics were carried to Europe from places as distant as Peking along the famous Silk Route. Surviving specimens of richly coloured silks testify to the range of natural colorants available to Chinese and Japanese dyers of the period. The "Arte di Calimala," the cloth industry of Florence, was famed for its excellent dyeing and finishing. In Venice the charter of the Dyers' Guild, the "*Mariegola dell'arte dei Tintori*" contains recipes used by the guild. Guilds of dyers sprang up in France, Flanders, Germany, Britain, and European countries.

The natural dyes used up to the middle of the 19th century, the greater part of them of vegetable origin, involved many problems, both in respect to sources and to processes. Tyrian purple mysteriously disappeared about the middle of the 15th century. The use of natural indigo involved protracted fermentation processes to liberate the dye in a soluble, colourless form. After cloth had been steeped in a vat, oxidation in air gave blue dyeings on cotton and wool. Many natural dyes have no affinity per se for textile fibres until such fibres have been treated with aluminum, iron, or tin compounds to receive the dye (mordanting). Chief among the mordant dyes was alizarin derived from madder root, widely cultivated at one time in Europe, Turkey, and India. It yielded a fast red on aluminum mordant, a purplish black on iron mordant; intermediate shades were obtained by using a mixture of mordants. The discovery of the New World and the establishment of trade routes to the Americas led to the introduction into Europe of new mordant dyes such as brazilwood, giving reds on aluminum, browns on iron, and a rose-pink on tin mordants, and logwood, or haematoxylin, giving fine blacks on a chromium oxide mordant. Scarlets were obtained either from kermes, which consists of the dried bodies of certain female insects, or cochineal, produced in Mexico from the dried female insect *Coccus cacti*, which lives on the Indian fig tree or nopal. Until 1954 the traditional scarlet of British Guards' uniforms was obtained by dyeing tin-mordanted wool with cochineal.

Synthetic dyes. The year 1856, in which William Henry Perkin (1838–1907), English chemist, discovered the synthetic dye mauveine, marks the beginning of synthetic dyestuffs, which ultimately supplanted natural dyes. The synthetic-dye manufacturing industry was founded by Perkin in 1857, when he set up facilities near London for the commercial production of mauveine and, later, of other synthetic dyes. Other dye-making factories followed both in the U.K. and continental Europe, and new dyes began to appear on the market. Research was directed toward determining the structures of important natural dyes and then to synthesizing them. The first success was the synthesis of artificial alizarin from coal-tar anthracene in 1868. Indigo was first synthesized in 1880 though an economic manufacturing process was not devised until almost 20 years later. The cultivation of madder and indigo rapidly declined as the synthetic products came to be accepted by dyers. Organic chemistry flourished in the research centres of Europe, and new dyes, many of them owing nothing to nature, were perfected by dye makers. These included azo dyestuffs,

The discovery of mauveine

certain organic compounds having the general formula RN = NR′ (1862), which today comprise the largest group of synthetic dyes; congo red, the first direct cotton dye requiring no mordanting or fixing process (1884); azoic dyes, a group of organic dyes derived from azo benzene, mostly reds and yellows (1911); acetate dyes, mainly azo or anthraquinone dyes, highly dispersed to render them capable of penetrating acetate fibres (1922–23); and fibre-reactive dyes, those which combine chemically to form a link with the material (1956). Dyestuffs research continues, as indicated by the large annual volume of patent literature.

RAW MATERIALS FOR MANUFACTURE OF DYES

The primary raw materials from which the dye maker synthesizes dyestuffs and pigments are the coal-tar chemicals called aromatic hydrocarbons, chiefly benzene, toluene, orthoxylene, metaxylene, paraxylene, naphthalene, and anthracene. Dye manufacture thus involves the process of first deriving these intermediate chemicals from raw materials and then building up dyes from the intermediates.

Coal tar, the traditional source of dye intermediates, is obtained from bituminous coal by distillation. Coal tar is a complex mixture of substances, which, on distillation, is commonly separated into light oil (boiling point up to 170° C), middle oil (boiling point 170°–230° C), heavy oil (boiling point 230°–270° C), anthracene oil (boiling point above 270° C), and pitch (nonvolatile still residue). The light oil is redistilled and separated into fractions of narrower boiling-point range: 80°–110° C, 110°–140° C, and 140°–170° C. These fractions are purified by successive contact with caustic soda (sodium hydroxide) solution, and concentrated sulfuric acid. The treated fractions are again distilled to yield respectively 90 percent benzol, 50 percent benzol, and solvent naphtha. By fractional distillation the 90 percent benzol may be separated into its constituents, benzene, toluene, and the xylenes. The commercial benzene thus obtained is not pure enough for successful synthesis into dye intermediates and must be purified even further. This is accomplished by a freezing process, followed by treatment with sulfuric acid. Toluene and the xylenes are brought to a similar degree of chemical purity by a combination of chemical and physical means.

Naphthalene is obtained from middle oil (170°–230° C), from which it separates, on cooling, as crystals, which are pressed, treated with caustic-soda solution, then with sulfuric acid, and finally distilled or sublimed (vaporized and recondensed to solid form).

Anthracene is separated from anthracene oil in the form of crude crystals. These are then dissolved in naphtha, distilled in superheated steam, and recrystallized in pure form.

New sources of hydrocarbons

Although coal tar remains an important source of aromatic hydrocarbons, the petrochemicals industry is an increasingly important supplier; in the United States it is the principal source. Distillation of crude oil yields substances that can be reformed either by thermal or catalytic processes and suitable purification into the aromatic hydrocarbons. Largest users are the synthetic fibre, synthetic rubber, detergent, and plastics industries, but sizable quantities are consumed by the dye-making industry (see also PETROLEUM REFINING).

MANUFACTURING PROCESSES FOR INTERMEDIATES

All synthetic organic dyes and pigments contain a ring structure of atoms. The same formation is present in dye intermediates. It is usual also for the dye intermediate to have one or more of the groups of atoms that react chemically to form salts.

A simple example of an intermediate is β-naphthol, which is obtained from naphthalene by heating it with 96 percent sulfuric acid at 160° C (320° F), adding sodium sulfate, and volatizing the remaining naphthalene with steam. Sodium naphthalene-2-sulfonate, left behind after this steam treatment, is then fused with caustic soda to yield β-naphthol. The β-naphthol is separated from the mixture by diluting it with water and treating it with sulfuric acid. These reactions, known respectively as sulfonation and alkali fusion, are given the name unit processes.

Unit processes

A number of such unit processes are available to the chemical manufacturers, and by means of various permutations, it is possible to synthesize the several hundred intermediates needed for the manufacture of the modern range of colorants.

Nitration. In nitration, a nitro group is substituted for a hydrogen atom in an aromatic hydrocarbon molecule by the action of nitric acid. Benzene yields nitrobenzene by reaction with a mixture of nitric and sulfuric acids at a temperature not exceeding 50° C (122° F). Toluene, with mixed acid at 30°–35° C (86°–95° F), yields orthonitrotoluene, metanitrotoluene, and paranitrotoluene. Technical quality nitrotoluenes are obtained from the crude nitration product by physical separation methods including fractional distillation; 1-nitronaphthalene is obtained when naphthalene is nitrated.

Nitration is a strongly exothermic, or heat-releasing, reaction; in large-scale nitrations the temperature must be maintained between established limits and efficient means of cooling provided. Cast-iron, mild steel, or stainless-steel vessels are used, arranged to permit agitating the mixture. The addition of nitric acid is carefully controlled by instruments that detect incipient rises in temperature. Other safeguards stop the addition in case the stirring mechanism is halted by a power or mechanical failure.

Reduction. Amines, chemical compounds formed by substituting organic radicals for the hydrogen atoms of ammonia, may be obtained from nitro compounds by reduction; that is, replacement of oxygen by hydrogen. This is accomplished by mixing the nitro compound with iron borings and a minimum quantity of aqueous hydrochloric acid in a cast-iron reducer with a powerful agitator. The end products are aniline, or other amines, and ferric oxide. The aniline, which separates as an upper layer above the aqueous suspension of iron and oxide, is removed mechanically and purified by steam distillation followed by fractionation. Analogous processes are employed in making other amines from their corresponding nitro compounds.

Halogenation. Halogenation is the introduction, by direct or indirect means, of a halogen (fluorine, chlorine, or bromine) into molecules of dye intermediates. In many cases direct substitution can be effected using elemental chlorine or bromine with or without a catalyst. In other cases a diazotized amine is treated with cuprous chloride or bromide. Indirect means must be employed for introducing fluorine. Hydrogen fluoride (HF), for example, is used to displace chlorine in chloro compounds.

Amination. The amination, or conversion of chloro compounds to amines, can be accomplished by ammonolysis, a process that involves heating the chloro compound with aqueous ammonia in a steel autoclave (equipment for working at pressures above atmospheric). Chloronitrobenzene, for example, is transformed into nitroaniline by this process. Several industrial amination processes exist. In the Dow Process monochlorobenzene is converted into aniline by heating it with aqueous ammonia at 240° C (464° F) in the presence of a copper oxide catalyst. In the Halcon Process phenol undergoes ammonolysis to yield aniline.

Hydroxylation. Hydroxylation is the introduction of a hydroxyl, or OH, group into the dye-intermediate molecule. Among the methods used for this are fusion by sodium hydroxide, Bucherer reaction (conversion of aromatic amines into phenolic compounds by aqueous sulfite or bisulfite), hydrolysis of chloro compounds under various reaction conditions according to the reactivity of the chloro compound, decomposition of diazonium salts by hot aqueous sulfuric acid, hydrolysis of sulfo groups, and oxidation (described below); for example, the compound cumene yields phenol and acetone.

Phenol and acetone

Oxidation. Oxidation, the combination of a substance with oxygen, or, generally, any reaction in which an atom

loses electrons, is a ubiquitous chemical reaction in dye chemistry; only a few examples can be given here. The conversion of methyl groups into carboxylic acids, and more especially into styryl compounds, is effected by various oxidants such as sodium hypochlorite, potassium permanganate, and sodium dichromate. The catalyzed oxidation of naphthalene with air leads to phthalic anhydride. Other important reactions include oxidation of leuco compounds to dyes and the formation of complex polycyclic substances from simpler molecules.

Benzidine rearrangement. The benzidine rearrangement consists of the conversion of nitrobenzene and its derivatives into derivatives of biphenyl in a two-stage process. The first stage consists of alkaline reduction with zinc dust to the hydrazobenzene; and the second of treatment with hydrochloric acid.

The above and other unit processes are used in various combinations in order to produce the important intermediate chemicals that are, in turn, used to manufacture the dyes themselves.

H Acid. H Acid, which is 1-amino-8-hydroxynaphthalene-3,6-disulfonic acid, an important dye intermediate, is produced from naphthalene by a combination of the unit processes of sulfonation, nitration, reduction, and hydrolysis. H Acid is used in the manufacture of a large number of azo dyes.

Diaminostilbene disulfonic acid. 4,4′-diaminostilbene-2,2′-disulfonic acid is produced from *p*-nitrotoluene by a combination of the unit processes of sulfonation, oxidation, and reduction. This product is used in large quantities for manufacture of fluorescent brightening agents.

MANUFACTURING PROCESSES FOR DYES AND PIGMENTS

In the early synthetic dye plants, the equipment was primitive, consisting of simple iron vessels and wooden vats. In the latter half of the 19th century new processes arising from the discovery of an increasing number of new dyes led to improvements in the design of reaction vessels, autoclaves, pumps for the transfer of liquids, compressors, and vacuum pumps. The development of vitreous coating techniques, the use of lead and copper, and improvements in the centrifuge, drying stove, grinding, and mixing equipment also proceeded at that time and has continued to the present. The advent of stainless steel and the introduction of plastic and polymer materials and of the newer techniques, especially the rubber coating of mild steel, has resulted in the gradual disappearance of the wooden vat and has reduced the use of lead and copper as constructional materials. In modern plants multipurpose vessels constructed for easy linking to one another and to filters, separators, etc. are employed wherever possible. Rapid changes may be made from one combination to another for efficient production. Stainless steel and rubber-covered steel may be quickly and easily cleaned in contrast to the former laborious process of cleaning wooden vats. In the larger modern plants, flow rates, heating, cooling, and other services are usually automatically controlled. Inlet and outlet valves are actuated, usually pneumatically, by sensing probes that detect changes in temperature, pressure, and hydrogen ion concentration that have occurred in the reaction vessel or other part of the plant. Intermediates, dyes, and pigments are usually manufactured in batches, discontinuously, because the total annual requirements of each may not amount to more than 10–100 tons (9–90 metric tons). If an annual output greatly in excess of these quantities is required, semicontinuous or continuous processes may be adopted, relatively small amounts of reactants coming together at any one time. This increases the safety factor in strongly exothermic reactions, such as nitrations and sulfonations.

Discontinuous batch operation

Many of the unit processes used to produce intermediates are also used to make dyes and pigments. In the processes described below, only the penultimate and final stages leading to the particular colorant are given. For precision, the Part I Reference, and the Part II number, where known, are given from the *Colour Index* (C.I.), the standard reference work published jointly by the Society of Dyers and Colourists in Great Britain, and the American Association of Textile Chemists and Colorists in the United States, with the cooperation of manufacturers in Belgium, Brazil, Chile, France, Germany, Greece, India, Italy, Japan, The Netherlands, Poland, Spain, and Switzerland. Part I of the *Colour Index* classifies dyes and pigments according to usage while Part II classifies them according to chemical constitution.

Nitroso dyes. By treating β-naphthol with nitrous acid, α-nitroso-β-naphthol is obtained, which reacts with an iron salt such as ferrous sulfate to give Pigment Green B, C.I. Pigment Green 8, 10006.

Nitro dyes. Celliton Fast Yellow RR, C.I. Disperse Yellow 1, 10345, results from the condensation of 2,4-dinitrochlorobenzene with 4-aminophenol. The process may be carried out in alcoholic solution. The final product, like all disperse dyes, must be ground to a particle size of 1–10 microns (1,000 microns = 1 millimetre) in an aqueous medium containing a dispersing agent.

Azo dyes. Almost all azo dyes are made by diazotization of a primary aromatic amine in an acid solution by using nitrous acid in the presence of ice. The diazo compound so formed is coupled with a suitable component, such as an aromatic amine, naphthol, or other phenolic substance, to form an azo compound.

These reactions are usually carried out in aqueous solution in rubber-lined steel equipment of capacities varying from 2,000 to 5,000 gallons (7,600–19,000 litres). Isolation of the final dye is usually achieved by addition of common salt in sufficient quantity to precipitate the dye, while leaving small amounts of undesirable, often dulling, by-products in solution. The dye precipitate is then separated from the solution in a filter press. As "press cake," it undergoes the finishing processes usual for most dyes, such as wet milling, drying, grinding, and reduction to a standard strength by mixing with diluents. Monoazo colorants contain a single azo group. Examples are the wool dye Acid Orange II, C.I. Acid Orange 7, 15510; and Permanent Red FRR, C.I. Pigment Red 2, 12310. Examples of diazo colorants, the molecules of which contain two azo groups, include Congo Red, C.I. Direct Red 28, 22120; and the solvent-soluble Fat Red B, C.I. Solvent Red 24, 26105. Triazo dyes contain three, and tetrakisazo dyes contain four, azo groups.

Acid alizarine dyes. Valuable greens and blues are derived by the process of sulfonation of diaminoanthraquinones to give dyes with the requisite degree of water solubility. Processes of sulfonation, nitration, reduction, alkylation, and other steps are also used to obtain certain disperse dyes from anthraquinone.

The more complex vat dyes, often with elaborate molecular structures, are built up from simpler compounds by many of the common unit processes such as sulfonation, as well as more specialized chemical reactions.

Indigoid and thioindigoid dyes. The sodamide process (1901) for Indigo, C.I. Vat Blue I, 73000, begins with aniline, which, by reaction with formaldehyde and sodium cyanide, yields phenylglycine; the latter on fusion with a mixture of sodamide and caustic potash is changed into indoxyl, oxidation of which leads to indigo. The thioindigoid, Durindone Orange R, C.I. Vat Orange 5, 73335, is also synthesized by a complicated route.

Sodamide Indigo process

Triphenylmethane dyes. Crystal Violet, C.I. Basic Violet 3, 42555, has a typical structure containing a central carbon atom, which is introduced by various methods including the employment of phosgene, $COCl_2$, formaldehyde, HCHO, and intermediate chemicals containing the groups = C = O and = CH–OH.

Cyanine dyes. Dyes in this class commonly contain two ring systems of atoms linked by conjugate chains (N–C = C–C = N, for example). There are many different methods of preparation, among which are the condensation of complex aldehydes with amines.

Fibre-reactive dyes. Fibre-reactive dyes react chemically to combine with and become part of the fibre substance. The simpler types are monochlorotriazinyl and dichlorotriazinyl compounds. An amino compound

with an azo, phthalocyanine, or anthraquinone group in its molecule is condensed in an aqueous medium with cyanuric chloride.

Phthalocyanine pigments. Copper phthalocyanine, a blue pigment, is manufactured in hundreds of tons annually by heating phthalonitrile, or some other related substance derived from phthalic anhydride, with a copper (cuprous) salt. For a high-grade pigment quality, the raw pigment must be subjected to special conditioning processes to achieve the required physical form. Valuable green pigments are obtained from copper phthalocyanine by halogenation processes.

TECHNIQUES AND EQUIPMENT

The dyeing of a fibre or textile is carried out in a solution, generally aqueous, known as the dye liquor or dyebath. For true dyeing (as opposed to mere staining) to have taken place, the coloration must be relatively permanent, that is, not readily removed by rinsing in water or by normal washing procedures; moreover, the dyeing must not fade rapidly on exposure to light. The process of attachment of the dye molecule to the fibre is one of absorption, that is, the dye molecules concentrate on the fibre surface.

Binding forces. There are four kinds of forces by which dye molecules are bound to fibre: (1) ionic forces, (2) hydrogen bonding, (3) van der Waals' forces, and (4) covalent chemical linkages. In the dyeing of wool, which is a complex protein containing about twenty different α-amino acids, the sulfuric acid added to the dyebath forms ionic linkages with the amino groups of the protein. In the process of dyeing, the sulfate anion (negative ion) is replaced by a dye anion. In the dyeing of wool, silk, and synthetic fibres, hydrogen bonds are probably set up between the azo, amino, alkylamino, and other groups, and the amido, –CO–NH–, groups. Van der Waals' forces (the attractive forces between the atoms or molecules of all substances) are thought to act in the dyeing of cotton between the molecular units of the fibre and the linear, extended molecules of direct dyes. Covalent chemical links are brought about in the dyebath by chemical reaction between a fibre-reactive dye molecule, one containing a chemically reactive centre, and a hydroxy group of a cotton fibre, in the presence of alkali.

In any dyeing process, whatever the chemical class of dye being used, heat must be supplied to the dyebath; energy is used in transferring dye molecules from the solution to the fibre as well as in swelling the fibre to render it more receptive. The technical term for the transfer process is exhaustion. Evenness of dyeing, known as levelness, is an important quality in the dyeing of all forms of natural and synthetic fibres; it may be attained by control of dyeing conditions, that is, by agitation to ensure proper contact between dye liquor and substance being dyed, and by use of restraining agents to control rate of dyeing, or strike.

Solvent dyeing. Serious consideration has recently been given to methods of dyeing in which water as the medium is replaced by solvents such as the chlorinated hydrocarbons used in dry cleaning. There are a number of technical advantages in solvent dyeing, apart from the elimination of effluent (pollution) problems associated with conventional methods of dyeing and finishing. Advantages include more rapid wetting of textiles, less swelling, increased speed of dyeing per given amount of material, and savings in energy, because less heat is required to heat or evaporate perchlorethylene, for example, than is needed for water.

FIBRE PREPARATION

Cotton. The impurities associated with natural fibres, and the materials added during processing of natural, modified, and synthetic fibres into yarns and cloth must be removed if dyeing is to be successful. Cotton in the raw state contains organic impurities such as proteins, waxes, tannins, and colouring matters, along with inorganic compounds. During the course of being woven into cloth, cotton is treated with size, fungicides, and other materials, all of which must be removed before dyeing. The cloth is first passed through a singeing machine to burn off surface fibres, then through a bath of enzyme preparations to remove the starch used as a size. Next it is heated with aqueous alkali under pressure in a boiler, or kier, to saponify, or hydrolyze fats, break down proteins and tannic acid, and solubilize the tough sheath surrounding the fibres. Finally the cloth is rinsed, treated with dilute acid, and washed acid free with weak ammonia. Sometimes a second alkali treatment, mercerization, is given, using 18–30 percent sodium hydroxide solution; this causes the fibres to swell and to acquire greater dye affinity. Bleaching with hypochlorite then removes the natural yellowness that would otherwise cause dull shades on dyeing.

Mercerization

Wool. The chief impurities in raw wool are (1) vegetable matter, removed in part mechanically and in part by carbonizing, that is, the preferential destruction of cellulose-like matter by sulfuric acid; (2) suint, dried perspiration of sheep, removed by rinsing in water in which it is readily soluble; and (3) wool grease, the source of such valuable by-products as lanolin and cholesterol, removed for recovery by scouring, or washing, and emulsification, or conversion into a mixture with a liquid. In subsequent spinning and weaving, saponifiable oil such as oleine is added to replace the natural grease and must be removed later by scouring with sodium carbonate solution. Small quantities of mineral oils may be removed by emulsification; larger quantities may require the addition of a solvent to the cleaning solution. Knitting yarns are scoured by emulsification; carpet yarns, usually heavily loaded with oleine, are scoured by saponification with sodium carbonate and a synthetic detergent. Woollen goods are bleached with hydrogen peroxide. As noted, dyeing is not necessarily the final stage of processing; wool may be dyed at any point from raw wool (dyed in the wool) to finished article.

Other natural fibres. Silk is degummed with hot soap solution to remove the sericin that binds the filaments in the natural state.

The modified natural fibres, viscose and cellulose acetate, are sized with either starch or gelatin; these are removed by enzyme or mild aqueous treatments respectively.

Synthetic fibres. In general, synthetic fibres possess excellent mechanical strength. Even so, they are usually sized with starch, water-soluble plastics, gelatin, and other agents to strengthen them against the stresses of knitting and weaving. The kind of scouring applied to remove the sizing depends on the properties of the particular fibre. In the case of blended fibres, the scouring conditions are limited by the most sensitive fibre present; thus a wool-nylon blend is treated as wool.

Solvent scouring processes. Shorter, more economical scouring processes using chlorinated hydrocarbon solvents such as stabilized trichloroethylene are currently being developed. In these processes the solvent may be recovered and reused.

DYE APPLICATION

For each application the dyer selects the combination of dyes best suited to the particular fibre or blend he plans to dye, and best able to withstand the conditions the textile will encounter in further processing and in use in the finished article. In general, the higher the standard of fastness, the more expensive the dye, and the final choice may be a compromise between the desired fastness standards and the cost of the dyes. Fastness tests and standards have been the subject of work by the American Association of Textile Chemists and Colorists (AATCC), Europäisch-Continentale Echtheitsconvention (ECE), and the Society of Dyers and Colourists (SDC), Bradford, Yorkshire. Efforts have been made to set up a unified system by the International Organization for Standardization (ISO). Light fastness is assessed on a scale of 8; 1 represents the poorest fastness and 8 the best. Fastness to other agents, among them water, bleach, acid, alkali, detergent solution, and perspiration, is measured on a scale of 5.

Fastness standards

Dyes are generally used in combination to achieve a desired hue or fashion shade. If the substance to be dyed consists of only one type of fibre, such as wool, the dye mixture will be made up solely of wool dyes. But if the fabric contains more than one kind of fibre and they differ in dyeing properties, then mixtures of different application classes of dyes are used.

Table 1 shows the application classes most suited to various fibres; minor usages are indicated in parentheses.

Table 1: Fibres and Dyes

fibres	application classes
Natural fibres	
Animal	
Wool	acid, basic, mordant, reactive, (solubilized vat)
Wool blends (wool–cotton, wool–viscose, etc.)	acid, direct, mordant, reactive
Silk	acid, basic, direct, mordant, (reactive), (solubilized vat)
Vegetable	
Cotton	azoic, basic, direct, mordant, oxidation, reactive, sulfur, vat
Bast (linen, flax, hemp, jute, ramie)	acid, direct, (disperse), reactive, vat, solubilized vat
Modified cellulose fibres	
Viscose	direct, mordant, pigment, reactive, sulfur, vat, solubilized vat
Secondary acetate	disperse
Triacetate	disperse
Synthetic fibres	
Polyamide (nylon, Perlon, Rilsan)	acid, disperse, mordant, pigment, reactive
Polyester (Dacron, Terylene)	disperse, pigment
Polyacrylonitrile (Acrilan, Courtelle, Orlon)	basic, disperse, pigment
Polyvinyl chloride (Envilon, Thermovyl)	basic, disperse
Polyolefines (Meraklon, Prolene)	disperse
Elastomers (Glospan, Lycra)	acid, disperse, reactive, (wool), vat

Dyebath preparation. Details vary according to the class of dye and the nature and form of the substance dyed. Variables include the percentage of dye in the dyeing solution, the other chemicals added, the temperature of the dyebath, and the time of heating.

Acid-levelling dyes are heated in solution with sulfuric or formic acid after Glauber's salt (sodium sulfate) is added to restrain the reaction when dyeing wool. Nylon is dyed in a similar manner at as high a temperature as possible with acetic or formic acid.

Direct dyes are applied to cotton from a solution having common salt (sodium chloride) or Glauber's salt added to promote the process of transfer of the dye from the solution to the material. In the case of certain dyes, the colourfastness may be improved by aftertreatments with bichromate, copper, or formaldehyde, and by development (diazotization and coupling with an added component).

Vat dyes and sulfur dyes are introduced into the fabric in the form of a soluble compound, which then reacts by oxidation with air, or an added chemical (bichromate or perborate) to form or regenerate the pigment in the fibre. Any loosely held pigment must then be carefully removed by soaping.

Disperse dyes usually require the addition of some dispersing agent to the solution to prevent aggregation. When polyester fibres, which are hydrophobic, are being dyed, chemicals such as chlorobenzenes, called carriers, are needed, or the process may be carried out quickly at high temperature in pressure vessels, without a carrier. Dry methods are also used, such as padding fabric with the disperse dye and then exposing it for a very short time to very high temperatures (200° C, or 392° F).

Textile printing. Textile printing can be regarded as localized dyeing. Colorants are used in the form of a printing paste containing a thickener such as gum tragacanth to prevent diffusion of the dye, and hence to preserve a sharp edge to the imprint. Various techniques are employed. In the ancient style the pattern was printed with a mordant fixed by steaming, and then the fabric was dyed; only the mordanted parts retained the colour. In direct style the fabric is printed with colorants or their precursors and then subjected to ageing and various fixing processes. Almost all classes of dyes are used in this style.

Finishing. Various finishing treatments may be applied to dyed fabrics and garments to impart shrink resistance, crease resistance, permanent pleating, or waterproofing. The processes may influence hue and fastness and may require special properties in the colorants.

MATERIALS

Forms in which textiles are dyed. Loose stock consists of randomly distributed wool or cotton fibres; tow is the corresponding term for synthetic fibres. Sliver is a more orderly arrangement of fibres in a loosely connected, continuous form suitable for spinning. It is wound into either hanks or tops, loose balls about one foot in diameter. After spinning, the yarn is either made up into hanks or into packages weighing about two pounds each, by winding the yarn round perforated metal tubes. The packages are curiously named, some according to their shapes; for example, cones, cheeses, cakes, beams, and rockets. Piece goods, woven cloth or textiles knitted in rope form, and garments, a term that includes stockings, tights, hose, and half hose, are also dyed as such.

Hanks and packages

Forms in which textiles are printed. The chief form of textiles for printing is piece goods, woven material. For special melange-printing effects, sliver or slubbing, slightly twisted roving, is printed and then spun, for example into carpet thread.

MACHINERY AND EQUIPMENT FOR DYEING AND PRINTING

Modern dyeing machines are made from stainless steels. Steels containing up to 4 percent molybdenum are favoured to withstand the acid conditions that are common. A dyeing machine consists essentially of a vessel to contain the dye liquor, provided with equipment for heating, cooling, and circulating the liquor into and around the goods to be dyed or moving the goods through the dye liquor. The kind of machine employed depends on the nature of the goods to be dyed. Labour and energy costs are high in relation to total dyeing costs; the dyer's aim is to shorten dyeing times to save steam and electrical power and to avoid spoilage of goods.

Machines for dyeing loose stock. A widely used machine is the conical-pan loose-stock machine; fibres are held in an inner truncated-conical vessel while the hot dye liquor is mechanically pumped through. The fibre mass tends to become compressed in the upper narrow half of the cone, assisting efficient circulation. Levelling problems are less important because uniformity may be achieved by blending the dyed fibres prior to spinning.

Hank-dyeing machinery. The Hussong machine is the traditional apparatus; it has a long, square-ended tank as dyebath into which a framework of poles carrying the hanks can be lowered. The dye liquor is circulated by an impellor and moves through a perforated false bottom that also houses the open steam pipe for heating. In modern machines, circulation is improved especially at the point of contact between hank and pole. This leads to better levelling and elimination of irregularities caused by uneven cooling.

The Hussong machine

Package-dyeing machines. Dye liquor may be pumped in either of two directions: (1) through the perforated central spindle and outward through the package, or (2) by the reverse path into the outer layers of the package and out of the spindle. In either case levelness is important. In the case of soluble dyes the dye liquor must be free of suspended matter. In the case of disperse dyes, in which particles of dye are dispersed in, rather than dissolved in, the solution, no gross aggregates can be allowed; otherwise the packages would retain undesirable solids on the outer and inner surfaces. Some package-dyeing machines are capable of working under pressure at temperatures up to 130° C.

Piece-dyeing machines. The winch is the oldest piece-dyeing machine and takes its name from the slatted roller that moves an endless rope of cloth or endless belt of

cloth at full width through the dye liquor. Pressurized-winch machines have been developed in the U.S. In an entirely new concept, the Gaston County jet machine circulates fabric in rope form through a pipe by means of a high-pressure jet of dye liquor. The jet machine is increasingly important in high-temperature dyeing of synthetic fibres, especially polyester fabrics.

Another machine, the jig, has a V-shaped trough holding the dye liquor and guide rollers to carry the cloth at full width between two external, powered rollers; the cloth is wound onto each roller alternately, that is, the cloth is first moved forward, then backward through the dye liquor until dyeing is complete. Modern machines, automatically controlled and programmed, can be built to work under pressure.

Padding mangles. Solutions or suspensions of colorants or their precursors may be padded onto piece goods by passing the cloth through a trough containing the liquor and then between rollers under pressure. Development and fixation processes such as steaming or dry-heat treatment can be carried out in other apparatus. The method is used in semicontinuous and continuous operations.

Garment-dyeing machines. Partly manufactured or finished garments, hosiery, and felt hats are dyed in drum or paddle type machines. In the rotating-drum machine an inner, perforated drum, rotating horizontally, contains the garments; the dye liquor and means of heating are housed in an outer, stationary drum. In the side-paddle machine, also known as a Gorrie, a vertical paddle wheel moves dye liquor and garments round an oval dyebath. To avoid entanglement, articles are placed in freely permeable bags. For hosiery dyeing and finishing, automated machines have been developed; stockings are held on metal formers onto which circulating dye liquor is sprayed.

Gorrie

Textile-printing equipment. The traditional methods of hand-block printing, batik, screen printing, and others are still employed, but the bulk of production is on roller-printing machines. Engraved rollers, one for each colour, apply dye paste to the fabric from the channels incised in the surface; surplus paste is removed from unmarked parts by a doctor blade, similar to that used in ink printing. Rollers are usually of copper or are copper-faced and may be chromium plated for durability. After receiving the imprint, the fabric enters a drying chamber, then a steamer where it is raised to the temperature at which fixation takes place. The fabric is then rinsed, soaped, rinsed again, and dried. Highly mechanized silk-screen printing is becoming widespread and is overtaking roller printing largely on cost grounds.

NONTEXTILE APPLICATIONS

Synthetic colorants are widely used to impart colour to materials other than textiles.

Paper. The coloration of paper is usually carried out by addition of colorants to the beater containing the cellulose pulp. The choice of colorants is determined by their cost, fastness, and the end-use requirements of the manufactured paper. Acid, basic, and direct dyes as well as pigments are used. Pigments are also applied as a surface coating to finished cardboard. Crepe paper is made by padding dye solution on to absorbent, white paper and then drying by a special process.

Leather. Acid, basic, direct, and mordant dyes are applied to leather by methods related to those of the textile dyer or by a technique that consists in spraying the leather with a pigment suspension or a dye solution.

Fur. Among the fur colorants are oxidation bases, reactive dyes, and acid dyes including azo–metal complexes. Woolled sheepskins are dyed with disperse dyes.

Oils. Oils are coloured by dyes that dissolve in solvents rather than those that dissolve in water. The oleate derivatives of certain triarylmethane dyes are also used, as is the fluorescent dye Fluorol 5G, C.I. Solvent Green 4, 45550.

Soaps. Colorants selected for tinting soap must have high strength and no tendency to stain skin or towels. Selected pigments and a few acid and basic dyes are used.

Food. The use of synthetic colouring agents in food is governed by law in many countries; certain products are prohibited. In general, nontoxic, water-soluble dyes proven to be noncarcinogenic, having extremely low arsenic and heavy-metal content, and designated as foodstuffs quality by dye makers are used.

Cosmetics. A few inorganic pigments are used to make eye cosmetics; for many years eosin, a red fluorescent dye, has been one of the colorants used in lipstick formulations. Hair dyes are chiefly in the class of substituted amines known as oxidation colours. Some countries regulate the types of colorants that may be used in cosmetics.

Printing inks. Modern organic pigments of good fastness are used (see PRINTING).

Plastics. High-grade modern pigments, able to withstand the high temperatures and severe conditions inherent in processing, are widely used in plastics of all kinds, including those for outdoor use.

Metal surfaces. Anodized aluminum is dyed, in sheets or as articles, with acid dyes, especially the 1:1 azo–metal complexes. Modern organic pigments of good fastness are used for automobile finishes.

Camouflage. Dyes and pigments fast to light and weather are required for camouflage. Their infrared spectral characteristics are important because by the use of infrared photography, an unsuitable green can be made to appear in sharp contrast to the vegetation with which it is a satisfactory match in ordinary daylight. Specialized, high-grade vat dyes have been developed for this purpose.

Coloured smokes. A number of relatively simple dyes made from anthraquinone sublime on heating to give strongly coloured smokes and are used for military, signalling, and other purposes. The dye Quinoline Yellow Spirit Soluble, C.I. Solvent Yellow 33, 47000, is also used.

Marker use

Fluorescent dyes. The powerful water-soluble dye fluorescein has long been used as a marker for tracing the course of underground streams and in sea-rescue operations. A range of basic dyes, in association with certain resins, for example toluene–sulfonamide–formaldehyde condensation products, give strikingly luminous effects and are applied in poster and display work. Fluorescent brightening agents, sometimes known as colourless dyes, selectively absorb ultraviolet radiation and emit light in the blue region of the spectrum; they are used in large quantities as ingredients in soap and washing powders, and also in papermaking.

Indicators. Some dyes such as phenolphthalein, methyl orange, and Congo red undergo a change in hue with change in acidity (pH); these are used in analytical and industrial laboratories as indicators in titrations. Vat Yellow 1 papers are used to test for hydrosulfite in vat dyeing, litmus papers for determining acidity and alkalinity.

Biological stains. A few natural dyes and some synthetic dyes are used for the differential staining of sectioned tissues for microscopic examination. Carmine, haematoxylin, eosin, and Safranine O are commonly used in this way.

Photography. There are four main uses of dyes in photography.

1. Sensitizing agents, chiefly cyanines and merocyanines, confer sensitivity on a photographic emulsion in certain spectral regions in which it would otherwise be absent.

2. Antihalation dyes, mostly modified triarylmethane dyes and cyanines of special structure, are painted on the back of the emulsion to prevent reflection of transmitted incident light and consequent spreading of the image.

3. Desensitizing agents such as Pinakryptol Green and special cyanine dyes.

4. Dyes for colour photography, for example, the azomethines, provide the subtractive primary colours yellow, magenta, and cyan, which, since each can absorb part of the daylight reflected by the others, may be combined to yield a maximum number of object colours.

Medicinal uses. Classical examples of dyes used in medicine include the mild laxative phenolphthalein; the

internal, bladder, and urinary-tract antiseptic methylene blue; and the surface antiseptics gentian violet and acriflavine.

STATISTICS ON MANUFACTURE AND USE

The production of synthetic dyes for the major producing countries for 1960 and 1970 is given in Table 2.

Table 2: Manufacture of Synthetic Dyestuffs*

country	production (metric tons)	
	1960	1970
United States	70,000	106,379
Germany, Federal Republic of†	56,234	104,563
Union of Soviet Socialist Republics	84,099	94,765
United Kingdom	40,200	52,600
Japan‡	26,625	52,403
Switzerland	...	§
France	14,396	24,610
Poland	9,783	20,980
Spain	...	16,754
Italy	18,258	14,100
German Democratic Republic‖	8,423	13,437
Czechoslovakia	5,163	9,616
Romania	3,695	9,312
Total¶	345,486	579,111

*Synthetic organic dyestuffs (including pigment dyestuffs), synthetic organic products of a kind used as luminophores, products of the kind known as optical bleaching agents, substantive to the fibre, natural indigo, colour lakes, in terms of 60 percent concentration basis. †Excluding luminophores, optical bleaching agents, natural indigo and colour lakes. ‡Excluding pigment resin colours and lakes. §In 1970 probably exceeded 40,000 metric tons. ‖Including lithopone. ¶Includes production and estimated production for other minor producers.
Source: *The Growth of World Industry, vol. II, Commodity Production Data*, editions of 1969 and 1971, UN, 1971 and 1973.

BIBLIOGRAPHY

History: Authoritative articles on dyes and dyeing from classical antiquity to the present may be found in the journal CIBA *Reviews* (1937–). The commemorative volume, *Perkin Centenary, London: 100 Years of Synthetic Dyestuffs* (1958), contains a chapter on the life and work of Sir William Perkin by J. READ. The article by C. PAINE surveys the development of the synthetic-dye-making industry.

Industrial dye chemistry: E.N. ABRAHART, *Dyes and their Intermediates* (1968); and H.E. FIERZ-DAVID and L. BLANGEY, *Fundamental Processes of Dye Chemistry* (1949, Eng. trans. from the 5th Austrian ed. of 1943), are textbooks written from the industrial viewpoint. Articles on special topics (*e.g.*, wool dyes, acetate silk dyes, etc.) may be found in the *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd ed., 22 vol. and supplements (1963).

Technology of dyeing: The standard reference work is the *Colour Index*, 3rd ed., 4 vol. (1971), and *Additions and Amendments* (annual), published jointly by the Society of Dyers and Colourists and the American Association of Textile Chemists and Colorists. Research papers and discursive articles are published in the *Journal of the Society of Dyers and Colourists* (1884–). Textbooks include: T. VICKERSTAFF, *The Physical Chemistry of Dyeing*, 2nd ed. rev. (1954), an advanced treatise on dye–substrate relationships; S.R. COCKETT, *Dyeing and Printing* (1964), and E.R. TROTMAN, *Dyeing and Chemical Technology of Textile Fibres*, 4th ed. (1970), good general introductions. The following specialized monographs are useful sources of information: S.R. COCKETT and K.A. HILTON, *The Dyeing of Cellulosic Fibres and Related Processes* (1961); C.L. BIRD, *The Theory and Practice of Wool Dyeing*, 3rd ed. (1963); H.U. SCHMIDLIN, *Preparation and Dyeing of Synthetic Fibres* (Eng. trans., 1963); EDWARD GURR, *Synthetic Dyes in Biology, Medicine and Chemistry* (1971).

(E.N.A.)

Dyestuffs and Pigments

Dyestuffs and pigments are intensely coloured substances used for imparting colour to other materials. Technically, dyestuffs are soluble in the medium in which they are applied, whereas pigments are insoluble.

Colour in materials originates either from organic or inorganic substances present in the material, or from a purely physical effect, such as that observed in oil films on water. Organic colouring matters are compounds of the element carbon, including animal or vegetable colouring matters, as well as synthetic material, like the coal-tar dyes. Colour in inorganic substances is commonly observed in rocks, minerals, ceramics, fired enamels, coloured glass, precious stones, and certain pigments. Colours produced by physical effects result from dispersion, refraction, or scattering of light. Examples of such physically produced colours are the rainbow; the blue of the sky; the iridescence of opal; and the blue, green, and violet colorations of birds, butterflies, fish, and insects. In some living organisms organic and physical coloration occur together.

Organic and inorganic colouring matters

There is scarcely any manufactured product that does not at some stage incorporate or require the use of products of the dyestuff and pigment industry. About 7,000 different dyestuffs are currently made, under some 35,000 different trade names, and about 200 new ones are introduced every year. The total annual world production is several hundred thousand tons, with a value of $2,500,000,000.

Reasons for the existence of the great number of commercial dyestuffs include the variety of fibres and other materials requiring coloration; also the fact that colour in textiles must withstand a variety of stresses, depending upon the methods of manufacture and use.

All dyestuffs and many pigments are complex organic compounds, mainly prepared synthetically from chemical products of the coal-tar and petrochemical industries. The syntheses consist typically of sequences of carefully regulated procedures carried out under strict scientific control; they are the result of extensive research. Dyeing itself is mostly carried out by means of water solutions, and dyestuffs are applied to textiles, paper, leather, and many other substances. Pigments are used as finely ground solid particles rather than as solutions. They are employed, for example, in paints, printing inks, and plastics. Some pigments are entirely inorganic in origin; *e.g.*, iron oxide. These are cheap but rather dull in hue. The brighter pigments are organic compounds, and in fact may simply be dyestuffs insoluble in the medium. Many products are thus used in both capacities; *i.e.*, in soluble form as dyes for dyeing textiles, paper, and other materials and in insoluble form as pigments for colouring paints, inks, and the like.

All organic coloured substances, whether the natural materials found in the animal and vegetable worlds or synthetic dyestuffs, rely for their colour on certain aspects of their chemical structures. As a class, organic dyestuffs and pigments, with which this article largely deals, are known as colorants.

HISTORY

Development of natural dyestuffs. Some thirty centuries ago Moses was instructed to accept gifts of "blue, and purple, and scarlet and fine linen and goats' hair." These names of colours most probably referred to materials dyed with dyes known as indigo, Tyrian purple, and kermes. Indigo was extracted from plants of the *Indigofera* genus, Tyrian purple from mollusks of the genera *Murex* and *Purpura*, and kermes from a dried insect *Kermes ilicis*. The Bible, therefore, illustrates the antiquity of the craft of dyeing and the variety of natural products used in olden times as sources of dyestuffs. Dyed fabrics from even earlier periods discovered in Egyptian tombs show that dyeing was practiced at least as early as the 25th century BC. It is believed that the craft originated in India (the name indigo being derived from the Greek word for Indian, *indikos*) and spread westward to Persia, Phoenicia, and Egypt.

The method of producing a variety of hues by mixing red, blue, and yellow dyestuffs was known in ancient times. Fabrics found in caves in the Dead Sea area and dating from AD 135 were found to have been dyed with various mixtures of safflower yellow on alum mordant (fixative), kermes on alum, and indigo blue.

Until the 19th century all dyestuffs were natural products, obtained mainly from vegetable sources. A wide variety of plants was used. The craft of dyeing demanded a high degree of skill, and details of its recipes were largely secret. Some of the most important natural dye-

stuffs that were in use until the advent of synthetic products are shown in Table 1.

With few exceptions these natural dyestuffs have no affinity for cellulose, and thus they could be applied to cotton only with the aid of an auxiliary inorganic chemical or mordant (Lat. *mordere*, "to bite"), which precipitated the colouring matter in a less soluble form in the fibre. Mordants were also used on wool and silk, both to improve fastness and to obtain a wide variety of hues, different mordants yielding different hues with the same dye. The mordants in common use in ancient times were mainly alum (potassium aluminum sulfate) and salts of iron, copper, or tin. Salts of chromium came into use about 1850. Eventually, on account of the superiority in fastness of the dyeings they produce, chromium mordants almost entirely displaced all other metal salts. Chromium mordants are still widely used for dyeing wool and to a lesser extent for silk and nylon, but are no longer of importance for cellulose fibres.

Ancient use of mordants

Table 1: Sources of Some Important Natural Dyestuffs

Yellow	
Persian berries*	the fruit of a *Rhamnus* species
Weld*	the leaves, seeds, and stem of the plant *Reseda luteola*
Fustic*	the wood of *Morus* or *Maclura tinctoria* (Central or South America)
Quercitron bark*	from *Quercus nigra* or *tinctoria* (U.S.)
Brown	
Catechu or cutch*	acacia, etc., wood (Far East)
Reds	
Kermes	a dried insect, *Kermes ilicis* (Mediterranean region)
Lac dye	the shellac-producing insect *Coccus lacca* (Far East)
Cochineal*	the dried female of the insect *Coccus cacti* (Mexico, etc.)
Archil and cudbear	prepared by fermentation of various lichens, *e.g.*, *Roccella tinctoria*
Madder (Turkey red)*	the root of *Rubia tinctorum* (Asia Minor)
Safflower	the dried flower head of the Dyer's Thistle (*Carthamus tinctorius*)
Blue	
Indigo	prepared from the leaves of various *Indigofera* plants, also the *Isatis tinctoria* or woad plant
Violet	
Tyrian purple	prepared from various mollusks; *e.g.*, *Murex brandaris*
Black	
Logwood*	the wood of the West Indian tree *Haematoxylon campechianum*†

*Used with mordants. †Before the discovery of America, blacks were prepared from tannins (*e.g.*, from oak galls or sumac) and salts of iron.

Following the introduction of synthetic dyes in the 19th century, natural dyestuffs have steadily declined in importance. Now only one, logwood, employed to dye nylon black, is used in any significant quantity.

Development of the chemistry of dyestuffs and dyeing. "The Liquors that Dyers imploy to tinge," wrote Robert Boyle, an English physicist, in 1664 in the first book dealing with the science of coloration, "are qualified to do so by multitudes of little Corpuscles of Colour . . . insinuating themselves into, and filling all the Pores of the Body to be Dyed, ..." Even today this remains a good description of the chemistry of dyeing.

As early as the mid-17th century, therefore, leading scientists were concerned with the processes of dyeing, and one of the first actions of the Royal Society after its foundation was to assist dyers by publication of a paper, "An Apparatus to the History of the Common Practices of Dyeing," read to them by Sir William Petty in 1662, describing the methods of dyeing then in use.

Petty, in his paper, classified dyestuffs into the fundamental colours, red, yellow, and "blew," and discussed the use of alum as a mordant, suggesting that it first is taken up by the fibre, where it then becomes combined with the dye. Thus, the dye is fixed in the fibre. This is also the modern view.

From that period onward the craft of dyeing was aided by the discoveries of chemists and, especially in the period 1670–1830, by a series of French chemists, whose interest appears to have stemmed from the action of Louis XIV's comptroller general of France, who, in order to establish high standards in the wool textile industry, laid down regulations controlling dyeing methods. Distinction was made at that time between *bon teint* and *petit teint*, or fast and nonfast colours; *i.e.*, dyeings that were fast or fugitive, respectively, to light and washing. When in 1729 the regulations were redrafted, the first system of controlled testing of fastness of dyed textiles was introduced.

Early chemical interest in dyeing

During the early part of the 19th century, dyeing continued to attract some scientific notice. Even greater interest was created by the advent of synthetic dyestuffs. William Henry Perkin's discovery (1856) of mauve, the first commercial synthetic dyestuff (see below), and the subsequent development of a great variety of dyestuffs of known chemical constitution initiated the modern development of the subject into a distinct branch of chemistry. This development has been particularly noteworthy since World War I. In addition to the availability of a continually growing number of dyestuffs, the reasons for this development include improved instrumentation, increased knowledge of the structure of fibres, and the application of classical chemical principles to the dyeing process.

The introduction of synthetic dyestuffs challenged organic chemists with the problem of determining the molecular structure of these complex materials and opened the way to the development of theories of the causes of colour in organic compounds in general. The idea of valence (fixed numbers of bonds for atoms of different kinds), and in particular the quadrivalence of carbon, had already been formulated when the new dyestuffs began to appear. Shortly afterward the ring structure of the benzene molecule and the three-dimensional nature of organic molecules in general were suggested. These advances greatly assisted in the understanding of the molecular structures of organic dyestuffs.

One of the most important contributions to the understanding of the relationship between colour and structure was the so-called "chromophore" theory proposed in 1876 by a German chemist, Otto N. Witt. In every dye, Witt identified a chromophoric ("colour-bearing") group together with one or more "auxochromes"; *i.e.*, associated groups whose function was to intensify the colour. In 1888 an English chemist, Henry Edward Armstrong, showed that the chromophores could generally be depicted as quinones (oxidized structures derived from the benzene ring). These theories are substantially those held today, although they have been broadened by recognition that the function of the chromophore is to produce a strong absorption of radiation, either in the ultraviolet or in the visible region of the spectrum, and that of the auxochrome is to shift the absorption to give more intense coloration (see below *Dependence of properties upon molecular structure*).

Chromophores and auxochromes

The development of the quantum theory in the 20th century also brought greater insight into the mode of interaction between light and organic molecules. In current terminology, the absorption of visible radiation by a dye raises the electronic energy of the molecule to a so-called "excited state." The energy must then be dissipated either as heat, as phosphorescence or fluorescence, or in a chemical reaction with the surroundings, in which case the dye decomposes and loses its original colour.

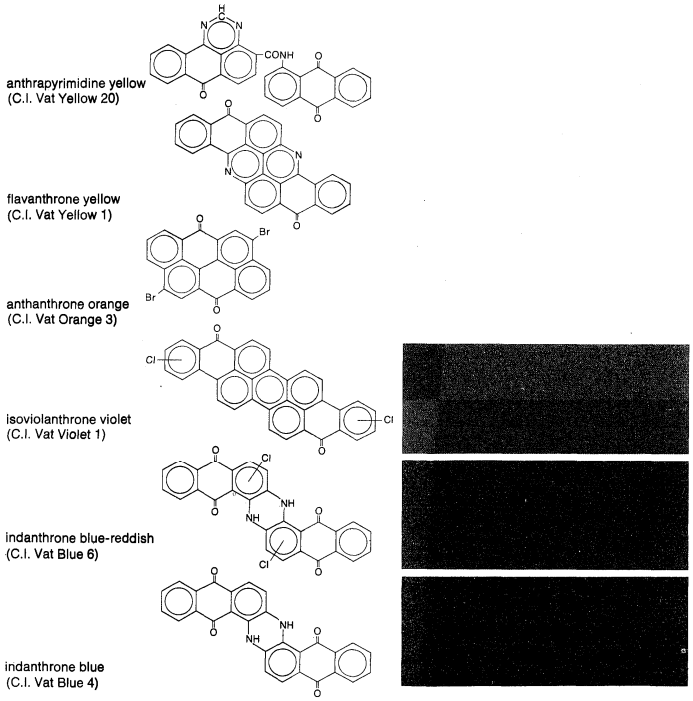
Development of synthetic dyestuffs. In 1856 William Henry Perkin, a chemistry student in London, discovered a purple colouring matter that had dyeing properties in a product he obtained by chemical treatment of aniline, a material from coal tar. This chance discovery, made in a simple laboratory Perkin had fitted out at his home, laid the foundation of the manufacture of synthetic dyestuffs and pigments, and indeed of the entire world-wide synthetic organic chemical industry.

The discovery of mauve

With foresight and courage, Perkin, in 1857, established a factory to manufacture the patented dye, at first called Aniline Purple, or Tyrian Purple, and later Mauve, a French word for the mallow, a plant with a purplish flower. Perkin, moreover, developed new methods of applying the dyestuff and visited dyers to show them how they should be used.

Mauve was an immediate success and the general method of its preparation, the oxidation of aniline, was varied and elaborated by other investigators to produce many new dyestuffs, first the crimson coloured Fuchsine or magenta, and later, blues, violets, and green. These were the first basic or cationic (positively charged) dyestuffs, a class that later became overshadowed in importance by others, though in recent years interest in it has revived. Because of their origin, these earliest synthetic dyestuffs were called aniline dyes, a term still used. Unfortunately, their somewhat fugitive nature caused all synthetic dyestuffs to be considered inferior to the natural products, though this is certainly not the case.

From *Journal of the Oil and Colour Chemists Association* (January 1963)



Colours obtained from various anthraquinone pigments. By modifying the benzene rings of the anthraquinone group, a range from yellow to purple can be produced. A great many other derivatives of anthraquinone are used for industrial purposes.

In 1868 a team of German chemists proved that alizarin, the colouring principle of madder root, is a derivative of the chemical compound anthraquinone. (The Figure shows the colours yellow to purple obtained with pigments manufactured by modifying the molecular structures of anthraquinone compounds.) Their synthesis of alizarin from anthraquinone was the first preparation of a natural vegetable dyestuff by purely chemical means. With the introduction of synthetic alizarin, the use of the natural madder root died out. But, at the same time, new and related products were introduced. The technique of sulfonation (that is, the introduction of sulfonic acid groups), developed earlier, was applied from 1893 onward to the development of a new class of valuable bright, fast dyestuffs for wool, the sulfonated anthraquinone derivatives. This class is one of the most important for wool, and the unsulfonated counterparts are used to dye modern man-made fibres.

Meanwhile, an entirely new type of colouring matter, the azo dyes, which now comprise the largest group of dyestuffs, was being developed. Johann Peter Griess, a young German, then working in the laboratory of an English brewery, discovered the diazonium compounds in 1858. These are very reactive substances, formed from the reaction of aromatic amines (nitrogen-containing organic compounds) with nitrous acid. Diazonium salts readily combine with aromatic amines or phenols to give deeply coloured substances containing the azo group (two nitrogen atoms joined by a double bond). The discovery was not, however, fully exploited until the 1870s, when the full tide of development of the azo dyestuffs set in, stimulated by the discovery that the addition of sulfonic acid groups gives highly water-soluble materials with excellent affinity for wool. By treating naphthalene derivatives with sulfuric acid, various sulfonated aminonaphthalene and hydroxynaphthalenes were made, from which azo dyestuffs could be prepared in a range of shades—from orange to black. And by use of derivatives of pyrazolone instead of naphthalene, yellow azo dyestuffs also were prepared. In 1884 Congo red was produced; this is a bisazo dyestuff; *i.e.*, one containing two azo groups in its molecules. Congo red was the forerunner of a long series of successful dyestuffs, still of importance, named "direct cotton" or "direct" dyestuffs, which can dye cotton without a mordant.

The azo dyestuffs

Indigo for a time was the sole example of a "vat" dyestuff; *i.e.*, one applied from a "vat" containing alkali and a reducing agent in which the insoluble dyestuff was temporarily solubilized during dyeing. In 1901, however, Indanthrene blue, the first of the modern type of vat dyestuff, based on anthraquinone, was introduced; and this class, including dyestuffs of the very greatest fastness, has become of great importance.

Vat dyestuffs

Four other important new classes of dyestuffs were introduced in the 20th century:

1. The azoic dyestuffs. As early as 1880, a substance called β-naphthol had been applied in alkaline solution to cotton and then coupled with a diazonium compound to produce an insoluble dyestuff in the cotton fibre. Little progress was made with this type of dyeing, however, until Naphthol AS (a more complex chemical compound derived from β-naphthol) was discovered in 1912 and used to replace β-naphthol. Later a whole series of such compounds—and new bases to be diazotized and coupled with them—appeared, giving bright, fast dyes, ranging in colour from yellow to black, but especially useful for oranges and reds.

2. The disperse or nonionic dyestuffs. These substances were discovered in 1923 and at first were used with the then new cellulose acetate fibre, but later also used for nylon and polyester fibres.

3. The phthalocyanine dyestuffs and pigments. Discovered originally in 1907 and 1927 by European chemists, who failed to realize their importance, phthalocyanine derivatives (a class of metal-containing organic compounds) were rediscovered by accident in Scotland in 1928. The first member of this class, a brilliant blue, was found to have a remarkable molecular structure, resembling those of chlorophyll and hemoglobin, natural colouring matters. It proved to be a pigment of outstanding properties. Chlorinated and brominated derivatives gave bright green dyes that together with the parent compound are used in very large quantities mainly as pigments; modified products are used for dyeing. Although no other useful shades have been obtained from phthalocyanine derivatives, quinacridone (a red compound discovered in 1955) is capable of preparation in red or violet forms as a pigment of fastness comparable to phthalocyanine.

4. The reactive dyestuffs for cellulose. First introduced in 1956, these dyes represent an entirely new principle, in that the dyestuff becomes a part of the actual fibre molecule and, therefore, is highly resistant to washing. Unlike other fast dyestuffs for cellulose, many reactive dyestuffs have very bright shades. Later, reactive dyestuffs adapted for wool and nylon were introduced.

The first reactive dyestuffs

DEPENDENCE OF PROPERTIES UPON MOLECULAR STRUCTURE

Much of the following discussion depends upon knowledge of the properties and behaviour of atoms and molecules. (Detailed information may be found in the articles

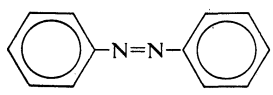
CHEMICAL BONDING; CHEMICAL REACTIONS; CHEMICAL COMPOUNDS, ORGANIC; and MOLECULAR STRUCTURE.)

To be useful, dyestuffs and pigments must have: (1) intense colour and (2) fastness; *i.e.*, resistance to the various chemical and mechanical stresses they meet in the manufacture or use of the finished coloured product. In addition, dyestuffs, but not pigments, must have: (3) solubility in the medium of application, which in almost all cases is water, and (4) ability to be adsorbed and retained by the fibre (substantivity) or to be chemically combined with it (reactivity). Pigments also have to meet unique requirements somewhat different from those of dyestuffs, including such physical and mechanical properties as flow properties in the media of application and insolubility in oily media. These properties are favoured by making the molecule large, and insolubility is also promoted by particular dispositions of the substituent groups—that is, the peripheral units attached to the main body of the molecule.

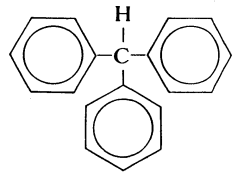
In all cases these properties depend in large part upon the molecular structures of the colorants themselves. The relationship between the properties of the dyestuffs and pigments and their molecular structure is dealt with in the following sections.

Colour. Organic colouring matters are complex unsaturated compounds—that is, compounds with multiple bonds between the atoms—and each such compound has a molecular structure containing one of the fundamental chromophores. Three of the most important chromophores in dyestuffs are called azobenzene, triphenylmethane, and anthraquinone. Their molecular structures are as follows:

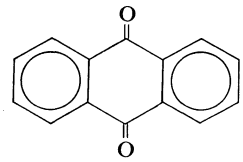
Structures of chromophores



azobenzene



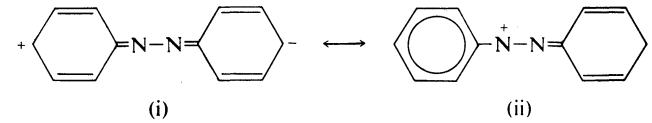
triphenylmethane



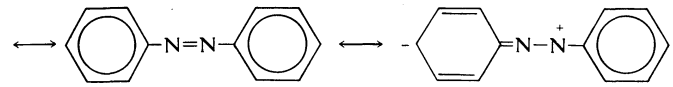
anthraquinone

in which the letters N, C, H, and O stand for atoms of nitrogen, carbon, hydrogen, and oxygen, respectively; the short single and double lines represent single and double bonds between them, and the hexagons with circles in them represent benzene rings—circular units each composed of six carbon atoms. The chromophoric compounds themselves usually have very weak colours because the conjugated chain (the system of alternate single and double bonds, including the benzene ring) in the molecules is relatively short. If this chain is greatly extended, as in the natural carotene pigments, for example (see below), deep colour develops. Such extended chains, however, are not conveniently introduced into synthetic dyestuffs, which instead rely upon the above-mentioned substituent groups called auxochromes for colour intensification. Auxochromes generally are polar substituents —that is, atoms or groups favouring separation of electrical charges—including such substituents as chloro, nitro, hydroxy, methoxy, amino, methylamino, dimethylamino, and trifluoromethyl groups.

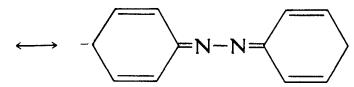
In the conjugated systems of alternate double and single bonds of the chromophores, one or more of the mobile electrons, called pi electrons, are able to move through the molecule. It is then possible to write various electronic structures of the molecule; *i.e.*, the so-called resonance forms. None of these structures describes the molecule as it really exists, but the molecule may be considered as a hybrid to which the imaginary structures contribute. Thus forms i–v below represent some of the forms contributing to the hybrid structure of azobenzene. Of these, iii is the most stable, because no energy is required to separate charges (as in the other forms).



(i) (ii)



(iii) (iv)

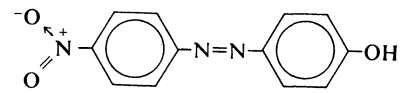


(v)

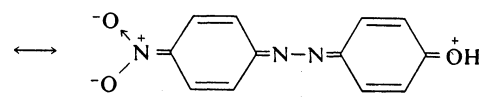
In this representation, double-headed arrows are used between resonance forms of a single hybrid structure.

The function of the auxochromic groups is to increase the stability of the alternative resonance forms. They do so because they can retain electrical charges more readily than carbon atoms can. For example, when two auxochromic groups are attached to an azobenzene molecule, the resulting structure may consist of new hybrid forms, as in the example below:

Influence of auxochromes



(vi)



(vii)

In this instance the alternative electronic configuration vii is more stable than i, ii, iv, or v above, because the charges are now held on oxygen and nitrogen atoms, which can retain them more readily than carbon. The result is that this compound has a more intense colour than azobenzene does.

With the increase in stability of the contributing electronic structures of a molecule, its electrons may be considered to move more readily along the chromophore. Thus its natural frequency of vibration is decreased. A simple analogy can be drawn with a violin string: the longer the string, the lower is the frequency and the longer is the wavelength of the note it emits when plucked. In general, the wavelength increases with the length of the oscillator. This applies to the absorption as well as the emission of energy, whether it be sound energy or light energy, because any oscillator absorbs energy most readily at the wavelength of its natural frequency of vibration.

The chromophores described above absorb most energy at short wavelengths—that is, in the ultraviolet region of the spectrum; and consequently they appear colourless or have only a faint yellow colour. When an auxochrome is present the wavelength of absorption increases and moves into the visible region, thus causing colour to appear. For example, anthraquinone is almost colourless, but 1-methylaminoanthraquinone has an intense red colour. The resulting hue is said to be "deeper."

Table 2 shows the changes that occur in the colour of a given chromophoric system with the progress of its main absorption band across the spectrum from short to long wavelengths. Such a succession of changes can often be produced by substitution of the chromophore with a

Table 2: Effect of Wavelength of Absorption on Colour Observed

region of maximum absorption*	colour seen†
Ultraviolet	none
Violet	yellow
Blue	orange
Green	bluish-red
Yellow	violet
Orange	blue
Red	blue-green

*Listed in order of increasing wavelength. †Listed in order of increasing depth of hue.

Light absorption and colour

succession of selected auxochromes, one or more at a time, if these auxochromes have increasing effectiveness in enhancing the mobility of the electrons of the chromophoric system. In the case of azo dyestuffs, for example, the hue can be deepened by substitution in either of the aromatic nuclei attached to the azo group, and if both are substituted, the effect is further enhanced.

The intensity of the colour of dyestuffs is among the highest found in any type of substance. The degree of this intensity, known as the molar extinction coefficient, is measured at the wavelength of maximum absorption of light. The molar extinction coefficient for dyestuffs is about 10^4. In practice, this means that a 1 gram percent solution of a dyestuff only 1/100th of a millimetre deep reduces the intensity of the most strongly absorbed light to about one-half of the original value. There is another way to visualize the colour intensity of dyestuffs; it requires from about 10 to 100 single layers of dyestuff molecules, placed one on top of another, to produce a visible colour; and, allowing for the minimum area the substance must have in order to be seen by the naked eye under the best conditions, a total of about 10,000 molecules, or about 10^{-18} gram, must be concentrated together before the colour becomes visible.

Fastness. Ability to withstand washing without changing colour, "wash fastness," and ability to withstand exposure to light, "light fastness," are two of the properties of most importance to the user of coloured fabrics. Dyestuffs that are chemically bonded to the fibre—*i.e.*, reactive dyestuffs—or those that are present as water insoluble particles of low solubility in soap solutions, such as vat dyestuffs, have the highest wash fastness.

Fastness to light and to washing

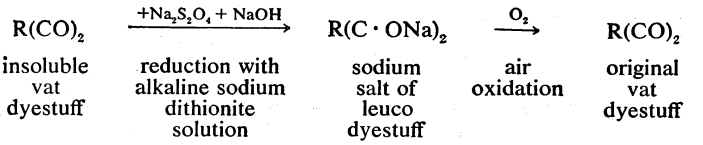
Water-soluble dyestuffs often have lower wash fastness, but the resistance to removal by soap varies greatly from one dyestuff to another, fastness generally being aided by high molecular weight of the dyestuff molecule and a minimum number of solubilizing groups.

All organic colouring matters fade when exposed to light; hardly any inorganic colours, and no physically produced colours, do. Light fading is the most complex of the reactions that dyestuffs undergo during use, and a great deal of research has been devoted to discovering its causes, which are not yet fully understood.

The light fastness of any dyed material depends not only on the nature of the dyestuff itself, but also on many other factors, including the method of application, the depth of shade, the type of fibre and its moisture content, the relative humidity and oxygen content of the air, and the nature of the illumination.

Solubility. Because normal dyeing takes place from aqueous solutions, the dyestuff must contain substituent groups conferring solubility in water. Table 3 includes the most important solubilizing groups used in dyestuff molecules.

Solubility in water may be needed only temporarily during the actual dyeing operation, the dye later being treated so that it becomes water-insoluble on the fibre. Such temporary solubility may be obtained with an insoluble, nonionic dyestuff molecule containing quinone groups, for example, by chemical reduction (addition of electrons) in the presence of alkali to produce a salt of the corresponding hydroquinone; the original quinone is then regenerated on the fibre after dyeing by oxidation. This reduction–oxidation procedure is the basis of the vat dyeing process. Usually sodium dithionite (also called sodium hydrosulfite, formula $Na_2S_2O_4$) is the reducing agent used, and sodium hydroxide is the alkali, with the result that the sodium salt of the reduced form of the dye, known as the leuco compound, is formed. The entire process of alkaline reduction is known as vatting. The leuco compound is generally unstable in air, being rapidly oxidized back to the original dye. This sequence of chemical reactions involved may be shown by the following equations:



In the method of dyeing cellulose fibres known as azoic coupling, a single soluble component, usually a naphthol, is applied to the fibre from solution in sodium hydroxide; this component is then converted into a water-insoluble azo dyestuff on the fibre by coupling with a soluble diazotized base. The procedure is illustrated in the following reaction sequence:

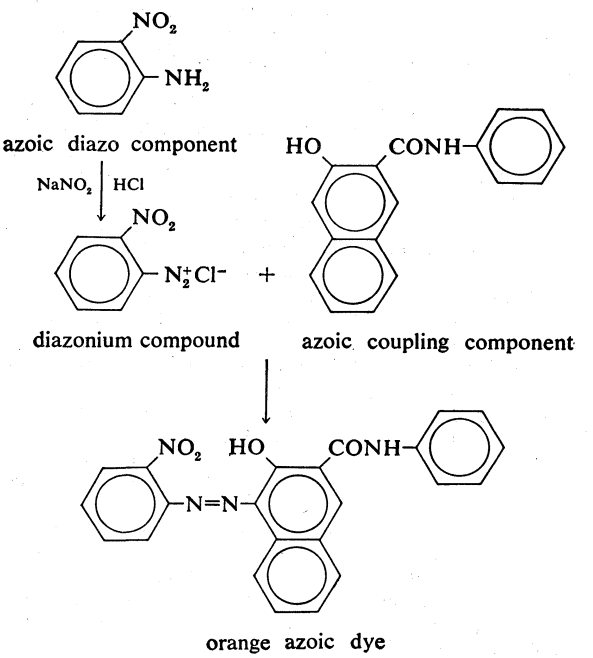


Table 3: Solubilizing Groups Used With Various Classes of Dyestuffs

name	structure*	type of dyestuff
Permanent		
Sodium sulfonate	$-SO_3Na$	direct cotton, acid wool, chrome mordant; 1:1 dye–metal complex for wool
Amine hydrochloride	$-\overset{+}{N}H_3Cl^-$, $-\overset{+}{N}R_3Cl^-$	basic, for cellulose, wool, silk, and acrylic fibres
Hydroxy, amino, and sulfamido groups	$-OH$, $-NH_2$, $-SO_2NH_2$	disperse, for cellulose acetate, nylon, polyester fibres; 2:1 dye–metal complexes for wool and nylon
Temporary		
Sodium phenolate	$-O^-Na^+$	certain classes of dyestuffs for cellulose; *i.e.*, naphthols for subsequent azo coupling on the fibre; and vat dyestuffs
Isothiouronium groups	$-CH_2SC(NR_2)=\overset{+}{N}R_2\ Cl^-$	phthalocyanine dyestuffs for cellulose

*R = an alkyl or aryl group.

The so-called disperse dyestuffs used to colour cellulose acetates, nylon, and polyester fibres are employed as dispersions—that is, as fine suspensions of the solid in liquid—rather than as true solutions. These substances, however, are slightly soluble in water due to the presence of several nonionic water-attracting substituent groups on their molecules, and dyeing probably takes place from the very dilute solutions they produce.

There is considerable interest in processes for dyeing certain man-made fibres with disperse dyestuffs from solutions in organic solvents instead of water. Another process also being considered is the use of vaporized dyestuffs, at high temperature (above 200° C, 392° F).

Attachment. Many solid materials adsorb dyestuffs from their solutions, but fibres do so particularly well because of their porosity and their ability to exert chemical forces upon the dyestuff molecules. There is an immense number of submicroscopic pores in a fibre, mainly directed parallel to the length of the fibre; there are, in fact, about ten million in the cross section of a normal textile fibre. The total surface of the walls of these pores is extremely high, and in natural fibres, such as cotton or wool, the pore surface amounts to no less than about five acres per pound. This value is about one thousand times as great as the external surface of the fibres.

Porosity and surface area

Since 1,000 to 10,000 single layers of dyestuff molecules placed one above another are needed to produce a deep coloration, a single layer of dyestuff molecules adsorbed on the walls of all the pores of a fibre would be more than enough to produce a satisfactory depth of shade. Actually, a large proportion of the pores are too small to allow a dyestuff molecule to enter, but, even so, enough dye is adsorbed to produce satisfactory depth of shade. For this action to occur, however, forces on the molecular scale must operate between the fibre molecules exposed in the pore surfaces, and the dyestuff molecules in the contiguous solution. The strongest of these forces are the forces of true chemical bonding by which the so-called reactive dyestuffs are fixed. Other types of dyestuff are attracted by a variety of weaker forces that operate between molecules. As already mentioned, some types of dyestuff are solubilized only during the dyeing operation; in such cases, after the subsequent insolubilization, the dyestuff particles remain held in the fibre by purely mechanical action.

With regard to the dyestuff molecule, attachments to the fibre resulting simply from intermolecular attraction are considered substantive, whereas those involving chemical bonding are considered reactive.

Substantivity. The presence of one or more of certain specific groups in a dyestuff molecule determines its substantivity, or affinity, for any given type of fibre, and hence its dyeing method. These groups include the following:

1. Anionic or cationic groups. These (respectively) negatively- or positively-charged groups confer solubility on dyestuffs and also affinity for the appropriate fibres with ionic properties.
2. Polar groups. Although these substituents are electrically neutral they do contribute to charge separation within a molecule. This effect increases the affinity of nonionic (disperse) dyestuffs for man-made fibres and considerably influences the colour.
3. Groups that induce flatness (planarity) in the molecule as a whole. Substituents promoting planarity are essential in producing useful affinity in dyestuffs for cellulose acetates, or polyesters.
4. Lengthy conjugated chains. If a dyestuff molecule is planar, its affinity for cellulose fibres increases regularly with the length of its chain of alternate single and double bonds. To have useful affinity it should have at least eight conjugative double bonds.
5. Phenolic groups. When situated in dyestuff molecules adjacent to certain other groups, phenolic groups confer the ability of combination with a metal mordant.
6. Quinone groups. In the absence of permanent ionic groups, quinones enable a dyestuff to be used for dyeing cellulose fibres by temporary solubilization.
7. Primary amino groups. Primary amino groups on the bases used for azoic dyestuffs can be diazotized and coupled with naphthols (or other materials) already adsorbed on the fibre.
8. Paraffinic chains. Saturated hydrocarbon (paraffinic) chains are used in a few members of several classes of dyestuff. For example, when attached to an acidic molecule for dyeing wool, a paraffin chain of moderate length (containing four to 12 carbons) improves fastness to washing and other wet treatments. Also, when inserted between the polar group and aromatic nucleus of a disperse dyestuff molecule, a short (two- or three-carbon) chain improves the fastness to dry heat. And when inserted between a quaternary ammonium group and an aromatic nucleus, such a chain appears to improve the light fastness of cationic dies applied to polyacrylonitrile fibres.

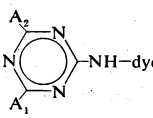
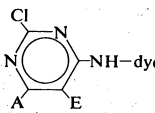
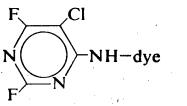
In an overall sense, a dye molecule that is either small in total volume, or slim in its cross section, makes rapid dyeing possible because of its ready accessibility to the pores of the fibres.

Reactivity. The reactive dyestuffs rely for their retention on the fibre on a principle different from that which controls the dyeing of most other types of dyestuff; that is, they form a covalent (nonionic) chemical bond with the fibre molecules. Such bonds are formed by a reaction between a specific chemical substituent group in the dyestuff molecule with part of the molecule of the fibre.

Formation of chemical bonds

The reactive groups associated with several classes of dyestuffs are shown in Table 4. Other groups also are employed, and new types of groups are introduced from time to time.

Table 4: Types of Reactive Groups in Dyestuffs

class	chemical structure	
Reactive dyestuffs for cellulose		
Triazine		A_1 = aryl or Cl; A_2 = Cl
Pyrimidine		A = Cl; E = H or Cl
Vinyl derivative	$HO_3SOCH_2CH_2-X-$dye	X = $-SO_2-$, $=-CONH-$, $=-SO_2NH-$
Reactive dyestuffs for wool		
Acrylamide	$CH_2=CHCONH-$(dye-metal complex) $CH_2=C(Br)CONH-$dye	
Vinylsulfonyl	$NaO_3SOCH_2CH_2SO_2-$dye	
Trihalopyrimidine		

In summary, the molecule of any colouring matter can be viewed as a structure built up step by step, each step designed to give the molecule certain properties and enable it to fulfill certain functions. In Table 5 details of chemical structure that confer the above listed properties on the molecule are presented, and the molecular constitutions of some typical important colouring matters, chosen from nature and from industry, illustrate these principles.

TECHNOLOGY

In this section will be found limited aspects of the classification, manufacture, use, and testing of dyes (for a de-

Table 5: Typical Dyestuffs and Pigments, Natural and Synthetic

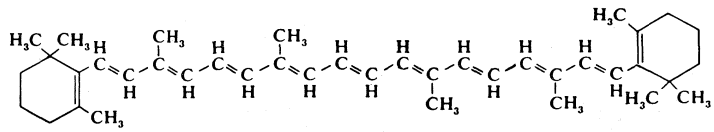
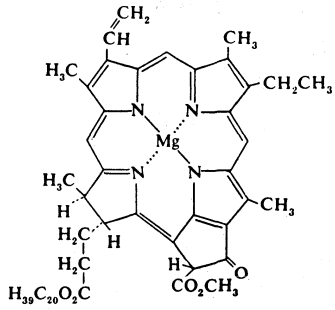
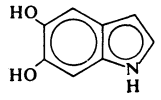
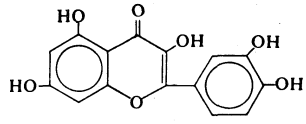
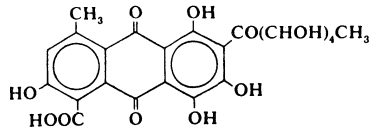
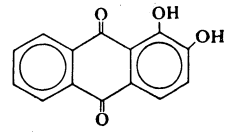
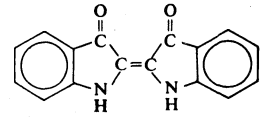
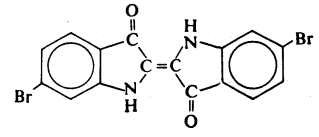
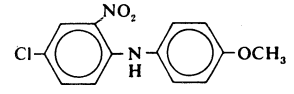
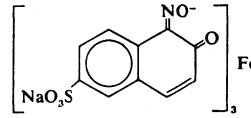
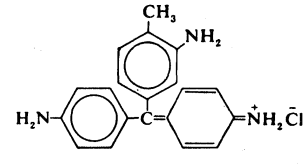
classification and name	*Colour Index* references	chemical structure	uses and properties
Natural biological colorants			
β-carotene	C.I. Natural Yellow 26, C.I. 75130		for fats and oils
Chlorophyll *a*	C.I. Natural Green 3, C.I. 75810		for soaps, oils, perfumes, foodstuffs
Melanin	C.I. Natural Brown 9	 complex; derived from 5, 6-dihydroxyindole	artists' watercolour
Derived biological colorants			
Quercetin (from quercitron bark)	C.I. Natural Yellow 10, C.I. 75670		with mordant, for cellulose fibres or wool
Cochineal (from dried female insects, *Coccus cacti*)	C.I. Natural Red 4, C.I. 75470		with mordants, for wool and silk
Alizarin (from root of madder plant)	C.I. Natural Red 8, C.I. 75330		with mordants, for cellulose fibres, silk, and wool
Indigo	C.I. Natural Blue 1, C.I. 75780, 75790 (C.I. Vat Blue 1, C.I. 73000)		for cellulose fibres and wool
Tyrian purple	C.I. 75800		artists' pigments
Synthetic colorants			
Nitro Amacel Golden Orange III, SRA East Golden Orange III, etc.	C.I. Disperse Orange 15 C.I. 10350		for cellulose acetate and nylon
Nitroso Naphthol Green B, Calcocid Green B, etc.	C.I. Acid Green 1, C.I. 10020	 (as iron complex)	for wool
Triphenylmethane Magenta	C.I. Basic Violet 14, C.I. 42510		for cellulose fibres and miscellaneous materials

Table 5: Typical Dyestuffs and Pigments, Natural and Synthetic (continued)

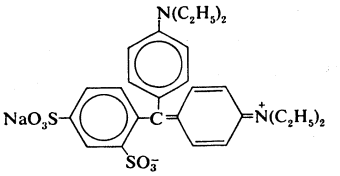
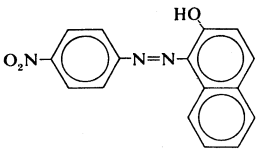
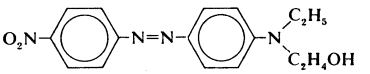
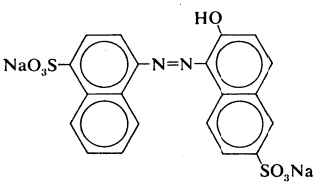
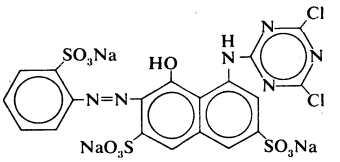
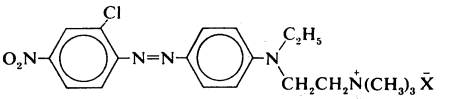
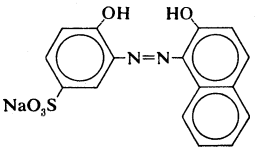
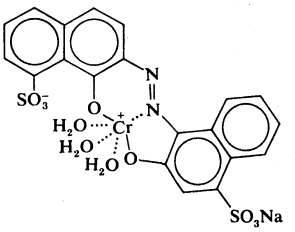
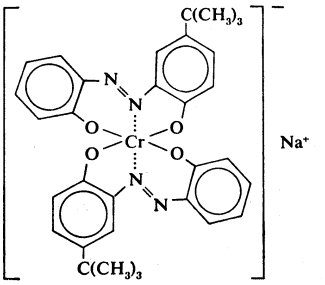
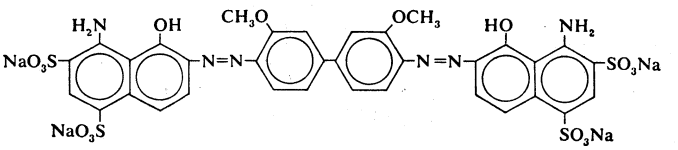
classification and name	*Colour Index* references	chemical structure	uses and properties
Patent Blue V, Kiton Pure Blue V, etc.	C.I. Acid Blue 1, C.I. 42045		for wool, silk, and miscellaneous materials
Azo Para Red, Lake Red, Federal Red, etc.	C.I. Pigment Red 1, C.I. 12070		for miscellaneous materials
Cibacet Scarlet 2B, Celliton Scarlet B, etc.	C.I. Disperse Red 1, C.I. 11110		for cellulose acetates and nylons
Fast Red E, Naphthalene Red EA. etc.	C.I. Acid Red 13 (C.I. Food Red 4), C.I. 16045		for wool and silk (permitted for food in U.K.)
Red reactive dye			for cellulose fibres
Astrazon Red GTL	C.I. Basic Red 18		for acrylic fibres
Chrome Violet B, Acid Alizarin Violet, etc.	C.I. Mordant Violet 5, C.I. 15670		with chromium mordant, for wool
Palatine Fast Blue GGN, Ultralan Blue GG, etc.	C.I. Acid Blue 158, C.I. 14880		for wool
Neutral dyeing metal-complex dye			for wool or nylon
Niagara Sky Blue 6B, Direct Sky Blue Green Shade, etc.	C.I. Direct Blue 1, C.I. 24410		for cellulose fibres, silk, and nylon

Table 5: Typical Dyestuffs and Pigments, Natural and Synthetic (continued)

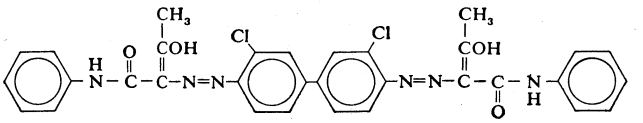
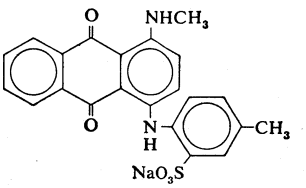
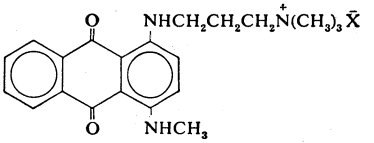
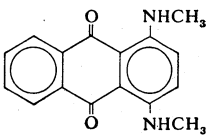
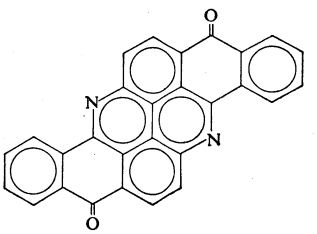
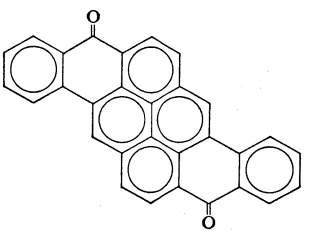
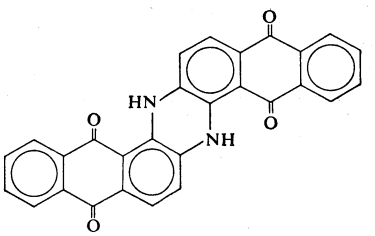
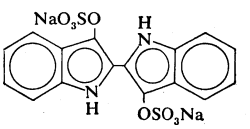
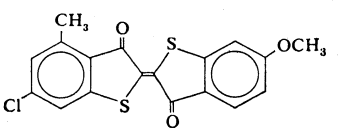
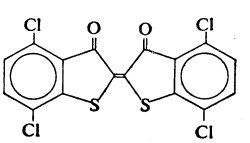
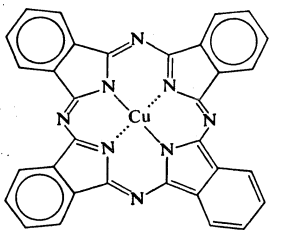
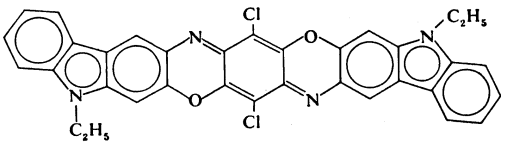
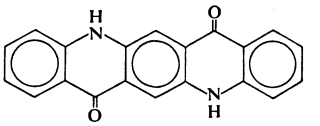
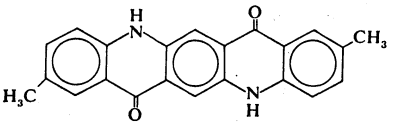
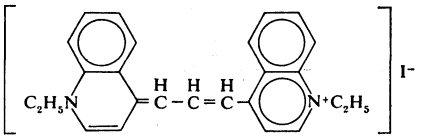
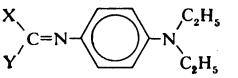
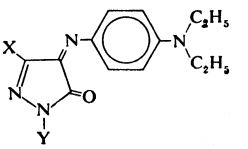
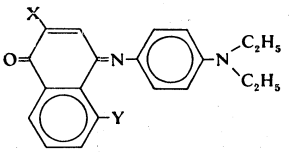
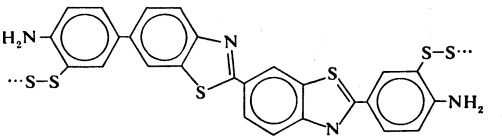
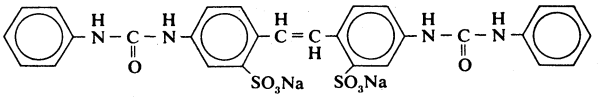
classification and name	*Colour Index* references	chemical structure	uses and properties
Benzidine Yellow	C.I. Pigment Yellow 12, C.I. 21090		good fastness
Anthraquinone Kiton Fast Blue 3G, Acilan Astrol B, etc.	C.I. Acid Blue 27, C.I. 61530		for wool and silk
Cationic dye			for acrylic fibres
Cibacet Blue BR, Celliton Fast Blue B, etc.	C.I. Disperse Blue 14, C.I. 61500		for cellulose acetate
Flavanthrone	C.I. Vat Yellow 1, C.I. 70600		for cellulose fibres
Pyranthrone	C.I. Vat Orange 9, C.I. 59700		for cellulose fibres
Indanthrone	C.I. Vat Blue 4, C.I. 69800		for cellulose fibres; also as a pigment
Indigoid and thioindigoid Tinosol Blue O, Indigosol O, etc.	C.I. Solubilized Vat Blue 1, C.I. 73002		for cellulose fibres and wool
Ciba Scarlet 3B, Indanthren Scarlet B, etc.	C.I. Vat Red 6, C.I. 73355		for cellulose fibres
Thioindigo Bordeaux			as a pigment; good fastness

Table 5: Typical Dyestuffs and Pigments, Natural and Synthetic (continued)

classification and name	*Colour Index* references	chemical structure	uses and properties
Phthalocyanine Bermuda Blue, Pigment Fast Blue	C.I. Pigment Blue 15, C.I. 74160		as a pigment; good fastness
Dioxazine Carbazole Dioxazine Violet			as a pigment; good light fastness
Quinacridone Quinacridone Red			as a pigment; good fastness
Quinacridone Magenta			as a pigment; good fastness
Cyanine Kryptocyanine			sensitizer for photographic emulsions
Indoaniline and azomethine	...		formed in the emulsion during colour development of coloured prints or transparencies (yellow layer)
	...		formed in the emulsion during colour development of coloured prints or transparencies (magenta layer)
	...	 the X and Y groups have various undisclosed constitutions	formed in the emulsion during colour development of coloured prints or transparencies (cyan-[blue-green] layer)
Sulfur dyestuffs Yellow sulfur dye	...	 dyestuffs consisting mainly of this polymeric substance	for cellulose fibres
Fluorescent brightening agents Blancol C Leucophor R, etc.	C.I. Fluorescent Brightening Agent 30, C.I. 40600		for brightening white cellulose fibres (including paper), wool, and nylon

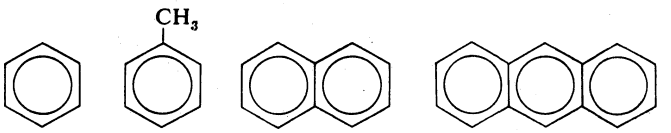
tailed account of the technology of dyeing, see the article DYES AND DYEING).

Classification. Dyestuffs and pigments are classified by chemical constitution, by method of application, by hue, or even by the name of the manufacturer. Generally, the commercial name of a particular colorant consists of: (1) a brand name denoting application class (each one specific to each manufacturer); (2) the name of the hue; (3) suffix letters (often German in origin)—*e.g.*, B, G, or R (blau, gelb, rot); (4) a strength indication—*e.g.*, 150 percent, 250, denoting brands with less diluent (usually salt) than the normal.

Colour Index references

In addition, there is a *Colour Index* which lists (and describes) all commercial dyestuffs, by application class, and by chemical constitution. For example, the dye Orange II has a *Colour Index* Part I reference as "C.I. Acid Orange 7" and a Part II reference as "C.I. 15510." In the first it is described as a dye for wool, silk, or nylon, used for garment dyeing and for colouring coir and sisal fibres. Also included are details of the grading for fastness to thirteen different agents such as washing, milling, chlorination, light, and so on. Further, the dye is shown as being marketed by about 60 different manufacturers under about 45 commercial names. The Part II reference gives the chemical constitution, method of manufacture, and colour reactions of the dye with common reagents, together with a number of literature references and the information that the dyestuff was discovered by one Z. Roussin in 1876. The manufacturing process for this dyestuff is outlined below.

Manufacture. The dyestuff industry uses hundreds of different chemical compounds to produce the thousands of different colorants it markets. These are manufactured by complex and often lengthy series of chemical reactions and utilizing a variety of techniques. The starting materials are relatively few in number, mainly the aromatic hydrocarbons, of which benzene, toluene, naphthalene, and anthracene are the most important.



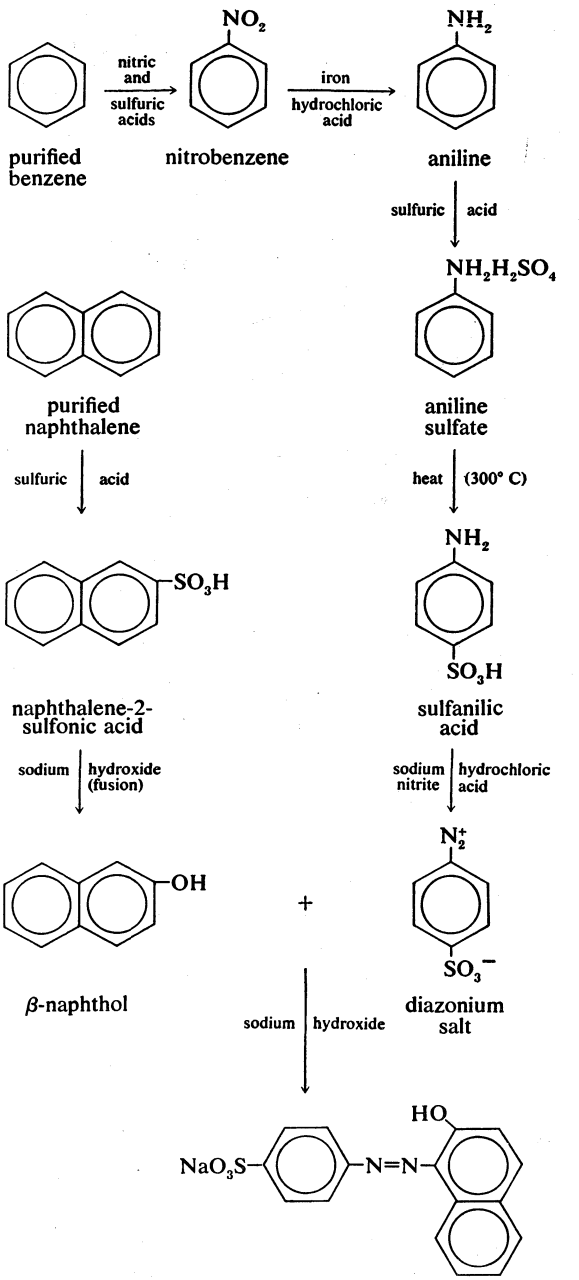
benzene toluene naphthalene anthracene

Manufacture of intermediates

From these compounds many hundreds of other substances, known as intermediates, are manufactured. The intermediates usually are not dyestuffs but from them, ultimately, the dyes are derived. In general, intermediates are formed by one or more of a variety of chemical operations, such as nitration, sulfonation, chlorination, acylation, and so on. Among the most important of the products thus formed are aniline, alkylaniline, acylaniline, chloroaniline, nitroaniline, alkoxyaniline; a large variety of sulfonated aminonaphthalenes and hydroxynaphthalenes; and aminoanthraquinone and hydroxyanthraquinone.

To illustrate, the various reaction sequences used to prepare dyestuffs are outlined below. The first is a relatively simple procedure for making an azo dyestuff, Orange II; the other is a complex series of processes, given here only in a shortened form, for making an anthraquinone vat dye, Caledon Jade Green.

Orange II. Numerically, the azo class of dyes is by far the largest group, containing examples of all hues and including products suitable for colouring all types of material. The preparation of azo dyes is based on the diazo-coupling reaction already described, in which an aromatic amine is treated in an aqueous solution with nitrous acid, thus forming a diazonium salt. The diazonium salt is unstable, and without separation from solution it is allowed to couple with a phenolic substance slowly added in an alkaline solution. The result of the coupling reaction is a highly coloured azo compound. The formation of the acid wool dye, Orange II, from the intermediates sulfanilic acid and β-naphthol is a good example. The reactions involved are shown below:

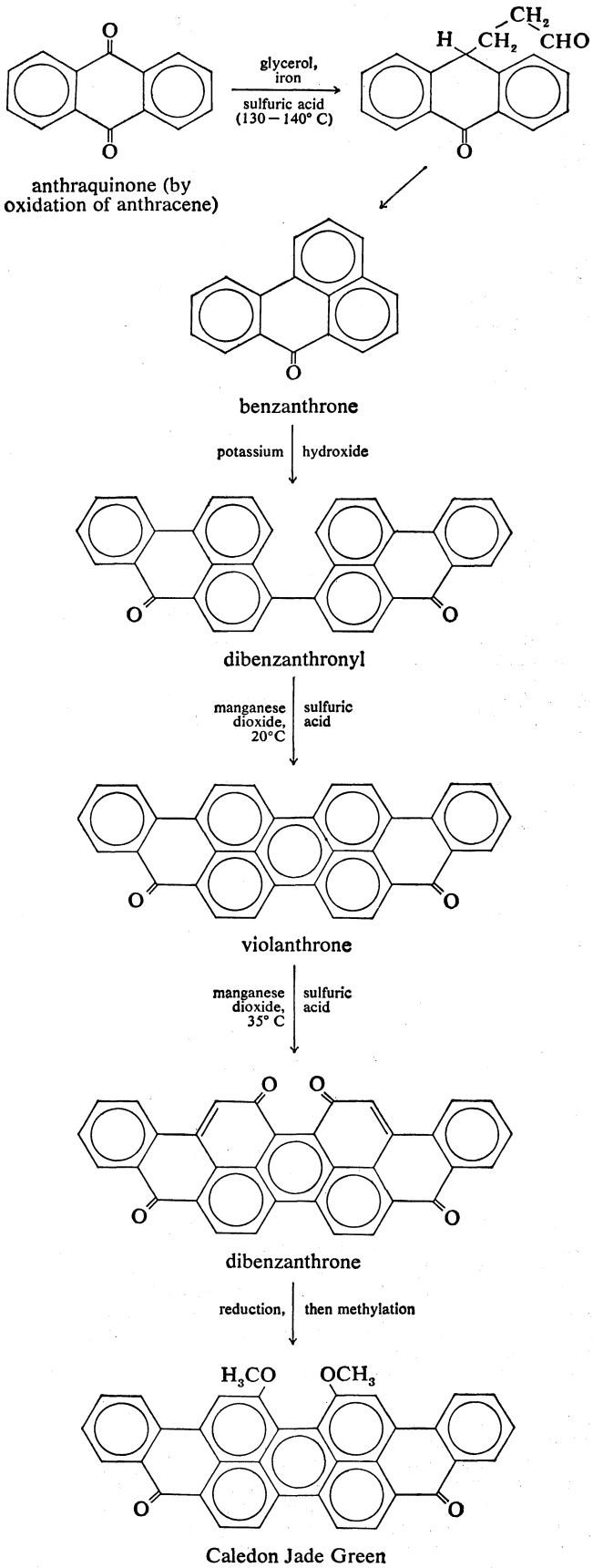


Orange II

Caledon Jade Green. Caledon Jade Green was discovered in 1920 in the laboratories of a Scottish dye maker; it has been one of the outstanding dyestuff inventions of all time, remarkable both for its attractive bright shade and for its excellent all-round fastness properties. The chemical reactions involved in its synthesis are shown below in somewhat abbreviated form.

Uses. Although the principal use of dyestuffs is in the dyeing of fibres and fabrics, dyestuffs and pigments are used in large quantities for many other purposes, including the printing of textiles and the colouring of many materials other than textiles, such as leather, paper, and plastics. Pigments are widely employed in paints, varnishes, printing inks, artists' colours, and writing and copying inks. Several colorants are used in foodstuffs, oils, and cosmetics. Dyestuffs also are important as histological stains and for measuring the surface area and porosity of industrial powders.

Special types of dyestuff (chiefly cyanine dyes) are used in making photographic emulsions and others in colour photography. Furs and human hair are dyed by processes involving the application of dyestuffs or related chemical



compounds. For foodstuff colouring, specially purified dyestuffs, free from poisonous metals, are used. Only a few dyestuffs are permitted in foods in most countries; these have been rigorously screened, particularly for freedom from carcinogenic properties.

Plastics may be mass-coloured; *i.e.*, the plastic may be ground with a pigment before the final shaping operation, a process that gives an opaque product, or they may be surface-dyed after shaping, by immersion in a dye solution. The latter method does not produce opacity and is suitable for colouring rolls of transparent plastic film or small, transparent articles, such as toothbrush handles.

Standardization, testing, analysis, and identification. After manufacture, water-soluble dyestuffs can be separated from the solution in which they have been prepared by salting out—that is, made insoluble by the addition of salts—after which the dyestuffs are filtered off and dried. Insoluble dyestuffs, which must be in a form in which they can easily be suspended in water, are finely ground in water with a dispersing agent before drying.

Commercial dyestuff products must meet standards of shade, strength, wettability, and so on, so that successive deliveries of any particular material are the same. Even with the best control of manufacturing processes, however, the shade varies slightly from one manufactured batch to another, and the products in their final processing may have to be adjusted to the shade and strength of a standard by addition of a little colouring matter of a different hue (but with similar properties) or of a diluent such as salt.

Methods of standardization

For normal dyeing operations on a technical or a laboratory scale, commercial dyestuffs are used without pretreatment, but for laboratory research purposes it is often necessary to use specially purified dyestuffs, and to check their purity by chemical analysis. Most water-soluble dyestuffs are not as readily purified as are most organic compounds because dyestuffs often are strongly ionic and therefore insoluble in organic solvents. They cannot, therefore, be recrystallized from such solvents. Recrystallization from water often is not satisfactory either because the dyestuff is not sufficiently soluble, or because it does not separate readily on cooling (or if it does, it may do so in a form difficult to remove by filtration). Moreover, most commercial dyes contain large quantities of diluent material, usually common salt, which is difficult to remove completely. For chemical research, therefore, the so-called batch dyestuffs, *i.e.*, dyes obtained directly from the manufacturing plant and unmixed with diluent, are preferable. To purify a water-soluble dyestuff for which a process has not been previously worked out, recrystallization usually is attempted, first from water alone and then, if this is not suitable, from a simple water-miscible solvent, such as acetone, ethanol, propanol, or butanol, or from mixtures of these with water. If all such methods fail, repeated salting out of the dyestuff from water, by additions of sodium acetate, is used, followed by filtration and washing of the precipitate with ethanol, which removes any residual sodium acetate. This procedure gives a product of high purity but in a rather low yield, because much of the dyestuff is lost in the filtrates and in the washings.

Methods of analysis

Dyestuffs and pigments can be analyzed by the normal methods of analysis used in organic chemistry; *i.e.*, by determination of the total content of the various elements present—carbon, hydrogen, oxygen, sulfur, and so on. With improvement in speed and accuracy, these methods, which now require very small amounts of the test sample, have tended to displace older procedures. They do not, however, detect traces of the coloured impurities, which are often present even after careful purification and which may or may not interfere with the use of the product. Detection of these is best made by the separation method known as chromatography, which is based upon adsorption of the material to be separated on a solid support.

Often it is necessary to identify dyes on samples of textile or other coloured material. This usually involves a systematic analysis in which the material is treated in succession with a variety of chemical reagents and solvents, the changes in colour produced or the amount of colour dissolved by the solvent being observed. Such schemes of analysis often are designed merely to identify the dyeing class to which the colouring matter belongs,

and this may be all the information needed. Other schemes, however, enable a precise identification of many individual dyes.

BIBLIOGRAPHY

Classification, trade names, uses, properties, and chemical constitutions of dyestuffs and pigments: SOCIETY OF DYERS AND COLOURISTS, *The Colour Index*, 3rd ed. (1971).

History of dyeing: F.W. GIBBS, "Invention in Chemical Industries," in CHARLES SINGER *et al.* (eds.), *A History of Technology*, vol. 3, ch. 25 (1957); E.J. HOLMYARD, "Dyestuffs in the Nineteenth Century," *ibid.*, vol. 5, ch. 12 (1958).

Descriptions of Peter Griess's life and work: F.A. MASON, "Johann Peter Griess, 1829–1888," *J. Soc. Dyers Colour.*, 46:33–39 (1930); J. BOULTON, "Peter Griess-Biography," *ibid.*, 75:277–278 (1959).

Chemistry of dyeing: THOMAS VICKERSTAFF, *The Physical Chemistry of Dyeing*, 2nd ed. (1954); C.H. GILES, "An Outline of the Chemistry of Dyeing Processes," *Chem. Ind.*, no. 3–4, pp. 92–101, 137–150 (1966).

Chemistry and manufacture of synthetic dyestuffs and pigments: J.C. CAIN and J.F. THORPE, *The Synthetic Dyestuffs*, 7th ed. by J.F. THORPE and R.P. LINSTEAD (1933); H.E. FIERZ-DAVID and L. BLANGEY, *Fundamental Processes of Dye Chemistry* (1949), includes laboratory experiments; H.A. LUBS (ed.), *The Chemistry of Synthetic Dyes and Pigments* (1955); K. VENKATARAMAN, *The Chemistry of Synthetic Dyes*, 4 vol. (1963–71); E.N. ABRAHART, *Dyes and Their Intermediates* (1968); R.L.M. ALLEN, *Colour Chemistry* (1971); DAVID PATTERSON (ed.), *Pigments: An Introduction to Their Physical Chemistry* (1967).

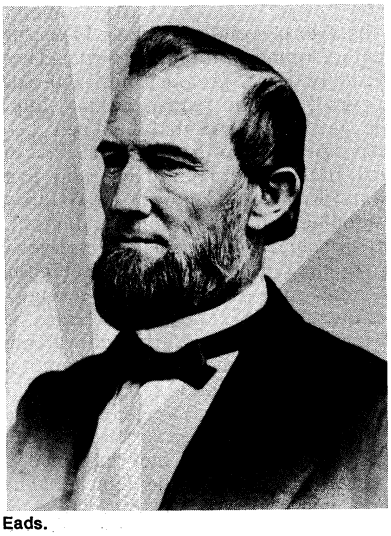
Analysis of colorants: ELLIS CLAYTON, *Identification of Dyes on Textile Fibres and Detection of Metals in Fibrous Materials, Dyes and Organic Pigments*, rev. ed. (1963); P.A. KOCH, *Rezeptbuch für Faserstoff-Laboratorien* (1960; Eng. trans., *Microscopic and Chemical Testing of Textiles: A Practical Manual*, 1963); WALTER GARNER, *Textile Laboratory Manual*, 3rd ed., 6 vol. (1966–67).

(C.H.Gi.)

Eads, James Buchanan

James Buchanan Eads, one of the outstanding civil engineers of the 19th century, is renowned chiefly for his construction of the steel triple-arch Eads Bridge at St. Louis, Missouri, and for successfully deepening and fixing the ship channel at the mouth of the Mississippi.

By courtesy of the Library of Congress, Washington, D.C.



Eads.

Born in Lawrenceburg, Indiana, on May 23, 1820, Eads was named for his mother's cousin James Buchanan, a young Pennsylvania congressman who later became president of the United States. The boy spent a migrant youth with a minimum of formal education, for his father's never very successful business ventures took the family first to Cincinnati, then Louisville, and finally St. Louis. James Eads educated himself by reading through the library of his first employer, a St. Louis dry-goods merchant. At 18 he became purser on a Mississippi riverboat. Not long after, he began to consider means to recover by salvage the heavy losses from the frequent riverboat disasters. When he was 22, he invented a salvage boat, which he called a submarine; actually it was a surface vessel from which he could descend in a diving bell that he had also designed and walk the river bottom. He recovered lead and iron pigs and other valuable freight; on one occasion he retrieved a cargo that included a large crock of butter in a good state of preservation. So successful was his equipment that in 12 years of operations on the Mississippi and its tributaries he made his fortune.

Retiring from the river to marry and settle down, Eads set himself up briefly as a glass manufacturer, but the promising enterprise, the first glass factory in the West, was ruined by the Mexican War; by 1848 he was back in the salvage business. He built three new submarines, the third of which was capable of pumping out and raising a sunken hull from the bottom. Within a few years he had ten boats in his fleet.

As the Civil War threatened, Eads foresaw the struggle that would take place for control of the Mississippi system, and he advanced a radical idea. He proposed that ironclad steam-powered warships of shallow draft be built to operate on the rivers. The U.S. government was slow to take up his offer to build such a flotilla; when it did, he built the ships in record time, working 4,000 men on day and night shifts seven days a week. The novel craft he set afloat spearheaded Grant's offensive against Forts Henry and Donelson, the first important Union victories of the war. They continued to play a conspicuous role under Andrew Foote and David Farragut at Memphis, Island No. 10, Vicksburg, and Mobile Bay. The vessels were the first ironclads to fight in North America and the first in the world to engage enemy warships. (The "Monitor" and "Merrimac," both ironclads that battled in the U.S. Civil War, were the first such vessels to close against each other in combat.) Immediately after the war, Eads was chosen to direct a construction project of extraordinary difficulty, the bridging of the Mississippi at St. Louis.

Contract for the Eads Bridge

From his knowledge of the river and of the fabrication of iron and steel, he secured, against opposition, some of it unscrupulous, a contract for a steel triple-arch bridge over the river at St. Louis, which he began on August 20, 1867. Its three spans, 502, 520, and 502 feet (152, 158, and 152 metres), respectively, consisted of triangularly braced 18-inch (46-centimetre) hollow steel tubes linked in units and set in piers based on bedrock. Since the rock lay some 100 feet (30 metres) below the river surface, reaching it posed major problems. The work of digging through the mud bottom had to be carried on under compressed air, and some of the men developed caisson disease (the bends). Following the deaths of two workers on March 19, 1870, Eads established a floating hospital, provided nourishing food for his workers, insisted on slow decompression on emerging from the caissons, and installed a lift.

The steel used in construction of the bridge was subject to similar rigorous standards; it was inspected at the works and on the site. Indeed, its supplier, the famed industrialist Andrew Carnegie, was forced to take back some batches for rerolling three times, and some were still rejected as not conforming to the specified strength of 60,000 pounds (27,000 kilograms) per square inch. Many other problems arose. To construct his first steel arches without disturbing navigation on the river, Eads used timber cantilevers to support them, with the halves of each arch held back by cables passing over the top of towers built on the piers. To join the two halves of the middle arch, Eads's deputy, Col. Henry Flad, had planned to hump the middle arch slightly to bring the two halves together; then, with the cantilevering removed, the arch would assume its normal shape. Eads, on the other hand, had prepared a wrought-iron plug fitted with threads; the last two arch ribs could be shortened by five inches each and cut with screw threads to receive the plug, which would close the distance between the ribs. Because of an unusual mid-September warm spell, which warped the bridge arches toward the north, Flad could

not close the arches by the method he had chosen and, after trying to cool the steel tubes with ice packs, fell back on Eads's screw–plug connection. The first arch was closed on September 17, 1873.

The Eads, or St. Louis, Bridge, the largest bridge of any type built up to that time, was recognized throughout the world as a landmark engineering achievement, with its pioneering use of structural steel, its foundations planted at record depths, and its cantilevering technique used for raising the arches. The bridge was officially opened on July 4, 1874.

New Orleans navigation channel

Soon after, Eads's rare understanding of the Mississippi was enlisted at New Orleans to provide a year-round navigation channel for the city. Despite widespread skepticism, he successfully altered the sedimental behaviour of the river by building a series of jetties, and within five years, by 1879, he had created a practical channel for shipping. In this important work he employed a technique of carrying out the project at his own expense, simply on the basis of guarantees if successful. On the same conditions he sought to promote a ship-carrying railway across the Isthmus of Tehuantepec, in Mexico, as a more economic and viable alternative to a canal across the Isthmus of Panama. Two bills to promote the railway failed in Congress, doubtless a great aggravation to one of such drive and dedication, and Eads's doctors ordered him to rest. He went to Nassau in the Bahamas, where he died on March 8, 1887.

James B. Eads was the first U.S. engineer to be honoured with the Albert Medal of the Royal Society of Arts in London. He had been a consultant for Liverpool docks, as well as for installations in Toronto and in Vera Cruz and Tampico, Mexico. Twice married, he had two daughters and three stepdaughters.

BIBLIOGRAPHY. ESTILL MCHENRY (ed.), *Addresses and Papers of James B. Eads* (1884); LOUIS HOW, *James B. Eads* (1900), biography by Eads's grandson; H.J. HOPKINS, *A Span of Bridges* (1970), offers a good account of the building of the bridge; FLORENCE DORSEY, *Road to the Sea and the Mississippi River: The Story of James B. Eads* (1947), the standard work, a well-documented biography, based on research in technical and family papers.

(W.H.G.A.)

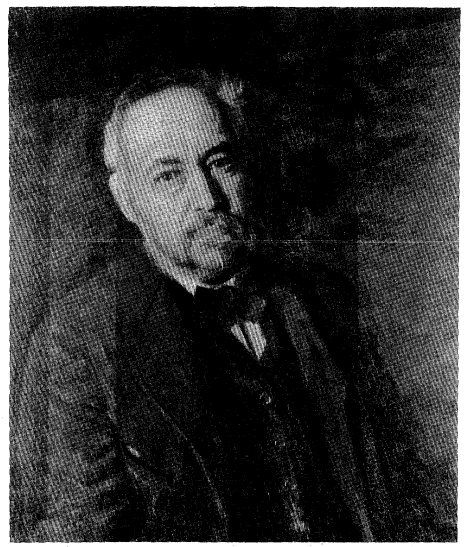
Eakins, Thomas

One of the greatest artists of the United States, Thomas Eakins worked in the mainstream tradition of Realism in American art, carrying it in the last three decades of the 19th century to what may have been its highest achievement. With his mastery of mechanical drawing, perspective, and anatomy, Eakins realistically depicted the physical world and the beings who inhabit it. But more importantly, he sought and captured the inner reality of scenes and individuals, expressing in memorable images his understanding of the world in which he lived.

Early life and first works. Thomas Eakins was born on July 25, 1844, in Philadelphia. Except for one extended study trip abroad and a brief trip to the West, virtually his entire life was spent in that city. From his father, a writing master, Eakins inherited not only the manual dexterity and sense of precision that characterizes his art but also the love of outdoor activity and the commitment to absolute integrity that marked his personal life. Young Eakins did well in school, especially in science and mathematics. As his interest in art developed, he studied at the Pennsylvania Academy of Fine Arts. Concerned particularly with the human figure, he reinforced his study of the live model at the academy by attending lectures in anatomy at Jefferson Medical College and eventually witnessing and participating in dissections.

Study in France and travel in Spain

Eakins went to France in 1866. He enrolled at the École des Beaux-Arts and studied with the leading academic painter Jean-Léon Gérôme for over three years. Unaffected by the avant-garde painting of the Impressionists, Eakins absorbed a solid academic tradition with its emphasis on drawing. After completing his study, Eakins went to Spain late in 1869, where he was greatly influenced by the 17th-century paintings of Diego Velázquez and Jusepe de Ribera. Perhaps reacting against the rigours of his academic training, he preferred artists who used paint and brush boldly to express their sense of life, creating what he called "big work." In Spain, his student days behind him, Eakins undertook his first independent efforts at oil painting.



Thomas Eakins, self-portrait, oil on canvas, 1902; in the National Academy of Design, New York.
By courtesy of the National Academy of Design, New York

Eakins returned to Philadelphia in the summer of 1870. His earliest artistic subjects were his sisters and other members of his family and the family of his fiancée, Katherine Crowell. Redolent with the character of each individual in an intimate and personal domestic setting—pensive young ladies at the piano, children engrossed with toys scattered on the floor, Katherine playing with a kitten in her lap—these rich, warm portraits seem to express in colour and mood the essence of what Lewis Mumford called "the Brown Decades." Close family ties were important to Eakins, and the intimate harmony of his home life was seriously disrupted and saddened by the death, first of his mother, and, later, of Katherine Crowell.

Eakins resumed the vigorous outdoor life of his earlier years—hunting, sailing, fishing, swimming, rowing. These activities, like his family circle, provided him with subject matter for his art. A candid realist, Eakins simply painted the people and the world that he knew best, choosing his subjects from the life that he lived. Like the poetry of his aged friend Walt Whitman, who lived across the Delaware River in Camden, New Jersey, Eakins' art was autobiographical, "a song of himself." Eakins, in fact, often included himself as an observer in his own paintings—sculling in the background behind his friend in "Max Schmitt in a Single Scull," peering intently at a surgical operation in "The Agnew Clinic," or treading water next to his setter dog Harry and watching a group of students swimming nude in "The Swimming Hole." Each of the early outdoor scenes, natural and informal at first glance, was, in fact, carefully composed on a perspective grid, with each object precisely located in pictorial space. Each image is further informed by Eakins' personal knowledge of the scene depicted. Thus colour, composition, and the play of lights and darks subtly convey to the viewer a fuller understanding of and feeling for the concentrated energy of a sculler propelling his boat through the water, or the taut equilibrium of the moment when a hunter standing in his boat balances himself, sights his target, and slowly squeezes the trigger.

Eakins' masterpiece

In 1875 Eakins, who had yet to become well known, decided to paint a major picture for the Centennial Exposition to be held in Philadelphia the following year. Eakins took as his subject a scene that had become familiar to him—Samuel Gross of Jefferson Medical College operating in his clinic before his students. Gross was a magnetic teacher and one of the country's greatest

surgeons. Eakins often selected moments that reveal multiple aspects of a scene and in this picture depicts Gross as both surgeon and teacher. Gross stands in the centre of a sombre amphitheatre, starkly top-lighted by a flood of cool daylight cascading down from a skylight above; he is dressed in black street clothes. He has opened an incision in the leg of the anesthetized male patient stretched out before him. While his assistants probe the wound, the doctor turns, one hand holding a scalpel covered with blood, to tell his students what he has done and what he will do next. At the left a seated woman, perhaps the patient's mother, flings an arm across her face, shielding her eyes from the scene, her fingers clawing the air in anguish. Her emotion and the note of pain and suffering inherent in the subject contrast strikingly with the cool professionalism of Gross, whose calm features reflect assurance and determination as well as compassion. The painting objectively records a realistic drama of contemporary life, full of feeling but free of sentimentality. "The Gross Clinic" is generally agreed to be Eakins' masterpiece.

Rejection by critics and public. To Eakins' dismay "The Gross Clinic" was rejected for the art exhibition at the Centennial Exposition, and he had to exhibit it in a medical section. Critics and public alike responded to the painting unfavourably. While they could accept historical scenes of grisly martyrdoms or bloody massacres without qualm, "The Gross Clinic" represented blood and pain and suffering as immediate facts in Philadelphia. That was offensive and unacceptable. Viewers could not appreciate a picture that was neither entertaining nor ennobling but simply a frank statement of contemporary reality. The rejection of the painting was the first of many rebuffs Eakins was to receive from Victorian contemporaries who shared his world but not his values.

Interest in anatomy

From his earliest student days, Eakins had been primarily interested in studying and portraying the human figure. His early sculling scenes displayed the musculature of athletic men, and "The Gross Clinic" dealt directly with the subject of human anatomy. But Eakins found few subjects in contemporary Philadelphia that afforded opportunities for portraying the undraped human figure, especially females. He circumvented this by painting repeatedly a partly imaginary scene of William Rush, a much earlier Philadelphia sculptor, carving his statue of the "Nymph with Bittern" from a naked female model in the presence of a chaperon, which provided him with a pictorial pretext for portraying a nude woman.

In the late 1870s Eakins began to teach at the Pennsylvania Academy of Fine Arts, where he became professor of drawing and painting in 1879. A popular and influential teacher, Eakins stressed anatomy and drawing from live, nude models as opposed to the study of plaster casts of antique sculpture. The fame of the Pennsylvania Academy as a centre for the best art instruction in the country spread among young artists. Yet notoriety accompanied repute, and objections were voiced increasingly from outside the academy to Eakins' unrestrained use of nude models in front of mixed classes. The suspicious were unable to accept Eakins' assurance that the relationship between artist and model was as innocent, objective, and professional as that between doctor and patient. Eakins continued to insist on the importance of teaching from nude human models and was finally forced to resign in 1886. Teaching had become a major part of his life, and this was another severe blow. He continued to teach sporadically at the newly formed Art Students League in Philadelphia and at the National Academy of Design in New York, and his personal relationships with young artists remained close. One bright moment during these difficult years occurred in 1884, when he married one of his pupils, Susan Macdowell.

Concern with locomotion

As a corollary to his interest in anatomy, Eakins was fascinated with locomotion—human and animal figures in motion. A commission in 1879 to paint Fairman Rogers driving his four-in-hand coach through Fairmount Park in Philadelphia (Philadelphia Museum of Art) led him to an intensive study of horse anatomy, and he made a number of sculpted wax sketches of horses in motion. He developed a serious interest in sculpture, an aspect of his art that only became appreciated much later. His interest in locomotion led to familiarity with the experiments in sequential photography being made in California by Eadweard Muybridge. By 1884 Eakins himself was experimenting with multiple-image photography of moving athletes and animals. And in later years his interest in the human figure in motion led him to make a series of impressive paintings of boxing scenes.

Eakins' interests ranged widely—sports, anatomy, locomotion, music, sculpture, photography—in directions often reminiscent of his great French contemporary Edgar Degas but without that artist's innovative stylistic concerns. There is no evidence, however, that Eakins was aware of the work of Degas. Eakins' art does demand comparison with that of Winslow Homer, the contemporary he most admired and his principal rival claimant to the title of the greatest American artist of the 19th century. Homer, also an objective realist, was similarly interested in outdoor sports and such sporting subjects as hunting, canoeing, and fishing. He also had a similar love for and identification with a specific place—in Homer's case, Prouts Neck, Maine. Homer's art is cool, detached, impersonal, and ultimately pessimistic in its view that man is at the mercy of a deterministic universe. Eakins' art, although often sad in its reflection of the buffeting each human receives in the course of his years, still is ultimately optimistic in its humanism, in its message that man, through his individual actions—a doctor with a knife, a sculler with an oar, a hunter with a gun, a boxer with his gloved fist, a musician with his instrument, a singer with her voice, a chess player with his pieces, a scientist with his instruments—can act, do things, have an effect in this world. Despite the wide variety of his subject matter, almost all of Eakins' art is portraiture, images of real people whom he knew and loved or respected. In his representations of the physical world, Eakins combined a technical ability to depict the external aspect of things with a probing for the essence of each scene. In his portraits of individuals, he similarly combined the faithful representation of the external and anatomical realities of each person with a deeper probing into the subject's inner being and character. The people he portrays have lived, and often their experiences are etched on their faces. The wear and tear of years is not glossed over but celebrated in staring eyes, wrinkles, and slumping torsos.

Eakins' art as portraiture

Significance and influence. Although always respected for his ability, Eakins remained throughout his years something of an outcast. His contemporaries, rather than allow themselves to be shaken by his frank statements of the human condition and his joyous appreciation of the human body, ignored him. He sold few pictures, but fortunately a small private income matched his modest needs. Unfettered by the demands of clients, Eakins was free to paint what and, more importantly, whom he wished. His art was never compromised by the need to flatter patrons or sitters, and honesty was his only policy. Good friends and faithful followers rather than fame and fortune were his lot. Not until the year of his death (June 25, 1916) was one of his paintings acquired by a museum ("Pushing for Rail," Metropolitan Museum of Art), and the first major exhibition of his work was held the following year (Metropolitan Museum of Art). But Eakins' art had its long-range effect, serving as a model and an impetus for the burst of realism in American painting during the early years of the 20th century, especially in the work of George Bellows and the group called the Ashcan School of painters. And despite the increasing dominance of abstract art during the middle years of the 20th century, a pervasive and stubborn substream of realism surfaced periodically—Regionalism, Pop art, the figurative work of artists such as George Segal and Leonard Baskin—to manifest the continuing debt of American art to the achievement of Thomas Eakins.

MAJOR WORKS

All paintings in oils unless otherwise noted. "A Street Scene in Seville" (1870; Mrs. John R. Garrett, Sr. Collection);

"Home Scene" (1871; Brooklyn Museum, New York); "Max Schmitt in a Single Scull" (1871; Metropolitan Museum of Art, New York); "Katherine" (1872; Yale University Art Gallery, New Haven, Conn.); "The Pair-Oared Shell" (1872; Philadelphia Museum of Art, Philadelphia); "The Biglin Brothers Turning the Stake" (1873; Cleveland Museum of Art, Cleveland, Ohio); "John Biglin in a Single Scull" (water colour, 1873; Metropolitan Museum of Art, New York); "Sailing" (1874; Philadelphia Museum of Art); "Pushing for Rail" (1874; Metropolitan Museum of Art, New York); "The Gross Clinic" (1875; Jefferson Medical College, Philadelphia); "Will Schuster and Blackman Going Shooting" (1876; Yale University Art Gallery, New Haven, Conn.); "Chess Players" (1876; Metropolitan Museum of Art, New York); "William Rush Carving His Allegorical Figure of the Schuylkill River" (1877; Philadelphia Museum of Art); "The Fairman Rogers Four-in-Hand" (1879; Philadelphia Museum of Art); "The Crucifixion" (1880; Philadelphia Museum of Art); "The Pathetic Song" (1881; Corcoran Gallery of Art, Washington, D.C.); "The Swimming Hole" (1883; Fort Worth Art Center Museum, Fort Worth, Texas); "Lady with a Setter Dog (Mrs. Eakins)" (1885; Metropolitan Museum of Art, New York); "Walt Whitman" (1887; Pennsylvania Academy of the Fine Arts, Philadelphia); "Letitia Wilson Jordan Bacon" (1888; Brooklyn Museum, New York); "Miss Van Buren" (*c.* 1889; Phillips Collection, Washington, D.C.); "The Agnew Clinic" (1889; University of Pennsylvania, Philadelphia); "Professor Henry A. Rowland" (1891; Addison Gallery of American Art, Andover, Mass.); "The Concert Singer" (1892; Philadelphia Museum of Art); "Frank Hamilton Cushing" (1894–95; Thomas Gilcrease Institute of American History and Art, Tulsa, Oklahoma); "Taking the Count" (1898; Yale University Art Gallery, New Haven, Conn.); "Salutat" (1898; Addison Gallery of American Art, Andover, Mass.); "Between Rounds" (1899; Philadelphia Museum of Art); "Benjamin Eakins" (1899; Philadelphia Museum of Art); "Mrs. William D. Frishmuth" (1900; Philadelphia Museum of Art); "The Thinker: Louis N. Kenton" (1900; Metropolitan Museum of Art, New York); "Self-Portrait" (1902; National Academy of Design, New York); "Mrs. Edith Mahon" (1904; Smith College Museum of Art, Northampton, Mass.); "Monsignor Diomede Falconio" (1905; National Gallery of Art, Washington, D.C.).

BIBLIOGRAPHY. LLOYD GOODRICH, *Thomas Eakins: His Life and Work* (1933), is the principal monograph on Eakins. Other books of interest on the subject are FAIRFIELD PORTER, *Thomas Eakins* (1959); and SYLVAN SCHENDLER, *Eakins* (1967). The largest collection of works by Thomas Eakins is to be found in the Philadelphia Museum of Art. There are also substantial collections at the Metropolian Museum of Art in New York and at the Yale University Art Gallery.

(J.D.Pro.)

Ear and Hearing, Human

The ear is an organ capable of receiving and responding to sound—that is, to the minute vibratory disturbances of the air molecules. Not only does the ear detect even faint sounds and noises, so that it can serve as a sentinel, but it also analyzes them. After the mechanical energy of vibration has been converted in that portion of the inner ear called the cochlea (see below) to an electrochemical form (the process of conversion is known as transduction), the ear sends to the brain coded information about the frequency, intensity, and complexity of the sounds. The sensations evoked are those of pitch, loudness, and timbre, all of which contribute to the recognition of the particular kind of sound heard. Equally important in the processing of sound stimuli, especially those of speech, is the ear's accurate recording of their temporal sequence, undistorted by fatigue, adaptation, or afterimages.

The ears are remarkably sensitive to minute differences in the intensity of a sound and the time of its arrival at each ear. This fine discrimination makes it possible not only to localize a source of sound with considerable accuracy but also to attend to certain sounds while ignoring or suppressing others that are of less interest or significance.

The inner ear also contains, in addition to the cochlear transducer, a second set of sense organs that make up the vestibular system. This system is concerned with equilibrium—*i.e.*, with balancing the body by adjustments of the position of the head and the posture. From a functional point of view, equilibrium is more closely related to vision than to hearing, despite its intimate anatomical relationship to the cochlea and the similarity of the sensory cells in the two sets of organs.

The present article contains a description of the structure of the ear and the physical relations of its parts, including both the cochlea and the vestibular organs; an account of the transmission of sound waves to the cochlea and to its sensory elements contained in the organ of Corti; and a description of the auditory nerve and the auditory pathways of the central nervous system. Tests of hearing are briefly surveyed. The article concludes with sections on the role of the vestibular system and on the tests used to detect disturbances of vestibular function.

Structure of the ear

Parts of the ear

The human ear, like that of other mammals, has three distinguishable parts: the external, the middle, and the inner ear. The external ear consists of the portion projecting from the side of the head, called the auricle or pinna, and the external auditory canal, which ends blindly at the eardrum. The middle ear is a narrow, air-filled space within the temporal bone, separated from the outside by the tympanic (eardrum) membrane and crossed by a chain of three tiny bones, the auditory ossicles. The inner ear is a complicated system of fluid-filled passages and cavities, deep in the rock-hard petrous portion of the temporal bone. It contains the sensory organs of hearing and equilibrium, the specialized endings of the auditory, or eighth cranial, nerve (Figure 1).

By courtesy of Abbott Laboratories, North Chicago, Illinois

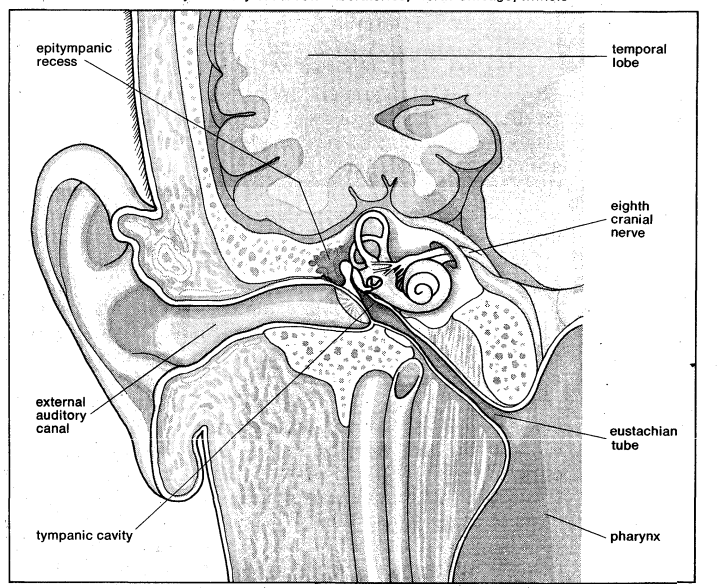


Figure 1: Section through the right side of the head showing the parts of the ear and the auditory area in the temporal lobe of the cerebral cortex.

EXTERNAL EAR

The most striking differences between the ear of man and that of many other mammals are seen in the auricle itself. In man the auricle is an almost rudimentary, usually immobile shell, more or less closely applied to the side of the head. It consists of a thin plate of yellow fibrocartilage covered by closely adherent skin. The cartilage is molded into the characteristic shape with clearly defined hollows, ridges, and furrows, forming an irregular, shallow funnel (Figure 2). The deepest depression, leading directly to the external auditory canal, or meatus, is called the concha. It is partly covered by two small projections, the tongue-like tragus in front and the antitragus behind. Above the tragus, a prominent ridge, the helix, arises from the floor of the concha and continues as the incurved rim of the upper portion of the auricle. An inner, concentric ridge, the antihelix, surrounds the concha and is separated from the helix by a furrow, the scapha (also called the fossa of the helix). In some ears

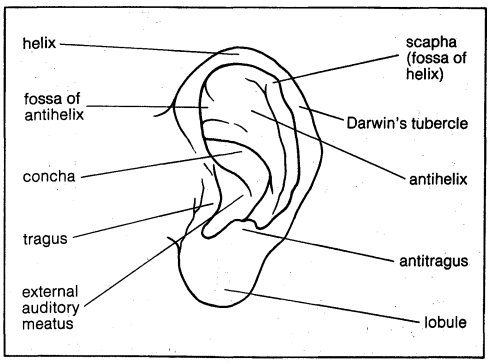


Figure 2: Structure of the outer ear.

a little prominence, Darwin's tubercle, is seen along the upper, posterior portion of the helix, the vestige of the folded-over point of the ear of a remote prehuman ancestor. The lobule, the fleshy, lower portion of the auricle, is the only part of the external ear that contains no cartilage. The auricle also has several small rudimentary muscles, which connect it to the skull and scalp. They are usually without function but in some persons are capable of limited movements.

External auditory canal

The external auditory canal is a slightly curved tube, extending inward from the floor of the concha and ending blindly at the tympanic membrane. In the outer third the wall consists of cartilage; in the inner two-thirds, of bone. The entire length of the passage (24 millimetres, or almost one inch) is lined with skin, which also covers the surface of the eardrum membrane. Fine hairs directed outward and modified sweat glands producing earwax, or cerumen, discourage insects from entering the canal.

TYMPANIC MEMBRANE AND MIDDLE EAR

Tympanic membrane. The tympanic membrane lies obliquely across the end of the external canal in the form of a flattened cone with its apex directed inward. Its edges are attached to a ring of bone, the tympanic annulus. The diameter of the membrane is about nine millimetres (about one-third inch). The appearance and mobility of the membrane are important for the diagnosis of middle-ear disease. The healthy membrane is pearl gray; sometimes it has a pinkish or yellowish tinge.

The drum membrane is made up of three layers. The outer layer is continuous with the skin of the external canal; the inner with the mucous membrane lining the middle ear. Between them is a layer of circular and radial fibres that give the drum its stiffness and tension. It is well supplied with blood vessels and with sensory nerve fibres that make it acutely sensitive to pain.

The cavity, including the eustachian tube. The cavity of the middle ear is a narrow, air-filled space. A slight constriction divides it into two chambers, the tympanum (eardrum) proper, or atrium below, and the epitympanum, or attic above. Its outer wall is formed largely by the tympanic membrane, its inner wall by the bony capsule of the inner ear. Its roof and floor are thin plates of bone separating it from the cranial cavity and the brain above and from the carotid artery and the jugular vein below. Its posterior bony wall has an opening leading to a second air space, the tympanic antrum, and to the small air cells of the mastoid process, the portion of the temporal bone directly in back of the external auditory canal. In the narrow anterior wall is the opening of the eustachian or auditory tube.

The eustachian tube and its function

The eustachian tube, about 45 millimetres (1.75 inches) in length, leads downward and inward from the tympanic cavity to the nasopharynx, the space that is behind and continuous with the nasal passages and above the soft palate. At its upper end the tube is rather narrow and surrounded by bone. Toward the pharynx it widens and becomes cartilaginous. Its mucous lining is continuous with that of the middle ear. The cilia (small motile hair-like projections) that cover it help to speed the drainage of mucous secretions from the middle ear to the pharynx.

The major function of the tube is to ventilate the middle ear and thus maintain equal air pressure on both sides of the drum membrane. The tube is closed at rest and opens during swallowing, so that minor pressure differences are adjusted without conscious effort. During a dive or a rapid descent in an airplane the tube may remain tightly closed. The discomfort that is felt inside the ear as the atmospheric pressure increases outside can usually be overcome by attempting a forced expiration while the mouth and nostrils are held tightly shut. This manoeuvre, which raises the air pressure in the pharynx and causes the tube to open, is named for the Italian anatomist Antonio Maria Valsalva (1666–1723), who recommended it for clearing pus from an infected middle ear.

Auditory ossicles. Three auditory ossicles with somewhat fanciful names form a short chain crossing the middle ear. From the outside inward, they are the malleus (hammer), the incus (anvil), and the stapes (stirrup). In appearance the malleus more nearly resembles a club, and the incus a premolar tooth with widely spreading roots. The stapes, on the other hand, is unmistakably a stirrup.

The head of the malleus and the body of the incus lie in the attic above the upper margin of the drum. The two small bones have a tightly fitting joint between them. The handle of the malleus is firmly attached to the upper half of the drum membrane. Its head is anchored to the walls and roof of the attic by three small ligaments. Another ligament fixes the short process (projection) of the incus in a slight depression in the rear wall of the cavity. Its long process is bent near its lower end and carries a small bony knob that forms a loose joint with the head of the stapes. This last, the smallest and lightest of the ossicles, is about three millimetres (about 0.1 inch) in height and weighs scarcely three milligrams (0.0001 ounce). It lies almost horizontally, at right angles to the long process of the incus. Its footplate fits nicely in the oval window, one of the two openings in the wall of the bony labyrinth, where it is held in place by a ring-like ligament called the annular ligament.

Two minute muscles are found in the middle ear. One, the slender muscle called the tensor tympani, emerges from a bony canal just above the opening of the eustachian tube, runs backward, changes direction as it passes over a pulley-like projection, and attaches to the upper part of the handle of the malleus. Its contractions tend to pull the malleus inward and thus increase the tension of the drum membrane. The other, called the stapedius, rises in the posterior wall and sends its minute tendon forward to attach to the neck of the stapes. Its contractions tend to pull the footplate out of the oval window by tipping the stapes backward.

When the healthy tympanic membrane is examined with the otoscope (an instrument designed for visual inspection of the interior of the ear), through the membrane the handle of the malleus is clearly seen projecting from above downward and backward and dividing the upper portion of the membrane into two almost equal parts. Behind and parallel to it, the long process of the incus can sometimes be made out. Above the malleus is a small triangular area in which the membrane is thin and slack. Behind this area, called the pars flaccida (as opposed to the pars tensa, which makes up the much larger portion of the membrane), lies the bare chorda tympani, a slender branch of the facial nerve that passes through the middle ear on its way to join the lingual nerve. It carries important secretory fibres to the parotid (salivary) gland, and sensory fibres to the taste buds of the tip of the tongue.

INNER EAR

Structure as a whole. The inner ear is enclosed in a bony case called the otic capsule, a part of which forms the inner wall of the middle ear. There, two openings between the middle and inner ear are found: the oval window above, which is filled by the footplate of the stapes, and the round window, which is covered by a thin membrane, sometimes referred to as the secondary

tympanic membrane. Between them is a bulge called the promontory.

Because of its complicated galleries and chambers, the inner ear as a whole is referred to as the labyrinth. There are, in fact, two labyrinths, one inside the other. The passages hollowed out in the otic capsule constitute the bony labyrinth. In it is suspended a delicate system of ducts and sacs that constitutes the membranous labyrinth (Figure 3).

By courtesy of Abbott Laboratories, North Chicago, Illinois

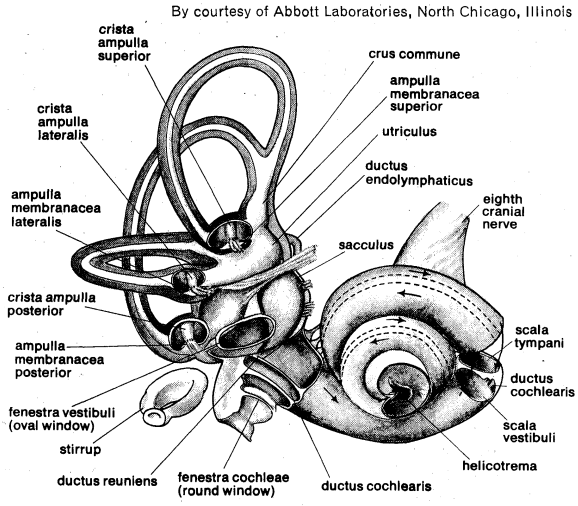


Figure 3: The membranous labyrinth, within the bony labyrinth. The stapes has been removed from the oval window.

The bony labyrinth

The bony labyrinth consists of a central chamber called the vestibule, the three semicircular canals, and the spirally coiled cochlea. The last strongly resembles the shell of a snail, and its name is derived from the Greek word for a snail. The canals are designated, according to their position, superior, lateral, and posterior. The superior and posterior canals are in diagonal vertical planes that intersect at right angles. The lateral canal is often referred to as the horizontal because it lies approximately in that plane. Each canal has an expanded end, the ampulla, which opens into the vestibule. The ampullae of the lateral and superior canals lie close together, just above the oval window, but the ampulla of the posterior canal opens on the opposite side of the vestibule. The other ends of the superior and posterior canals join to form a common stem, or crus, which also opens into the vestibule. Nearby is the mouth of a canal called the vestibular aqueduct, which opens into the cranial cavity. The lateral canal has an independent opening at its opposite end. Thus the vestibule, in effect, completes the circle for each of the semicircular canals.

The cochlea, a tube consisting of two and one-half spiral turns around a hollow central pillar, the modiolus, forms a cone approximately nine millimetres (0.4 inch) in diameter at its base and five millimetres (0.2 inch) in height. The tube is approximately 30 millimetres (1.2 inches) in total length, tapering from a width of two millimetres (0.08 inch) where the basal coil joins the vestibule and ending blindly in the cupula at the apex. The hollow centre of the modiolus is a spiral canal, containing the fibres and ganglion cells of the cochlear nerve; this enters the base of the modiolus through an opening in the petrous portion of the temporal bone called the internal meatus. A thin bony shelf, the osseous lamina, winds around the modiolus like the thread of a screw. It projects about half way across the cochlear canal, partly dividing it into two ramps, or galleries, the scala vestibuli above and the scala tympani below. The round window is in the wall of the scala tympani at its basal end. Nearby is the opening of the narrow cochlear aqueduct, through which passes the perilymphatic duct, connecting the interior of the cochlea with the posterior cranial fossa (the rear portion of the floor of the cranial cavity).

Suspended in the bony labyrinth, the membranous labyrinth occupies only a fraction of the available space. Each of the semicircular canals contains a narrow membranous semicircular duct, which by no means fills the channels or lumen of the canal, with a rounded expansion, the ampulla, at one end. In the same way that the semicircular canals communicate with the vestibule, the semicircular ducts open into an elongated tubular sac, the utricle, located in the upper part of the vestibule. The utricle is also connected, through a narrow duct, with a similar structure, the saccule, which occupies the lower part of the vestibule. The cochlear duct is a coiled, tapered tube, suspended between the scala vestibuli and scala tympani of the cochlea, ending blindly at both its basal and its apical ends. It communicates with (opens into) the saccule through a narrow connecting duct (the ductus reuniens) near the basal end. The duct that unites the utricle and the saccule is also connected to another narrow channel, the endolymphatic duct, which passes through the vestibular aqueduct and ends in a pouch on the cranial surface of the petrous bone, the endolymphatic sac.

Perilymph and endolymph

The space within the bony labyrinth that is not occupied by the various parts of the membranous labyrinth is filled with a watery fluid, the perilymph, which in composition closely resembles the cerebrospinal fluid, the aqueous humour of the eye, and other extracellular fluids of the body tissues. Like them, it is apparently formed locally from the blood plasma by ultrafiltration through the walls of the minute blood vessels called capillaries. Since it is possible for cerebrospinal fluid to enter the cochlea by way of the perilymphatic duct, a portion of the perilymph may come from that source.

The membranous labyrinth is filled with a second watery fluid of different composition, the endolymph. Chemical analysis has shown that perilymph is like most other extracellular fluids, being high in its concentration of sodium ions (about 150 milliequivalents per litre) and low in its concentration of potassium ions (about five milliequivalents per litre), but that endolymph is unique among the extracellular fluids in that it has potassium as the dominant ion (about 140 milliequivalents per litre), with a much reduced concentration of sodium (about 15 milliequivalents per litre). Contrary to the statement sometimes found in textbooks, mammalian endolymph is not a viscous fluid. Recent studies indicate that its viscosity is approximately that of water.

The process of formation of the endolymph and the maintenance of the difference in composition between it and perilymph is not yet completely understood. The tissue known as the stria vascularis, in the wall of the cochlear duct, is thought to play an important role in its secretion, but other tissues of the cochlea and the vestibular organs are probably also involved. Since the membranous labyrinth is a closed system, the question of removal of the endolymph is also of interest. Reabsorption is thought to occur from the endolymphatic sac, but this is probably only part of the story. In all probability other tissues also have important roles in regulating the inner-ear fluids.

Utricle and saccule

Parts of the inner ear. *Vestibule.* The vestibule includes the utricle and saccule, each of which contains a single sensory patch called a macula. In the utricle the macula projects from the anterior wall as an oval spot on a small rounded shelf. In the saccule, the macula lies against the medial (inner) wall of the vestibule directly overlying the bone. It is more elongated than the utricular macula and somewhat resembles the letter J in shape. Each macula is covered by neuroepithelium. This covering consists of sensory cells—called hair cells because of their hairlike projections—and supporting cells. The cells are separated by a basement membrane from the underlying connective tissue. Fibres from the vestibular branch of the vestibulocochlear (eighth cranial, or auditory) nerve enter the macula and pierce the basement membrane to end either at the base of the hair cells or as cuplike formations called calyces surrounding their cell bodies. The hair cells are thus the essential sensory receptor elements of the macula. Each hair cell

is topped by a bundle consisting of about 100 fine, nonmotile "hairs" (stereocilia) of graded lengths and a single motile hairlike projection called a kinocilium. Covering the entire macula is a delicate acellular structure, the otolithic (or statolithic) membrane, which is often described as gelatinous although it has a distinct fibrillar (minute-fibre-like) pattern. The surface of the membrane is covered by a blanket of rhombohedral crystals, the otoconia (or statoconia), consisting of calcium carbonate in the form of calcite. These crystalline particles, which range in length from one to 20 microns (there are about 25,000 microns in an inch), have a specific gravity almost three times that of the membrane itself and thus add considerable mass to it. The hair cells of the maculae are of two types. One type has a rounded body enclosed by a nerve calyx, the other a cylindrical body with nerve endings at its base. The surfaces of the utricular and saccular maculae show characteristic organization and orientation of the two types of cells in definite mosaic patterns. The larger, rounded cells tend to have a curvilinear arrangement near the centre of the macula, with the cylindrical cells around the periphery. The significance of these patterns is poorly understood, but they are presumed to favour increased directional sensitivity to movements of the head (Figure 4).

From H.H. Lindeman, "Studies on the Morphology of the Sensory Regions of the Vestibular Apparatus," *Ergebnisse der Anatomie und Entwicklungsgeschichte*, vol. 42.1 (1969); Springer-Verlag, New York

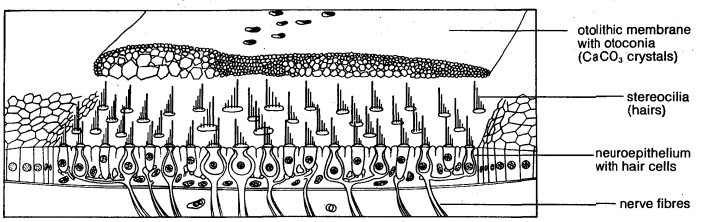


Figure 4: Structure of the macula of the utricle.

The utricle and saccule with their maculae are often referred to as otolith organs. Because they respond to gravitational forces, they are also called gravity receptors.

Sensory end organ of semicircular duct

Semicircular canals. The sensory end organ of each semicircular duct is located in the expanded end, or ampulla. It consists of a saddle-shaped ridge of tissue covered with a sensory epithelium containing the same types of cells as do the maculae. The ridge, called the crista, extends across the ampulla from side to side, at right angles to the direction of the duct (Figure 5). It

From H.H. Lindeman, "Studies on the Morphology of the Sensory Regions of the Vestibular Apparatus," *Ergebnisse der Anatomie und Entwicklungsgeschichte*, vol. 42.1 (1969); Springer-Verlag, New York

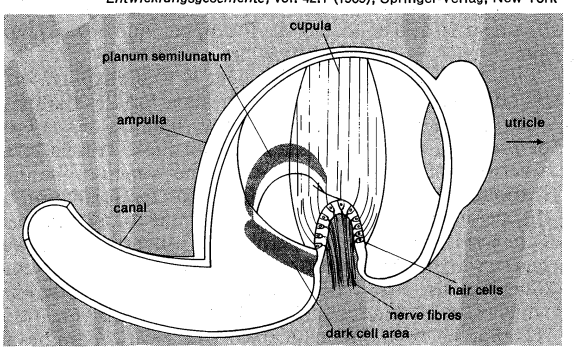


Figure 5: Ampulla of semicircular duct, showing hair cells and cupula.

is covered by a gelatinous structure, the cupula, which extends to the roof of the ampulla immediately above it, dividing the interior of the ampulla into two approximately equal parts. On each of the hair cells of the sensory epithelium of the crista there are about 100 stereocilia, or "hairs," anchored in the dense cuticular plate. The single kinocilium springs from a small area at one side of the surface of the cell, where it is covered

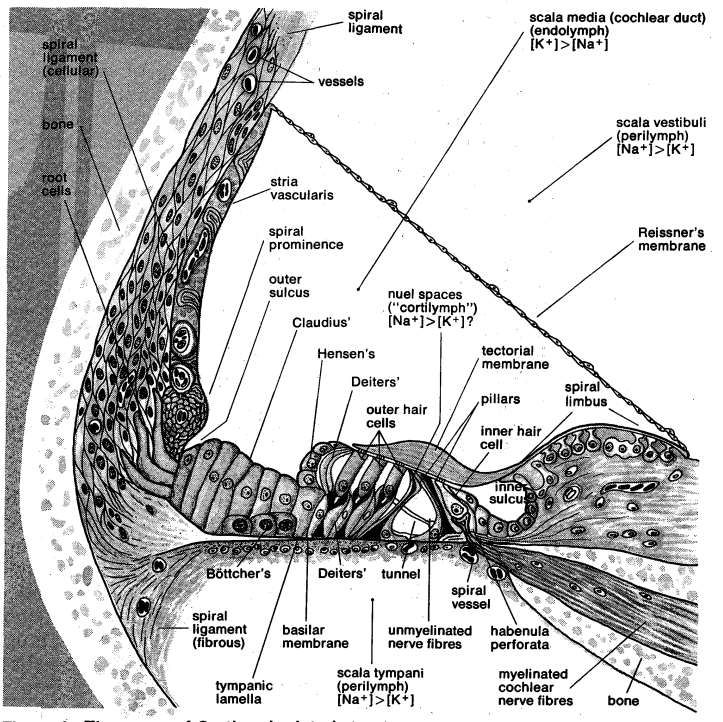


Figure 6: The organ of Corti and related structures as shown in a cross section through the cochlear duct of a guinea pig (see text).

From C.H. Best and N.B. Taylor, *The Physiological Basis of Medical Practice*, 8th ed. (© 1966); Williams and Wilkins Co., Baltimore

only by the thin plasma membrane. The stereocilia stand straight, in orderly rows. The longest are those nearest the kinocilium, which is motile. The height of the others decreases in stepwise fashion away from the kinocilium. Together with the longest stereocilia, the kinocilia extend well up into the substance of the cupula, occupying fine parallel channels. The cupula is thus attached at its base to the crista, but is free to move in either direction in response to pressure in the endolymph toward or away from the utricle. Its motion is transmitted to the hairs, causing them to bend just above the cuticular plate. This bending is the stimulus to the hair cells and elicits impulses in the vestibular fibres of the auditory nerve. Each macula and each crista is supplied by fibres from the vestibular nerve, one of the two parts of the vestibulocochlear (auditory, or eighth cranial) nerve.

Cochlea. The interior of the cochlea is divided longitudinally into three spiral ramps or scalae: the scala vestibuli, which communicates with the vestibule; the scala tympani, which ends blindly at the round window; and the scala media, or cochlear duct, which lies between the other two scalae and ends blindly both at the round window and at the apex. The scala vestibuli and scala tympani communicate at the apex, through an opening called the helicotrema (Figure 3).

The scala media

In cross section the scala media resembles a right triangle. Its base is formed by a thin bony shelf, the spiral osseous lamina, and by the basilar, or supporting, membrane, which separate it from the scala tympani, and its side by two important tissues lining the bony wall of the cochlea, the delicate stria vascularis and the tough, fibrous spiral ligament. The hypotenuse is formed by the thin membrane of Reissner, consisting of only two layers of flattened cells. A low ridge, the spiral limbus, rests on the margin of the osseous lamina. Reissner's membrane stretches from the inner margin of the limbus to the upper border of the stria, behind which it is inserted (Figure 6).

Upon the surface of the basilar membrane the sensory cells are arranged in more or less orderly rows. Together with their supporting cells they form a complex neuroepithelium called the basilar papilla or organ of Corti, named after the Italian microscopist, the Marchese Al-

fonso Corti, who first described it, in 1851. Viewed in cross section its most striking feature is the arch, or tunnel, of Corti, formed by two rows of pillar cells, or rods. The pillar cells furnish the major support of Corti's organ. They separate a single row of larger, more or less pear-shaped inner hair cells on the inner side of the tunnel from three or more rows of smaller, cylindrical, outer hair cells on the outer side. The inner hair cells are supported and enclosed by the inner phalangeal cells, which rest on the thin outer portion, called the tympanic lip, of the spiral limbus. On the inner side of the inner hair cells and the cells that support them is a curved furrow called the inner sulcus. This is lined with more or less cuboidal cells.

Each outer hair cell is supported by a cell known as a phalangeal cell of Deiters, which holds the base of the hair cell in a cup-shaped depression. From each Deiters cell a projection extends upward to the stiff membrane, the reticular lamina, that covers the organ of Corti. The top of the hair cell is firmly held by the lamina, but the body is suspended in fluid that fills the so-called space of Nuel and the tunnel. Although this fluid is sometimes referred to as the cortilymph, it appears to have the same composition as the perilymph. Beyond the hair cells and the cells of Deiters are epithelial cells of three other types, usually called the cells of Hensen, Claudius, and Böttcher, after the 19th-century microscopists who first described them. Their function has not been established, but they are assumed to help in maintaining the composition of the endolymph by secretory and absorptive activity.

The basilar membrane consists of two zones, the pars arcuata under the arch, and the pars pectinata extending from the feet of the outer pillars to the spiral ligament. The fibres of the membrane are fine and inconspicuous in the pars arcuata. Under the feet of the outer pillars the membrane splits into two distinct layers, the upper one containing stout, parallel, radial fibrils that give the membrane its stiffness. These fibrils decrease in calibre and increase in length from the basal end of the cochlea to the apex, so that the basilar membrane as a whole decreases in stiffness as it increases in width from base to apex.

Beneath the fibrillar layer is the acellular ground substance of the membrane, which is covered in turn by a single layer of mesothelial cells forming the tympanic lamella on the tympanic surface.

The spiral vessels

Capillary blood vessels are found beneath the tympanic lip of the limbus and, in many species, beneath the tunnel. These vessels, called spiral vessels, do not enter the organ of Corti but are thought to furnish the major part of the supply of oxygen and nutrients to its various cells, including the hair cells. Although the outer spiral vessel is seldom found in adult animals of certain species such as the dog and cat and is somewhat irregularly distributed in man, it is always present during fetal development. To judge by its size in the fetus, it plays a major part in bringing blood to the rapidly differentiating organ of Corti.

Each hair cell is capped by a plate, which bears the stereocilia or hairs. On the inner hair cells 40 to 60 stereocilia are arranged in two or more irregularly parallel rows. On the outer hair cells approximately 100 stereocilia form a W pattern. Below the notch of the W the plate is incomplete. At this point is the base of the kinocilium, although the motile "hair" is absent.

The stereocilia are about three to five microns in length and extend upward from the cuticular plate to the underside of the tectorial membrane. This is an acellular, gelatinous structure that covers the top of the limbus spiralis as a thin layer and extends outward over the inner sulcus and the reticular lamina. Its fibrils extend radially and somewhat obliquely to end in finger-like projections at its free margin; these make contact with the stereocilia of the outermost hair cells.

The myelinated (sheathed) fibres of the vestibulocochlear nerve fan out in spiral fashion from the modiolus to pass into a channel near the root of the osseous lamina called the canal of Rosenthal. There, the bipolar cell bodies of these neurons, or nerve cells, are located, forming the spiral ganglion. Beyond the ganglion their dendrites extend radially outward in the lamina beneath the limbus to pass through the small pores of a structure called the habenula perforata, directly under the inner hair cells. At this point the fibres abruptly lose their multilayered coats of myelin and continue as thin, naked, unmyelinated fibres into the organ of Corti. They form a longitudinally directed bundle running beneath the inner hair cells and another at the foot of the inner pillars just inside the tunnel. The majority of the fibres end beneath the inner hair cells, but some of them cross the tunnel to form longitudinal bundles beneath the outer hair cells, on which they eventually terminate.

The endings of the nerve fibres beneath the hair cells are of two distinct types. The larger and more numerous endings contain many minute vesicles (liquid-filled sacs), which are related to impulse transmission at neural junctions. As described below these endings belong to a special bundle of nerve fibres arising in the brainstem and constituting an efferent system or feedback loop. The smaller and less numerous endings contain few vesicles or other cell structures (organelles). They are the terminations of the afferent fibres of the cochlear nerve, transmitting impulses from the hair cells to the brainstem.

The outer membranous wall of the cochlear duct is lined, as has been mentioned above, by the stria vascularis, a dense cellular layer containing a network of capillary blood vessels. The surface cells of the stria, called dark cells, are of secretory type. Interspersed among the dark cells is an intermediate layer of cells called light cells. Several layers of flat basal cells bound the stria and separate it from the spiral ligament.

Reissner's membrane is inserted at the upper margin of the stria. At the lower margin of the stria is the spiral prominence, a low ridge containing its own set of longitudinally directed capillary vessels. Below the prominence is a depression called the outer sulcus, the floor of which is lined by cells of epithelial origin that send long projections into the substance of the spiral ligament. Between these so-called root cells, capillary vessels descend from the spiral ligament. This region appears to have an absorptive rather than a secretory function, and it may be concerned with the removal of waste materials from the endolymph.

The spiral ligament

The spiral ligament lies between the stria vascularis and the bony wall of the cochlea. Extending above the attachment of Reissner's membrane, it is there in contact with the perilymph in the scala vestibuli; and extending below the insertion of the basilar membrane, it is there in contact with the perilymph in the scala tympani. It contains many stout fibres that anchor the basilar membrane, and numerous cells, mainly connective tissue cells (fibrocytes). Behind the stria the structure of the spiral ligament is denser than near the upper and lower margins.

Like the stria, the spiral ligament is well supplied with blood vessels. It receives the radiating arterioles that pass outward from the modiolus in bony channels of the roof of the scala vestibuli. Branches from these vessels form a network of capillaries above Reissner's membrane that is thought to be largely responsible for the formation of the perilymph as an ultrafiltrate of the blood plasma. Other branches enter the stria, and still others pass behind it to the spiral prominence and the floor of the outer sulcus. From these separate capillary networks, which are not interconnected, venules (small veins) descending below the attachment of the basilar membrane collect the blood and deliver it to the spiral vein in the floor of the scala tympani.

Viewed from above, the organ of Corti with its covering, the reticular lamina, forms a well-defined mosaic pattern. In man the arrangement of the outer hair cells in the basal turn of the cochlea is quite regular, with three distinct and orderly rows; but in the higher turns of the cochlea, it becomes increasingly irregular, as scattered cells appear representing incomplete fourth and fifth rows. The spaces between the outer hair cells are filled by the oddly shaped extensions (phalangeal plates)

of the supporting cells. The double row of head plates of the inner and outer pillar cells cover the tunnel and separate the inner from the outer hair cells. The reticular lamina extends from the inner border cells near the inner sulcus to the Hensen cells but does not include either of these cell groups. When a hair cell degenerates and disappears as a result of injury, its place is quickly covered by the adjacent phalangeal plates, which form an easily recognized "scar."

The basilar membrane

The length of the basilar membrane (and of the organ of Corti that covers it) is about 35 millimetres (about 1.4 inches) in man. Its width varies from less than 0.001 millimetre near its basal end to 0.005 millimetre near the apex. The membrane is not under tension but decreases remarkably in stiffness from the base to the apex of the cochlea. Furthermore, at the basal end the osseous lamina is broader, the stria vascularis wider, and the spiral ligament stouter than at the apex. The mass of the organ of Corti, on the other hand, is least at the base and greatest at the apex. These considerations indicate that there is a certain degree of "tuning" provided in the structure of the cochlear duct and its contents. With greater stiffness and less mass, the basal end is more attuned to the higher frequencies of vibrations. Decreased stiffness and increased mass render the apical end more responsive to the lower frequencies.

The total number of outer hair cells in the cochlea has been estimated at 12,000, and the number of inner hair cells at 3,500. Although there are almost 30,000 fibres in the cochlear nerve, there is considerable overlap in the innervation of the outer hair cells. A single fibre may supply endings to many hair cells, which thus share a "party line." Furthermore, a single hair cell may receive nerve endings from many fibres. The actual distribution of the nerve fibres in the organ of Corti has not been worked out in any detail, but the inner hair cells appear to receive the lion's share of the afferent fibre endings (fibres bearing impulses to the brain), with less of the overlapping and sharing of fibres that are characteristic of the outer hair cells.

Functions of the ear

TRANSMISSION OF SOUND WAVES

Transmission to the inner ear. *Air conduction.* The auricle, or visible portion of the outer ear, because of its small size and virtual immobility in man, has lost most of the importance that it has in many animals as an aid in sound gathering and direction finding. For airborne sounds of relatively short wavelength—*i.e.*, those above 3,000 hertz (cycles per second)—the concha serves as a funnel, directing them into the canal. The canal itself contributes little of acoustic importance apart from a broad resonance centred at 3,800 hertz, which helps to determine the frequencies to which the ear is most sensitive.

Acoustic impedance

Sounds reaching the drum membrane are in part reflected and in part absorbed. Only that portion of the sound that is absorbed is effective in setting the drum membrane and the ossicles in motion and thus eventually reaching the inner ear. This tendency of the ear to oppose the passage of sound is called its acoustic impedance. The magnitude of the impedance depends upon the mass and stiffness of the drum membrane and the ossicular chain and on the frictional resistance that they offer. Direct or indirect measurement of the impedance of the ear in hard-of-hearing patients can give important information about the condition of the middle-ear mechanism.

The central portion of the drum membrane vibrates as a stiff cone in response to sound, at least at frequencies below 2,400 hertz. Its motion is transmitted to the handle of the malleus, the tip of which is at the umbo, or centre, of the membrane. At higher frequencies the motions of the drum membrane are no longer simple, and transmission to the malleus may be somewhat less effective. The malleus and incus are suspended by small elastic ligaments and are finely balanced, so that their masses are evenly distributed above and below their common axis of rotation. The head of the malleus and the body of the incus are tightly bound together, with the result that they move in and out as a unit with the movements of the drum membrane. At moderate sound pressures the stapes follows them, and the whole ossicular chain vibrates as a single mass. There may, however, be considerable freedom of motion and some loss of energy at the joint between the incus and the stapes because of the relatively loose coupling. The stapes itself does not move in and out, however, but rocks about the lower pole of its footplate as it transmits the vibrations to the perilymph that fills the vestibule. Its motion thus resembles that of a bell-crank lever rather than that of a piston (Figure 7).

Adapted from S.S. Stevens and H. Davis, *Hearing: Its Psychology and Physiology* (1938); John Wiley and Sons, Inc.

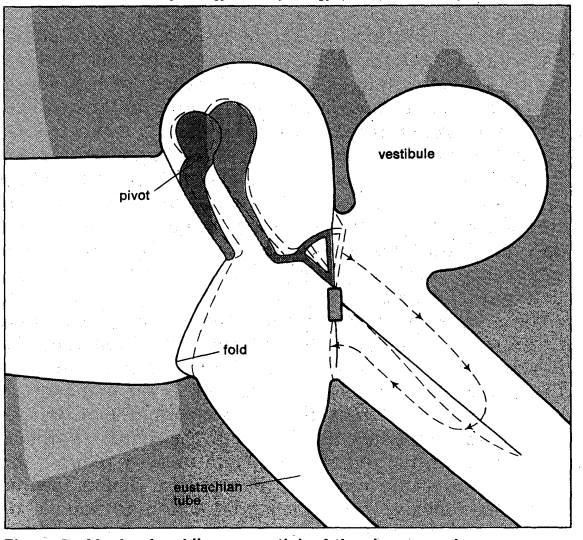


Figure 7: Mechanics (diagrammatic) of the drum membrane and ossicular chain (see text).

The acoustic problem solved by the middle-ear mechanism of drum membrane and ossicular chain is that of transmitting sound from the air to the watery perilymph of the inner ear. Because of the great difference in density between the gas and the liquid, there is a serious mismatch of impedance between them. Ordinarily, this mismatch causes 99.9 percent of airborne sound striking a water surface to be reflected, so that only 0.1 percent passes into the water. In the ear this would represent a transmission loss of 30 decibels, enough to seriously limit its performance, were it not for the transformer action of the middle ear. The matching of impedances is accomplished in two ways, primarily by the reduction in area between the drum membrane and the stapes footplate and secondarily by the mechanical advantage of the lever formed by the malleus and incus. Although the total area of the drum membrane is about 69 square millimetres (0.1 square inch), the area of its central portion that is free to move has been estimated at about 43 square millimetres (0.07 square inch). All of the sound that causes this area of the membrane to vibrate is transmitted and concentrated in the 3.2-square-millimetre (0.005-square-inch) area of the stapes footplate. Thus the pressure (*i.e.*, force per unit area) is increased at least 13 times over. The mechanical advantage of the ossicular lever, because the handle of the malleus is longer than the long projection of the incus, amounts to about 1.3. The total increase in pressure at the footplate is, therefore, not less than 17-fold, depending upon the area of the drum membrane actually vibrating. At frequencies in the range of 3,000 to 5,000 hertz, to which the ear is most sensitive, the increase may be even greater because of the resonant properties of the ear canal as a closed tube.

Matching of impedances

The ossicular chain not only concentrates sound in a small area but also applies sound preferentially to one window of the cochlea—*i.e.*, to the oval window, in which the footplate fits so neatly. If the oval and the round windows were equally exposed to airborne sound

crossing the middle ear, the vibrations in the perilymph of the scala vestibuli would be opposed by those in the perilymph of the scala tympani, and little effective movement of the basilar membrane would result. As it is, sound is delivered selectively to the oval window, and the round window moves in reciprocal fashion, bulging outward in response to an inward movement of the stapes footplate and vice versa. The passage of vibrations through the air across the middle ear from the drum membrane to the round window is of negligible importance.

Thanks to these mechanical features of the middle ear, the hair cells of the normal cochlea are able to respond at the threshold of hearing to vibrations of the tympanic membrane no greater than one angstrom unit (0.0000001 millimetre) in amplitude. On the other hand, when the ossicular chain is immobilized by disease, as in otosclerosis, which causes the stapes footplate to become fixed in the oval window, increase in the threshold of hearing of as much as 60 decibels is common. Bypassing the ossicular chain through the surgical creation of a new window, as in the so-called fenestration operation, can restore hearing to within 25 or 30 decibels of the normal. Only if the stapes is removed and replaced by a tiny artificial stapes can normal hearing be approached. Fortunately, defects of the middle ear causing so-called conductive impairment can often be corrected by surgical means so that useful hearing is restored.

Action of middle-ear muscles

The muscles of the middle ear, the tensor tympani and stapedius, can influence the transmission of sound by the ossicular chain. Contraction of the tensor tympani pulls the handle of the malleus inward and, as the name of the muscle suggests, tenses the drum membrane. Contraction of the stapedius tends to pull the stapes footplate outward from the oval window and thereby reduces the intensity of sound reaching the cochlea. The stapedius responds reflexly with quick contraction to sounds of high intensity applied either to the same ear or to the opposite ear. The reflex has been likened to the eyeblink or constriction of the pupil in response to light and is thought to have protective value. Unfortunately, the contractions of the middle-ear muscles are not instantaneous, so that they do not protect the cochlea against damage by sudden intense noise, as of gunfire. They also fatigue rather quickly and thus offer little protection against injury by sustained high level noise, as in industry. Because dangerously intense noises are rare in nature, it has been argued that the chief biological significance of the aural reflex is not protection of the inner ear but selective high pass filtering of sound. In other words, since the contractions stiffen the ossicular chain, they favour the transmission of the higher frequencies. These carry more information than the lower frequencies, whether in the consonant sounds of human speech or in the rustle of leaves and the snap of a twig that may signal the approach of an enemy in the wild. At the same time, they reduce the transmission of the lower frequencies, which may otherwise mask those of greater significance for survival. The intra-aural muscle reflexes may also have a role in listening and attention, but this function is by no means fully understood.

Bone conduction. There is another route by which sound can reach the inner ear—that is, by conduction through the bones of the skull. When the handle of a vibrating tuning fork is placed on a bony prominence such as the forehead or the mastoid process behind the ear, its note is clearly audible. Similarly, the ticking of a watch held between the teeth can be distinctly heard. When the external canals are closed with the fingers, the sound becomes louder, indicating that it is not entering the ear by the usual channel. Instead, it is producing vibrations of the skull that are passed on to the inner ear, either directly or indirectly.

The higher audible frequencies cause the skull to vibrate in segments, and these vibrations are transmitted to the cochlear fluids by direct compression of the otic capsule, the bony case inclosing the inner ear. Since the round window membrane is more freely mobile than the stapes footplate, the vibrations set up in the perilymph of the scala vestibuli are not cancelled out by those in the scala tympani, and the resultant movements of the basilar membrane can stimulate Corti's organ. This type of transmission is known as compressional bone conduction.

At lower frequencies—*i.e.*, 1,500 hertz and below—the skull moves as a rigid body. The ossicles, because of their inertia and because they are suspended in the middle-ear cavity and only loosely coupled to the skull, are less affected and move less freely than the cochlea and the margins of the oval window. The result is that the oval window moves with respect to the footplate of the stapes, which gives the same effect as if the stapes itself were vibrating. This form of transmission is known as inertial bone conduction. The fixed stapes in otosclerosis interferes with this type of bone conduction, but not with compressional bone conduction.

Hearing aids for conduction deafness

In the presence of middle-ear disease, hearing aids with special vibrators are sometimes used to deliver sound to the mastoid process (the part of the temporal bone that is behind the ear) and then by bone conduction to the inner ear. Bone conduction is also the basis of some of the oldest and most useful tests in the repertory of the otologist, those employing tuning forks to distinguish between a conductive impairment affecting the middle ear and amenable to surgery and a sensorineural impairment—that is, an impairment affecting the inner ear and the cochlear nerve—for which surgery is usually not indicated. In the Rinne test the sounding tuning fork is placed on the mastoid process, and the person being tested is asked to report when he can no longer hear it. The examiner then removes the fork quickly and holds the prongs close to the open ear canal. The normal ear continues to hear it for about 45 seconds, and this "positive" result occurs also with incomplete sensorineural impairment of hearing. When the result is "negative" and the fork is heard longer by bone conduction than by air conduction, a conductive type of deafness is present. In the Schwabach test the presence of a sensorineural impairment is indicated when the patient is unable to hear the bone-conducted sound as long as the examiner with normal hearing can. The patient with a conductive hearing loss, on the other hand, can hear the fork longer than can the examiner because the conductive lesion excludes the extraneous masking noise of the surroundings. A bone-conduction audiometer would give the same result.

For the Weber test, the fork is simply placed on the patient's forehead, and the examiner asks in which ear he hears it. If a sensorineural lesion is present in one ear, the patient will localize the sound in the opposite, or "better," ear. If a conductive lesion is present, he will localize it in the "worse" ear—*i.e.*, in the one that is protected from interference by extraneous sounds. This simple test is a valuable aid in the diagnosis of otosclerosis.

Transmission through liquid of inner ear to organ of Corti. The vibrations of the stapes footplate at the oval window are transmitted through the perilymph of the vestibule and that of the scala vestibuli to the cochlear duct. Normally, they do not affect the semicircular canals, the utricle, or the saccule. Only the cochlear fluids and the membranes of the cochlear duct are free to vibrate in response to alternating pressures at the oval window, because only the cochlea has the round window as a "relief valve."

What takes place in the cochlea is both an analysis or sorting out of the frequencies of a complex sound and a transduction of the mechanical vibrations into nerve impulses for transmission to the brainstem by the fibres of the cochlear nerve. The analysis occurs because the cochlea is a tuned structure, different parts of which vibrate in response to different frequencies; the transduction because the hair cells of Corti's organ are capable of changing minute amounts of mechanical energy into an electrochemical form that stimulates the endings of the nerve fibres.

The idea of the inner ear as a resonant structure was proposed by several authors in the 17th and 18th centuries but received its most explicit statement just over a

century ago by the German physicist and physiologist Hermann von Helmholtz. Inspired by the anatomical studies of Corti, Helmholtz postulated in the cochlea a series of resonators capable of analyzing complex sounds into their component frequencies. After considering various other structures, he concluded that the transverse fibres of the basilar membrane, which increase in length and decrease in stiffness from the basal end to the apex, were the resonators he sought. Although Helmholtz' resonance theory in its original form is no longer accepted, much experimental and clinical evidence supports the closely related "place theory," which holds that sounds of different frequency activate different regions of the basilar membrane and organ of Corti, there being an orderly progression from the basal end, where the highest tones are effective, to the apical end, where the lowest tones are received.

The ear's analysis of complex sounds

From the experiments of Georg von Békésy (for which he was awarded the Nobel Prize for Physiology or Medicine in 1961) it is clear that the frequency analysis performed by the cochlea occurs not because of a series of separate tuned resonators as postulated by Helmholtz but because the basilar membrane and organ of Corti vary continuously in stiffness and mass from base to apex, stiffness decreasing and mass increasing as the width of the basilar membrane increases. Vibrations reaching the basal end through the perilymph can be shown to proceed as travelling waves, attaining a maximum amplitude at a certain point along the membrane and then rapidly subsiding. The higher the frequency of the sound imposed, the shorter is the distance that the waves travel. Since the hair cells are arrayed in orderly ranks on the basilar membrane and the nerve fibres fan out in orderly fashion to innervate them, a tone of a given frequency causes a certain "place" on the basilar membrane to vibrate, the hair cells overlying it are stimulated, and the corresponding nerve fibres convey impulses to the brain. The brain can recognize the place on the basilar membrane and thus the pitch of the stimulating tone, depending upon the particular group of nerve fibres activated. For the lower frequencies the rate of stimulation is also an important indicator of pitch because the frequency of the nerve impulses tends to follow the frequency of the tone. For the higher frequencies, place alone seems to be decisive. The intimate nature of the events that occur at the hair cells when they are stimulated by movements of the basilar membrane is by no means fully understood. The up-and-down movements of the basilar membrane are thought to be converted into shearing movements between the reticular lamina and the tectorial membrane, which overlies it. The stereocilia, or hairs of the hair cells, in contact with the tectorial membrane are displaced or bent by the shearing forces, and it is this bending of the stereocilia that triggers the electrochemical events in the hair cells that excite the endings of the cochlear nerve fibres. Present understanding of the events occurring in the cochlea is based in part on von Békésy's experiments with cochlear models, in part on his direct microscopic observations by stroboscopic illumination of vibratory patterns in the human cochlea removed from the body after death and subjected to intense sound. Important insights have also been obtained through the study of the small electrical potentials that are produced by the living cochlea in response to sound. These alternating current potentials, first reported in 1929, reproduce the frequency and wave form of the stimulating sound and can be picked up by means of two electrodes, one of which is placed in contact with the round window membrane or with the bony wall of the cochlea. The potentials were shown to consist of two separate components, the microphonic potentials, which appear to be related to the transduction process at the hair cells, and the action potentials, which represent the nerve impulses in the terminal fibres of the cochlear nerve. There is also a direct current potential difference, the endocochlear potential of some 80 millivolts between the endolymph in the cochlear duct and the perilymph in the scala tympani. The alternating current potentials depend upon an intact organ of Corti, the direct current potential upon the stria vascularis. All require an adequate supply of oxygen to the cochlea and soon disappear when oxygen is lacking. When the organ of Corti is damaged by drugs or by intense sound, the potentials are diminished or abolished. In young children suspected of congenital deafness, the cochlear potentials can give important information about the state of the inner ear.

Electrical potentials of the cochlea

THE COCHLEAR NERVE AND THE CENTRAL AUDITORY PATHWAYS

The vestibulocochlear (acoustic, or eighth cranial) nerve consists of two anatomically and functionally distinct parts, the cochlear nerve, distributed to the organ of hearing, and the vestibular nerve, distributed to the organs of equilibrium. The fibres of the cochlear nerve have their peripheral terminals around the bases of the inner and outer hair cells and their central terminals in the groups of nerve cells called the dorsal and ventral cochlear nuclei, in the medulla oblongata, at the base of the brain. The cell bodies of these neurons (nerve cells), numbering some 30,000 in all, fill the spiral canal of Rosenthal at the root of the osseous lamina of the cochlea. Their peripheral, dendritic portions extend radially in the lamina to the habenula perforata, beneath the inner hair cells. At this point they lose their myelin sheaths and enter the organ of Corti as thin, unmyelinated fibres. The greater number are distributed as radial fibres directly to the inner hair cells, whereas others cross the tunnel either as radial fibres supplying a few outer hair cells or as spiral fibres that turn to run for relatively long distances and make contact with many outer hair cells. The central or axonal portions of the bipolar cochlear neurons unite to form the cochlear nerve trunk, the fibres of which are twisted like the strands of a rope. Leaving the modiolus through the internal meatus or passageway, they pass directly to the medulla. There each fibre divides into two, sending one branch to the dorsal and the other to the ventral cochlear nucleus.

Central auditory pathways

The central auditory pathways extend from the medulla to the cerebral cortex and consist of a series of nuclei (groups of nerve cell bodies) connected by fibre tracts made up of their axons. They form a more or less direct route for relaying acoustic information, encoded in the form of nerve impulses, directly to the highest cerebral levels in the cortex. At lower levels information as to pitch, loudness, and localization of sounds is processed, and appropriate responses, such as contractions of the intra-aural muscles, turning of the eyes and head, or movements of the body as a whole, are initiated.

Some fibres from the ventral cochlear nucleus pass across the midline to the cells of the superior olivary complex, whereas others make connection with the olivary complex of the same side. Together, these fibres form the trapezoid body. Fibres from the dorsal cochlear nucleus cross the midline to end on the cells of the nuclei of the lateral lemniscus. There they are joined by fibres from the ventral cochlear nuclei of both sides, and from the olivary complex. The lemniscus is a major tract, most of the fibres of which end in the inferior colliculus, the auditory centre of the midbrain, although some fibres may by-pass the colliculus and end at the next higher level, the medial geniculate body. From the medial geniculate there is an orderly projection of fibres to a portion of the cortex of the temporal lobe.

In man and other primates the primary acoustic area in the cerebral cortex is in the superior transverse temporal gyrus, a ridge in the temporal lobe, on the lower edge, or lip, of a cleft known as the sylvian fissure.

Because about half of the fibres of the auditory pathways cross the midline while others ascend on the same side, each ear is represented in both the right and the left cortex. For this reason both ears can continue to function normally, even if the auditory cortical area of one side is destroyed. Impaired hearing due to bilateral cortical injury involving both auditory areas has been reported, but it is extremely rare.

Parallel with the pathway ascending from the cochlear nuclei to the cortex is a pathway descending from the

cortex to the cochlear nuclei. In both pathways some of the fibres remain on the same side, while others cross the midline to the opposite side of the brain. There is also evidence of a "spur" line ascending from the dorsal cochlear nucleus to the cerebellum and another descending from the inferior colliculus, the auditory centre of the midbrain, to the cerebellum. The significance of these cerebellar connections is not clear, but they may antedate the development of the cerebral cortex. In general, the descending fibres may be regarded as exercising an inhibitory function by means of a sort of "negative feedback." They may also determine which ascending impulses shall be blocked and which shall be allowed to pass on to the higher centres of the brain.

From the olivary complex, a region in the medulla oblongata, there arises also a fibre tract called the olivocochlear bundle. It constitutes an efferent system or feedback loop, by which nerve impulses, thought to be inhibitory in nature, reach the hair cells. This system, which apparently utilizes acetylcholine as a chemical synaptic transmitter, is presumably involved in sharpening the analysis that is made in the cochlea.

Evidence of an orderly spatial representation of the organ of Corti at the lower levels of the auditory pathway has been reported by many investigators. Such a pattern would seem to be required by the place theory of cochlear analysis of sound.

Physiological evidence of "tuning" of the auditory system has also been obtained by recording the electrical potentials from individual neurons at various levels by means of microelectrodes. Most neurons of the auditory pathway show a "best frequency"—*i.e.*, a frequency to which the individual neuron responds at minimal intensity. With each increase in the intensity of the sound stimulus, the neuron is able to respond to a wider band of frequencies, thus reflecting the broad tuning of the basilar membrane. With sounds of lower frequency, the rate of response of the neuron tends to reflect the stimulus frequency. Increased intensity of stimulation causes a more rapid rate of responding. In general, pitch tends to be coded in terms of the neurons that are responding, loudness in terms of the rate of response and the total number of active neurons.

Auditory function of the cortex

Although extensive studies have been made of the responses of single cortical neurons, the data do not yet fit any comprehensive theory of auditory analysis. Experiments in animals have indicated that the cortex is not even necessary for frequency recognition, which can be carried out at lower levels, but that it is essential for the recognition of temporal patterns of sound. It appears likely, therefore, that in man both pitch and loudness are distinguished at lower levels of the auditory pathways and that the cortex is reserved for the analysis of more complex acoustic stimuli, such as speech and music, for which the temporal sequence of sounds is equally important.

Presumably it is also at cortical levels that the "meaning" of sounds is recognized and behaviour is adjusted in accordance with their significance. Such functions were formerly attributed to an "auditory association area" immediately surrounding the primary area, but such a term might equally well be interpreted as embracing the entire cerebral cortex, thanks to the multiple interconnections between the various areas.

The localization of sounds is known to depend upon the recognition of minute differences in intensity and in the time of arrival of the sound at the two ears. A sound that arrives at the right ear a few microseconds sooner than it does at the left or that sounds a few decibels louder in that ear is recognized as coming from the right. In a real-life situation the head may also be turned to pinpoint the source of sound by maximizing these differences. For low-frequency tones a difference in phase at the two ears is the criterion for localization, but for higher frequencies the difference in loudness caused by the sound shadow of the head becomes all-important.

Each cochlear nucleus receives impulses only from the ear of the same side. A comparison between the responses of the two ears first becomes possible at the superior olivary complex, which receives fibres from both ears. Electrophysiological experiments in animals have shown that some units of the accessory nucleus of the olivary complex respond to impulses from both ears. Others respond to impulses from one side exclusively, but their response is modified by the simultaneous arrival of impulses from the other side.

The superior olivary complex

This system appears to be capable of making the extraordinarily fine discriminations of time and intensity that are necessary for sound localization. By virtue of such complex neural interconnections in the brain, the two ears together can be much more effective than one ear alone in picking out a particular sound in the presence of a background of noise. They also permit attention to be directed to a single source of sound, such as one instrument in an orchestra or one voice in a crowd. This is the basis of the "cocktail-party effect," whereby a listener with normal hearing can attend to different conversations in turn or concentrate on one speaker despite the surrounding babble. Whether the muscles within the ear play a part in filtering out unwanted sounds during such selective listening has not been established.

HEARING TESTS: AUDIOMETRY

Before the development of electroacoustic equipment for generating and measuring sound, the tests of hearing at the disposal of the otologist gave approximate answers at best. A patient's ability to hear could be specified in terms of whether he could distinguish the ticking of a watch or the clicking of coins or at what distance he could understand a whispered voice. Alternatively, the examiner might note the length of time a patient could hear the gradually diminishing note of the tuning fork, comparing the performance with his own. Other specialized tuning fork tests have been described above.

The electronic audiometer, introduced in the 1930s, makes it possible to measure the patient's threshold of hearing for a series of pure tones ranging from a lower frequency of 125 hertz to an upper frequency of 8,000 or 10,000 hertz. This span includes the three octaves between 500 and 4,000 hertz that are most important for speech.

Components of an audiometer

The audiometer consists of an oscillator or signal generator, an amplifier, a device called an attenuator, which controls and specifies the intensity of the tones produced, and an earphone. The intensity range is usually 100 decibels in steps of five decibels. The "zero dB" level represents "normal hearing" for young adults under favourable, noise-free laboratory conditions and was established in 1964 as an international standard.

For pure-tone audiometry, the patient wears the earphone and is asked simply to indicate when he hears a tone. The examiner proceeds to determine the lowest intensity for each frequency at which the patient reports that he is just able to hear the tone 50 percent of the time. If, for example, he hears 4,000 hertz only at 40 decibels, he is said to have a 40-decibel hearing level for that frequency—*i.e.*, a threshold 40 decibels above the normal threshold. A graph showing the hearing level for each frequency is called an audiogram. The shape of the audiogram for a hard-of-hearing patient can give the otologist or audiologist important information for diagnosing the nature and cause of the defective hearing. (The otologist is concerned with diseases of the ear; the audiologist focusses his attention on the measurement of hearing impairment.)

With the Békésy automatic recording audiometer, the patient himself controls the level of the tone presented. He is required to press a button so long as he is able to hear the tone, which is automatically reduced in intensity as the button is pressed. When it is released the intensity increases until the patient again signals that he hears it by pressing the button. In this way the patient's threshold is "tracked." At the same time the frequency is slowly increased and a graphic recording is made. Thus a complete, continuous audiogram can be obtained in less than ten minutes.

A calibrated bone conduction vibrator is usually furnished with the audiometer so that hearing by bone con-

duction can also be measured. In the presence of otosclerosis or other conductive defect of the middle ear, there may be a sizable difference between the air-conduction and bone-conduction audiograms, the so-called "air-bone gap." This difference is a measure of the loss in transmission across the middle ear and indicates the maximum improvement that may be obtained through successful corrective surgery. When the defect is confined to the organ of Corti, the bone-conduction audiogram shows a loss similar to that for air conduction. In such cases of sensorineural impairment, surgery is seldom if ever capable of improving the hearing, but a hearing aid may prove useful.

Although faint sounds may not be heard at all by the ear with a sensorineural impairment, more intense sounds may be as loud as to a normal ear. This rapid increase in loudness above the threshold level is called recruitment. When the opposite ear has normal hearing, recruitment can be measured by the alternate binaural loudness balance (ABLB) test. The subject is asked to set the controls so that the loudness of the tone heard in his defective ear matches that of the tone heard in his normal ear. By repetition of the comparison at several intensity levels, the presence or absence of recruitment can be demonstrated. When recruitment is excessive, the range of useful hearing between the threshold and the level at which loudness becomes uncomfortable or intolerable may be extremely narrow, so that the amplification provided by a hearing aid is of limited value to the patient.

Speech reception threshold and discrimination

Although hearing thresholds for pure tones give some indication of the patient's hearing for speech, direct measurement of this ability is of interest to the otologist. Two types of tests are most often used. In one, the speech reception threshold (SRT) is measured by presenting words of spondee pattern—words containing two syllables of equal emphasis, as in "baseball" or "cowboy"—at various intensity levels until the level is found at which the patient can just hear and repeat half of the words correctly. This level usually corresponds closely to the average of the patient's thresholds for frequencies of 500, 1,000, and 2,000 hertz. A more important measure of socially useful hearing is the discrimination score. For this test a list of selected monosyllabic words is presented at a comfortable intensity level, and the subject is scored in terms of the percentage of the words heard correctly. This test is helpful in evaluating certain forms of hearing impairment in which the sounds may be audible but words remain unintelligible. Such tests are usually carried out in a quiet, sound-treated room that excludes extraneous noise. They may give a mistaken impression of the ability of the patient with sensorineural impairment to understand speech in ordinary noisy surroundings. Because such persons often have increased difficulty in understanding speech in the presence of noise, speech tests are best carried out against a standardized noise background as well as in the quiet. A person with a conductive defect may be less disturbed by noise than may the normal subject.

Tests suitable for young persons

For very young persons or others who are unable to cooperate in the usual audiometric tests, thresholds for pure tones may be established by electrophysiological means. A brief tone causes a small variation in the electrical potentials that can be recorded from the scalp and constitute the electroencephalogram, EEG (a recording of the brain waves). By repetition of the stimulus up to 100 times and by an averaging of the responses in a small computer, the responses can be selectively enhanced while the more or less random background of electrical activity is cancelled out. In this way auditory thresholds can be established that closely approximate those obtained in conventional audiometry. If the responses indicate impaired hearing, they give little or no indication of the site of the defect.

Another form of electrophysiological hearing test is the electrocochleogram. Electrical potentials representing impulses in the cochlear nerve are recorded from the cochlea by means of a fine, insulated needle electrode inserted through the eardrum membrane to make contact with the promontory of the basal turn. This test gives a direct measure of cochlear function.

More elaborate tests, often involving speech or sound localization, are available for testing hearing impaired by defects of the central nervous system. The interpretation of the results is often difficult, and the diagnostic information furnished by the tests is seldom clear-cut.

A simple and objective means of testing hearing at the level of the cochlea and brainstem is supplied by impedance audiometry. Two small tubes are sealed into the external canal. Through one tube sound from a small loudspeaker is injected into the canal. The portion that is reflected from the drum membrane is picked up by the other tube and led to a microphone, amplifier, and recorder. When a sudden, moderately intense sound is applied to the opposite ear, the stapedius muscle contracts, the impedance is increased, and the recorder shows a slight excursion as more sound is reflected. This test can give valuable information not only about the condition of the cochlea and the auditory pathways of the medulla but also about the facial nerve that supplies fibres to the stapedius muscle. On the other hand, it does not give an actual measurement of the acoustic impedance of the ear, representing the state of the ossicular chain and the mobility of the eardrum. This information can be obtained by means of the acoustic bridge—a device that enables the observer to listen to a sound as reflected from the subject's eardrum and at the same time to a similar sound of equal intensity as reflected in an artificial cavity, the volume of which is adjusted to equal that of the external canal of the ear being tested. When the two sounds are matched by varying the acoustic impedance of the cavity, the impedance of the ear is equal to that of the cavity, which can be read directly from the scale of the instrument. Conductive defects of the middle ear, including discontinuity (disarticulation) of the ossicular chain and immobility of the malleus or stapes, can be recognized by the characteristic changes they cause in the impedance of the ear.

Conductive impairment of hearing can occur from so simple a cause as an excessive accumulation of wax in the external canal. It is a common result of infections of the middle ear (otitis media), especially when perforation of the drum membrane leads to an ingrowth of epithelial cells with the formation of a mass called a cholesteatoma. In such cases the ossicles may be eroded or destroyed. As previously mentioned, otosclerosis can seriously limit the transmission of sound to the cochlea by causing the stapes footplate to become sealed in the oval window.

Presbycusis

Sensorineural impairment involving the organ of Corti and the cochlear nerve fibres of the basal turn occurs in aging and is the most common form of hearing loss in elderly persons (presbycusis). In some older patients the central auditory pathways may also be affected by a reduced blood supply to the brain. Injury to the cochlea and its nerve supply occurs also as a result of exposure to intense noise, whether in industry, military activity, sport, or entertainment. Noise levels above 90 decibels are generally recognized as dangerous to the hearing. The higher the level and the more prolonged the exposure, the greater is the risk of permanent changes in the inner ear (Figure 8).

Several well-known antibiotics, including streptomycin, kanamycin, neomycin, and gentamicin, can cause irreversible damage to the organ of Corti. Other drugs that may cause sensorineural impairment of hearing include quinine, aspirin and other salicylates in large doses, and certain diuretics and antitumour agents.

In Ménière's syndrome, variable sensorineural loss and tinnitus (noise or ringing in the ear) are associated with severe attacks of vertigo. Tumours of the vestibulocochlear nerve can affect hearing by compressing the nerve fibres and impairing their ability to transmit impulses.

Profound sensorineural deafness can occur as a result of viral and other infections, including mumps, measles, and meningitis. Rubella (German measles) in the mother during pregnancy can cause the child to be born with a severely damaged organ of Corti and profound hearing

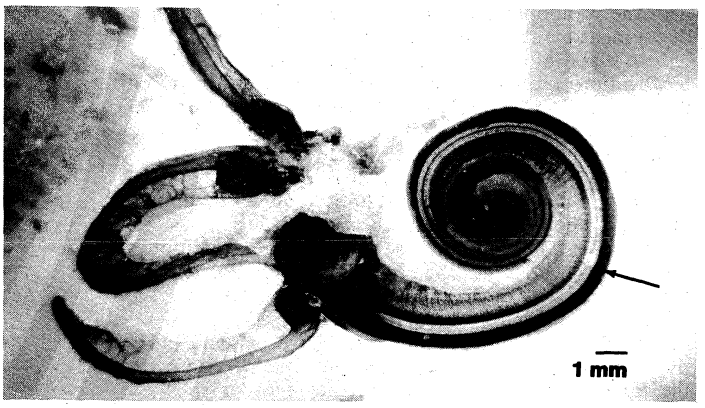


Figure 8: Dissection of the human cochlea and semicircular canals. The arrow indicates an area of degeneration of Corti's organ and the cochlear nerve fibres in the basal turn, corresponding to a loss of hearing for frequencies of 4,000 hertz and higher, presumably caused by noise exposure.
By courtesy of Lars-Goran Johnsson

loss. Cochlear abnormalities may be present also as a result of genetic defects. In all such cases of deafness in young children it is essential that the condition be recognized as early as possible so that special educational measures may be instituted to minimize the handicap. The newer electrophysiological hearing tests described above—*i.e.*, electroencephalographic audiometry and the electrocochleogram—offer the best possibilities for detecting such losses in infants.

VESTIBULAR FUNCTION

Equilibrium sense

The vestibular system represents the equilibrium sense and is concerned primarily with controlling the position of the head and the posture of the body. The adequate stimulus for its end organs is acceleration: angular acceleration for the semicircular canals; linear acceleration for the utricle and saccule. Functionally these organs are closely related to the cerebellum and to the reflex centres that govern the movements of the eyes, the neck, and the limbs. The information they deliver is proprioceptive in character, dealing with events within the body itself, rather than exteroceptive, dealing with events outside the body, as in the case of the responses of the cochlea to sound. Although the vestibular organs and the cochlea are derived embryologically from the same formation, the otic vesicle, their association seems to be a matter more of convenience than of necessity. Both from the developmental and from the structural point of view, the kinship of the vestibular organs with the lateral line system of the fishes is easily recognized. This system is made up of organs located in the skin at the side of the body and contains cristae with innervated hair ells and a cupula, like those of the semicircular canals. They are sensitive to waterborne vibrations and to pressure changes.

The anatomists of the 17th and 18th centuries were impressed by the orientation of the semicircular canals in the head, in three planes more or less perpendicular to one another. Since it was generally assumed that the entire labyrinth is devoted to hearing, the suggestion was put forward that the canals must perceive the direction of sound.

The first investigator to present evidence that the vestibular labyrinth is the organ of equilibrium was a French experimental neurologist, Marie-Jean-Pierre Flourens, who, in 1824, reported a series of experiments in which he had produced abnormal head movements in the pigeon by cutting each of the semicircular canals in turn. The plane of the movements was always the same as that of the injured canal. Hearing was not affected by destroying the nerve fibres to the vestibular organs, but it was abolished by cutting the cochlear nerve.

It was almost half a century later that the significance of these experiments was appreciated and the semicircular canals were recognized as specific sense organs concerned with the position of the head. The German physiologist Friedrich Goltz in 1870 suggested in his "hydrostatic concept" that the canals are stimulated by the weight of the fluid they contain, the pressure it exerts varying with the head position. Three years later the Austrian scientists Ernst Mach and Josef Breuer, and the Scottish chemist Crum Brown, working independently, proposed their "hydrodynamic concept." They held that head movements cause a flow of endolymph in the canals, which are stimulated by the fluid movement or pressure changes. A German physiologist, J.R. Ewald, showed that the compression of the lateral canal in the pigeon by a small pneumatic hammer caused endolymph movement toward the crista and turning of the head and eyes toward the opposite side. Decompression reversed both the direction of endolymph movement and the turning of the head and eyes.

Endolymph and the "hydrodynamic concept"

Later investigators have proved that the hydrodynamic concept is correct and that the cupula is deflected by endolymph movement in response to rotation. They were able to keep track of the deflection in live fish by means of a droplet of oil injected into the canal. At the start of rotation in the plane of the canal, the cupula was deflected in the opposite direction and then returned slowly to its position of rest. At the end of rotation it was deflected again, this time in the same direction as the rotation, and then returned once more to its upright resting position. These deflections were due to the inertia of the endolymph, which tended to lag behind at the start of rotation and to continue in motion after rotation had stopped. The slow return was a function of the elasticity of the cupula itself.

Other researchers found that they could keep the labyrinth of a cartilaginous fish, the thornback ray (*Raja clavata*), active for some time after it was removed from the animal and could record from the vestibular fibres of the vestibulocochlear nerve impulses arising in one of the ampullar cristae. At rest there was a slow continuous discharge, which was increased by rotation in one direction and decreased by rotation in the other. In other words, the level of excitation rose or fell depending upon the direction of rotation. Later studies with the electron microscope showed a remarkable polarization of the hair cells of the ampullar cristae. Each hair cell of the horizontal canals has its kinocilium facing toward the utricle, whereas each hair cell of the vertical canals has its kinocilium facing away from the utricle. In the horizontal canals deflection of the cupula toward the utricle—*i.e.*, bending of the stereocilia toward the kinocilium—depolarizes the hair cells and increases the rate of discharge. Deflection away from the utricle causes hyperpolarization and decreased discharge. In the vertical canals these effects are reversed.

Polarization of hair cells

The relation between the two labyrinths is reciprocal. When the head is turned to the left, the discharge from the left horizontal canal is increased while that from the right horizontal canal is decreased, and vice versa. Normal posture is the result of their acting both in cooperation and in opposition. When one labyrinth is injured, the unrestrained activity of the other causes a continuous false sense of turning (vertigo) and rhythmical, jerky movements of the eyes (nystagmus), both toward the normal side. When both labyrinths are injured or destroyed, as by the action of the antibiotic streptomycin, there may be a serious disturbance of posture and gait (ataxia) as well as severe vertigo and general disorientation. In younger persons the disturbance tends to subside as reliance is placed on vision an on proprioceptive impulses from the muscles and joints to compensate for the loss of information from the labyrinth. In elderly persons, the loss of labyrinthine function may be disabling.

The two otolith organs or gravity receptors, the maculae of the utricle and saccule, are more or less perpendicular to each other. The left and right utricular maculae lie in the same approximately horizontal plane; the saccular maculae in parallel, vertical planes. Both types of receptors are stimulated by shearing forces between the otolithic membrane and the cilia of the hair cells beneath it. The otolithic membrane is covered with a layer of crystals of calcite (otoconia), which add to its weight

and increase the shearing forces set up in response to a slight displacement when the head is tilted. These receptors, particularly the utricle, have an important role in the righting reflexes and in reflex maintenance of tonic contraction of the muscles that keep the body in the upright position. The role of the saccule is less completely understood, and some investigators have suggested that it is responsive to vibration as well as to rectilinear acceleration of the head in the sagittal (fore-and-aft) plane. Of the two receptors, the utricle appears to be the dominant partner.

Tests of vestibular function depend mainly upon stimulation of the semicircular canals. Rotation, which can cause vertigo and nystagmus, as well as temporary disorientation and a tendency to fall, stimulates both labyrinths simultaneously. Since the otoneurologist is usually more interested in examining the right and left labyrinths separately, he usually employs warmth as a stimulant. Syringing the ear canal with warm water at 44° C (111° F) or cool water at 30° C (86° F) elicits nystagmus by setting up convection currents in the horizontal canal. The duration of the nystagmus may be timed with a stopwatch, or the movements of the eyes can be accurately recorded by picking up the direct current potentials of the eyeballs by means of electrodes pasted to the skin of the temples (a process called electronystagmography, or ENG). An abnormal labyrinth usually gives a reduced response or none.

Caloric stimulation

Vertigo, the false sense of rotation, is experienced in Menière's syndrome and in many other abnormal conditions involving the labyrinth, the vestibular nerve, and the vestibular centres of the medulla oblongata. Though by no means uncommon, it can present a difficult diagnostic problem for the physician. In normal subjects, sudden movements of the head during rotation can elicit the so-called Coriolis reaction, which involves severe vertigo, disorientation, nystagmus, nausea, and vomiting. This is one of the potential hazards of daily life in a rotating space station.

BIBLIOGRAPHY. Reliable and readable introductory treatises are S.S. STEVENS and FRED WARSHOFSKY, *Sound and Hearing* (1965); and W.A. VAN BERGEIJK, J.R. PIERCE, and E.E. DAVID, *Waves and the Ear* (1960). HALLOWELL DAVIS and S.R. SILVERMAN, *Hearing and Deafness*, 3rd ed. (1970), although somewhat more technical, is also written for the non-specialist. A useful account of the anatomy of the ear may be found in WILLIAM BLOOM and D.W. FAWCETT, *A Textbook of Histology*, 9th ed. (1968). B.J. ANSON and J.A. DONALDSON, *The Surgical Anatomy of the Temporal Bone and Ear* (1967), presents the gross and microscopic structure of the ear from the point of view of the surgeon. Details of inner ear anatomy can be studied in SALVATORE IURATO, *Submicroscopic Structure of the Inner Ear* (1967); HANS ENGSTROM, H.W. ADES, and ANTON ANDERSSON, *Structural Pattern of the Organ of Corti* (1966); and HEINRICH SPOENDLIN, *The Organization of the Cochlear Receptor* (1966). For the comparative anatomy of the ear from fish to man, the illustrations in GUSTAF RETZIUS' two folio volumes *Das Gehörorgan der Wirbelthiere* (1881–84), are still unsurpassed, but available only in a few medical libraries. The physiology of hearing and of vestibular function is discussed in four chapters by JOSEPH E. HAWKINS in C.H. BEST and N.B. TAYLOR (eds.), *The Physiological Basis of Medical Practice*, 8th ed. (1966). Modern classics in the field of hearing are: S.S. STEVENS and HALLOWELL DAVIS, *Hearing* (1938, reprinted 1960), equally divided between psychological and physiological aspects of the subject; E.G. WEVER, *Theory of Hearing* (1949, reprinted 1970), including a good historical treatment of auditory theories developed through the centuries; and E.G. WEVER and MERLE LAWRENCE, *Physiological Acoustics* (1954), concerned mainly with middle-ear and inner-ear mechanics. G. VON BEKESY, *Experiments in Hearing* (1960), is the best source of information about the experimental work that won von Békésy the Nobel Prize, but it is not recommended for the novice. Biochemical aspects of cochlear physiology and pathology are treated in SIGURD RAUCH, *Biochemie des Hörorgans* (1964); and in *Biochemical Mechanisms in Hearing and Deafness*, ed. by MICHAEL M. PAPARELLA (1970). I.C. WHITFIELD, *The Auditory Pathway* (1967), gives a useful summary of our imperfect knowledge of auditory processes in the central nervous system.

The measurement of human hearing is described in numerous textbooks of audiology. *Audiometry: Principles and Practices*, ed. by ARAM GLORIG (1965), gives a good overview of the field. A thorough, sober, and factual account of studies concerning *The Effects of Noise on Man* has been written by K.D. KRYTER (1970). It is recommended as an antidote to some of the more hysterical popular writings on this controversial topic. There are few comprehensive publications devoted to vestibular function, but the symposium volume, ed. by R.J. WOLFSON, *The Vestibular System and Its Diseases* (1966), is a good source of information, as is the NASA series ed. by ASHTON GRAYBIEL, *The Role of the Vestibular Organs in Space Exploration*, 4 vol. (1965–68).

(J.E.H.)

Ear Diseases and Hearing Disorders

The ear has two important but quite different functions: the maintenance of equilibrium and hearing. For biped locomotion, the sense of equilibrium is especially necessary; and it is even more pressing in the dark, in which the eyes cannot help to maintain balance, or on uneven terrain, where sensations from the feet do not assist in maintaining the sense of equilibrium. Hearing, which developed later than equilibrium in the process of evolution, has become, with sight, one of the senses most needed by animals for flight and survival. For civilized man, hearing is also the chief means of receiving communication by language and is so taken for granted that its importance is rarely appreciated until it is impaired or lost. Indeed, Helen Keller, who was totally blind and deaf from early childhood, wrote:

> the problems of deafness are deeper and more complex, if not more important than those of blindness. Deafness is a much worse misfortune, for it means the loss of the most vital stimulus—the sound of the voice—that brings language, sets thoughts astir and keeps us in the intellectual company of man.

After a short survey of the development, structure, and functioning of the normal ear, this article deals with the more important diseases and disorders of the outer, the middle, and the inner ear. It concludes with brief sections on the causes of deafness and impaired hearing, on rehabilitation of hearing impairment and deafness, and on the social and economic implications of deafness.

The normal ear

EVOLUTION AND DEVELOPMENT OF THE EAR

The complicated structure of the mechanism for hearing is best understood by considering the way in which the human ear evolved from lower forms and how it develops in the human embryo, for one of the time-honoured apothegms in biology, "ontogeny recapitulates phylogeny," applies particularly to the ear, the development of which in the human embryo (ontogeny) retraces the stages of evolution from lower forms (phylogeny).

The earliest beginning of the human ear may be seen in the three-week-old embryo as a thickening of the epithelium (primitive skin) on each side of the head. The centre of this thickened disk soon grows inward to form a pit; then the mouth of the pit closes off, leaving a rounded fluid-filled structure called an otocyst; this is destined to become the inner ear, which will contain the nerve endings of the hearing (auditory) nerve and the balancing (vestibular) nerve.

Formation of the outer and the middle ear

Soon after the formation of the otocyst, the outer and middle ears begin to develop. These portions of the hearing mechanism are necessary for the transmitting of airborne sound from the outer world to the fluid-filled inner ear, in which the sound vibrations can be transformed to nerve impulses.

About 260,000,000 years ago, certain fishes, man's ancestors, left the oceans and swamps to live on dry land. To extract oxygen from the air, they developed lungs. The discarded gills became transformed into their outer and middle ear by one of the most remarkable salvage adaptations of the entire process of evolution.

These earliest lung-breathing fish that crawled out onto land were comparatively deaf, for only 0.1 percent of sound energy could reach the fluid of their inner ears, 99.9 percent being reflected away by the air–water sound barrier. This barrier is familiar to an underwater swimmer, who with his head under water clearly hears faint sounds originating underwater, like two stones clicked

together, but fails to hear a loud call from above water. Similarly, as a good fisherman can testify, fish are not disturbed by any but the loudest sounds produced above water, but instantly respond to sounds produced in the water, such as splashing.

To overcome the air–water sound barrier so that airborne sound could reach the fluid-filled inner ear, the no-longer-needed first gill cleft became the external ear canal and the middle-ear cavity, with a thin membrane in between, which is the eardrum membrane. The gill that connects with the fish's throat persists as the eustachian tube from the upper part of the throat to the middle ear. On either side of each gill, cartilage grew that lent stiffness to them. Similarly, the cartilage of the first gill cleft in the human embryo grows outward to form the outer ear, which protrudes from each side of the head. The cartilage of the second gill cleft in the human embryo becomes transformed into the ossicular chain, the chain of tiny bones that carries sound vibrations from the eardrum membrane to the fluid of the inner ear. The first of these bones, the hammer (malleus), is attached to the eardrum membrane by its handle. Attached to the head of the hammer by a tiny joint is the second tiny bone, the anvil (incus). Joined to the incus by another, even tinier joint, the smallest in the body, is the smallest bone in the body, the stirrup (stapes) bone, with its footplate inserted into an oval-shaped opening called the oval window, leading to the inner ear.

The air–water sound barrier, which caused our earliest land-based ancestor to be virtually deaf, is overcome in the human embryo by the mechanical device of the eardrum membrane and ossicular chain, in the following manner:

The large area of the eardrum membrane vibrates in response to sound with movements that are fairly large, but with small force. The ossicular chain transmits these movements to the much smaller area of the footplate of the stirrup, decreasing the size of the movements but increasing their force manyfold. The ossicular chain itself adds appreciably to the force of the vibrations of the stapes by its lever effect, the handle of the hammer being longer than the anvil. As a result, sound vibrations at the footplate of the stirrup in the oval window of the human ear are made 22 times smaller—and stronger—than at the eardrum membrane.

The human ear, which develops in the embryo from two distinct beginnings, the early otocyst for perception of sound in the inner ear and the primitive gills to form the sound-transmitting outer and middle ear, is the only part of the body that reaches adult size and form long before birth. By the fifth month of fetal life, just a little beyond halfway between fertilization of the egg and birth, the otocyst, by a series of outpouchings and infoldings, has fully developed into the complicated fluid-filled labyrinth of the inner ear. First to appear are three semicircular canals, one for each plane in space, and the vestibule with an elongation ending in a pouch, called the endolymphatic sac. As early as the seventh to eighth week after fertilization, the nerve endings for the vestibular nerve of equilibrium first appear, for in fishes and earlier aquatic forms equilibrium is more important than hearing. The function of hearing developed later in evolution, so that the hearing-nerve endings do not develop until the 12th week of embryonic life. By the fifth month of fetal life the entire labyrinth of the inner ear, including the semicircular canals and vestibule for balance and the snail-shaped cochlea for hearing, are fully formed and of adult size. The more recently acquired outer and middle ear for sound transmission are not fully formed and ready to function until the seventh month, two months before the infant is born.

Vestibular and hearing nerve endings

Not only is the sense of hearing acquired later in evolution and later in the embryo than the sense of equilibrium, but the sense of hearing is more fragile and more easily injured by viruses and other infections or by blows on the head or by excessive stimulation by extremely loud noise. Thus, infants born with defective hearing generally have a well-functioning sense of equilibrium, and children and adults whose hearing nerves have been damaged by viruses or other infections generally retain normal nerves for equilibrium.

STRUCTURE AND FUNCTION OF THE EAR

As noted above, the ear has three main parts, the outer ear, the middle ear, and the inner ear. The outer ear, formed of flexible, delicately curved cartilage covered by skin, includes the outer ear canal, formed partly by cartilage and partly by bone. The outer ear canal extends about three centimetres (1.2 inches) from the outer surface of the adult head to the thin, transparent eardrum membrane. The skin of the outer ear canal is supplied with hairs and with glands that secrete a waxy substance called earwax, or cerumen. Both the hairs and the earwax function to discourage insects from making their home and laying their eggs in the ear canal.

The outer ear

The eardrum membrane, averaging one centimetre (0.4 inch) in diameter, is semitransparent like a piece of waxed paper and is stretched tightly like the head of a drum, so that it vibrates readily in response to sound waves. The handle of the hammer is firmly embedded in it, so that vibrations of the eardrum membrane are communicated to the hammer and then to the anvil and stirrup, through the tiny joints that connect them. It is the vibratory movements of the stirrup footplate in the oval window that create vibratory motion of the fluid filling the inner ear, needed to stimulate the hearing nerve.

The middle ear is a small, air-filled cavity in bone, about the size of an aspirin tablet, separated from the outer ear canal by the eardrum membrane. The eustachian tube connects the middle ear with the upper part of the throat, above the palate. It opens with each yawn or swallow to equalize air pressure on both sides of the eardrum membrane. This function is most noticeable with rapid change in altitude as in a high-speed elevator, driving over mountains, or in an airplane. Failure of the eustachian tube to open results in a full or pressure sensation in the ear and a noticeable, although at first slight, impairment in hearing. This is more likely to happen during a head cold or an allergic reaction when the membranes of the eustachian tube are swollen. If the eustachian tube remains closed, fluid accumulates in the middle-ear cavity, replacing the air and causing an increased impairment of hearing.

Across the middle-ear cavity extends the chain of three tiny bones, mentioned earlier, the auditory ossicles. *Os* is Latin for "bone," and "ossicle" means "little bone." From the middle-ear cavity into the mastoid bone that lies behind the outer ear there extends a series of air cells, spongelike but with rigid, bony walls. These, like the air sacs in the bones of birds, are to lighten the weight of the skull. Infection of the middle-ear cavity can extend into these air cells and produce a mastoid abscess.

The inner ear

The inner ear, or labyrinth, consists of a complicated series of passageways embedded in rock-hard bone, the hardest bone in the body and therefore called the petrous bone, "petrous" being Latin for "rocklike." The passageways of the labyrinth are filled with fluid containing the nerve endings of the hearing nerve (auditory nerve) and the nerve of equilibrium (vestibular nerve). The vestibular nerve supplies the three semicircular canals. At one end of each canal a spherical enlargement is filled by a gelatinous caplike structure, the cupula, attached to hairlike endings of the vestibular nerve. When the head turns in the plane of a semicircular canal, inertia of the fluid within it causes the cupula to be pushed to one side, pulling on the vestibular-nerve endings and stimulating them. This stimulus travels back to the brain along the vestibular nerve, producing a sense of rotation or even vertigo. If a person is rotated slowly and steadily, as on a piano stool, the fluid within the semicircular canal begins to rotate. If the piano stool is suddenly stopped, the fluid continues to rotate, pushing the cupula to one side, stimulating the nerve of equilibrium, and causing vertigo. Strong stimulation of the vestibular nerve of equilibrium causes nausea, sometimes vomiting, as well as the sense of rotation with inability to stand and walk straight.

The vestibular nerve of equilibrium also supplies two

structures in a space called the vestibule, each consisting of a gelatinous membrane in which are embedded tiny stones about the size of grains of sand that rest upon hairlike endings of the vestibular nerve. One of these membranes lies in the horizontal plane, one in the vertical plane. Pressure exerted by gravity on these tiny stones, called otoliths, pulls on the hairlike endings of the vestibular nerve. In this way, one's position in space is known even in the dark or with the eyes closed.

The labyrinth of the inner ear, in addition to the more primitive balancing mechanism of the semicircular canals and vestibule, includes the more recently acquired organ of hearing, located in the snail-shaped cochlea (Latin: "snail shell"). Within the cochlea is a marvellously contrived structure first described in 1851 by an Italian anatomist named Alfonso Corti; thus, it is known as the organ of Corti. It consists of a membrane, broad at one end of the cochlea and narrow and tightly stretched at the other end, known as the basilar membrane, on which rests a large number of hair cells containing the ends of the hearing nerve. On top of the hair cells and attached to their projecting hairs rests the tectorial membrane, similar in its gelatinous composition to the cupulas in the semicircular canals. Like the basilar membrane, the tectorial membrane is large at one end of the cochlea and tiny at the other end, with all gradations in between.

Thus, the basilar and tectorial membranes resemble the strings on a piano or harp. It has been shown that for each tone in the musical scale, from the lowest audible rumbling tone of ten vibrations per second to the shrill tone of 24,000 vibrations per second, the highest tone audible to a young human ear, a certain section of the basilar and tectorial membranes in the cochlea is set into maximum vibration. The hairlike endings of the auditory nerve at that point are pulled upon and stimulated, and the stimulus travels back to the brain, which "hears" that particular tone. There is still debate as to whether the basilar membrane or the tectorial membrane is the chief one that vibrates in response to a particular tone. Since both membranes are graduated in size, large at one end of the organ of Corti and tiny at the other end, it seems likely that a section of both membranes vibrates in response to a particular tone.

The cochlea of the inner ear with its organ of Corti not only has a wide range of response but also distinguishes minute pitch differences, as small as a quarter tone for the trained ear, and, in addition to differences in pitch, the ear can distinguish differences in loudness from the faintest audible sound to a sound vibration 1,000,000,000,000 times stronger.

The person with normal hearing has another ability, that of suppressing or ignoring a large amount of background noise and concentrating on one speaker. This requires binaural hearing with two ears of approximately equal sensitivity. Binaural hearing is also needed to determine the direction from which sound comes. One's own voice is heard, and one automatically monitors its quality and intensity, so that unconsciously the voice is raised in the presence of noise and lowered in quiet surroundings. The harsh and toneless quality of the completely or profoundly deaf is due to loss of this feedback system.

Binaural hearing

Loss of hearing on one side causes difficulty in understanding speech in the presence of noise and loss of ability to tell the direction from which sound comes. In quiet there is little handicap in one-eared hearing unless the speaker happens to be on the side of the deaf ear.

Loss of the balancing vestibular nerve on one side causes temporary imbalance that is soon compensated for in children, more slowly in adults; it may never be completely compensated for in older persons. Loss of both balancing nerves causes imbalance noticed mostly in the dark and walking on grass or a soft carpet, where the eyes and the sensations from the soles of the feet cannot assist in maintaining equilibrium. A person with such a disability walks unsteadily, with the feet and legs apart and the hands and arms partly extended ready to catch hold of supports. Children compensate quickly and well for loss of both vestibular nerves, except in the dark and on uneven terrain, while adults compensate much more slowly and imperfectly.

Diseases and disorders of the ear

THE OUTER EAR

Each of the three main parts of the ear is afflicted by a particular set of diseases, according to the structure, tissues, and function of that part. Diseases of the outer ear are those that afflict skin, cartilage, and the glands and hair follicles in the outer ear canal. The sound-transmitting function of the outer ear is impaired when the ear canal becomes filled with tumour, infected material, or earwax, so that sound cannot reach the eardrum membrane. The most common diseases of the outer ear are briefly described in the following paragraphs.

Frostbite. The exposed position of the outer ear makes it the part of the body most frequently involved by freezing or frostbite. Humidity, duration of exposure, and, most of all, wind, in addition to degrees of temperature below freezing, predispose to occurrence of frostbite. The frozen area begins along the upper and outer edge of the ear, which becomes yellow-white and waxy in appearance, cold and hard to the touch, and numb with loss of skin sensation.

In treatment of frostbite the victim is placed as soon as possible in a warm room, but the frozen ear is kept cool until the returning blood circulation gradually thaws the frozen part from within. Massage of the frozen ear is avoided, for it is likely to injure the skin. Heat applied to the frozen area before circulation is established can result in clotting of the blood in the blood vessels. This in turn can result in death of that part of the ear, which turns black and eventually falls off, a process called dry gangrene.

Hematoma. Injury to the outer ear can cause bleeding between the cartilage and the skin, producing a smooth, rounded, nontender purplish swelling called hematoma. The accumulation of clotted blood is removed by a surgeon because, if it is left, it will become transformed into scar tissue and cause a permanent, irregular thickening of the outer ear commonly called cauliflower ear and seen in boxers and wrestlers whose ears receive much abuse.

Infections of the outer ear

Perichondritis. Infection of the cartilage of the outer ear, called perichondritis, is unusual but may occur from injury or from swimming in polluted water. It is due to a particular micro-organism, *Pseudomonas aeruginosa.* There is a greenish or brownish, musty or foul-smelling discharge from the outer ear canal, while the affected outer ear becomes tender, dusky red, and two to three times its normal thickness. Prompt medical treatment is necessary to prevent permanent deformity of the outer ear.

External otitis. Infection of the outer ear canal by molds or various micro-organisms occurs especially in warm, humid climates and among swimmers. The ear canal itches and becomes tender; a small amount of thin, often foul-smelling material drains from it. If the canal becomes clogged by the swelling and drainage, hearing will be impaired. Careful and thorough cleaning of the outer ear canal by a physician, application of antiseptic or antibiotic eardrops, and avoidance of swimming are indicated to clear up the infection.

Boil in the ear (furuncle). Infection of a hair follicle anywhere on the body is known as a boil, or furuncle. This can occur in a hair follicle in the outer ear canal, especially when there is infection of the skin of the canal. It always occurs because of a particular type of germ known as staphylococcus. Because the skin of the ear canal is closely attached to the underlying cartilage, a boil in the ear canal is especially painful, with swelling, redness, and tenderness but generally without fever. Heat applied to the outer ear by a hot-water bottle or electric pad helps the infection to come to a head and begin to drain. Antiseptic eardrops and careful cleaning of the outer ear canal are needed to prevent other hair follicles from becoming infected with a series of painful boils in the ear.

Erysipelas of the outer ear. Erysipelas is an infection in the skin caused by a particular type of streptococcus that causes a slowly advancing red, slightly tender thickening of the skin. It begins at the ear and spreads to the face and neck. Centuries ago erysipelas epidemics caused severe and often fatal infections. In AD 1089 one of the most severe epidemics was known as St. Anthony's fire; those who prayed to St. Anthony were said to recover; others, who did not, died. Today erysipelas is a rather mild and comparatively rare infection that clears up rapidly when sulfanilamide is taken by mouth or penicillin by injection.

Leprosy. Leprosy, seen rarely outside of the tropics today, was another scourge of ancient times that sometimes affected the outer ear. It is caused by the leprosy bacillus, *Mycobacterium leprae*, which causes a painless, slowly progressing thickening and distortion of the affected tissues. The diagnosis is made by examining a bit of the infected tissue under a microscope and finding the leprosy bacilli, which in appearance are not unlike the bacilli that cause tuberculosis. Fortunately, the antibiotics effective against tuberculosis are effective today in arresting the ravages of leprosy.

Eczema. Eczema of the skin of the outer ear, like eczema elsewhere, is an itching, scaling redness, sometimes with weeping of the affected skin. It is often the result of an allergy to a food or substance such as hair spray that comes in contact with the skin. The best treatment is discovery and avoidance of the causative agent.

Impacted earwax. The waxy substance produced by glands in the skin of the outer ear canal normally is carried outward by slow migration of the outer layers of skin. When wax is produced too rapidly it can accumulate as a hard plug, firmly filling the outer ear canal and blocking the passage of sound to the eardrum membrane, causing a painless impairment of hearing. Large plugs of earwax need to be removed by a physician. Smaller amounts may be softened by a few drops of baby oil left in the ear overnight, then syringed out with warm water and a soft-rubber baby ear syringe.

Cancer of the outer ear. Cancer of the outer ear occurs chiefly in areas exposed for many years to the direct sun. A small and at first painless ulcer, with a dry scab covering it, that slowly enlarges and deepens may be a skin cancer. Removal of a small bit of tissue from the edge (biopsy) and examining it under a microscope comprise the method of diagnosis. Complete removal by surgery or properly applied irradiation is needed for a cure. Cancer that begins in the ear canal is more serious, for it may deepen into the bone before it is diagnosed. It is then more difficult to cure by removal. Cancers of the ear canal are rather rare, while cancers of the skin of the outer ear are more common, as well as more readily cured by removal.

Osteoma of the bony ear canal. Osteoma of the bony ear canal is a bony knob that grows close to the eardrum membrane, especially in those who swim a great deal in cold water. It is not dangerous and does not need to be removed unless the bony overgrowth becomes large enough to block the ear canal.

Cyst of the ear. A cyst is a sac filled with liquid or semisolid material. A cyst of the ear is most often caused by a gland that lubricates the skin behind the earlobe, less often at the entrance of the ear canal. If the duct of this gland becomes stopped, the lubricating fatty material accumulates as a soft rounded nodule in the skin. Infection of the cyst causes a tender abscess to form and drain. The cyst will re-form unless it is removed completely by surgery.

Another type of cyst occurs above the ear canal, just in front of the outer ear or, rarely, in the neck behind and below the ear. This is a remnant of the primitive gill of the early embryo, a reminder of our ancient fishy ancestors. It may appear as a tiny pitlike depression that discharges a little moisture from time to time, or a cystic swelling may develop when the opening of the pit is closed, requiring surgical removal.

Keloid of the ear. In dark-skinned people, overgrowth of scar tissue from any skin incision or injury causes a thickened elevation of the scar called a keloid. Having the earlobes pierced for earrings sometimes results in a large, painless nodular keloid enlargement of the earlobe, harmless but unsightly. Keloids are removed surgically (see also SKIN DISEASES).

Enlargement of earlobe

Absence of the outer ear. Congenital deformity or absence of the outer ear, usually on one side, sometimes on both, is often accompanied by absence of the outer ear canal. This failure of the primitive gill structures to become properly transformed into the normal outer and middle ear is, in rare instances, hereditary. More often it occurs for no known reason. In some cases it can be traced to the damaging effects on the embryo of rubella (German measles) in the mother during the first three months of her pregnancy. Since the inner ear and nerves of equilibrium and hearing come from the otocyst, separate from the gill structure, in most cases of deformed or absent outer ear the hearing nerve is normal. Surgical construction of a new ear canal and eardrum membrane can then often improve the hearing, which has been impaired by the failure of sound conduction to reach the hearing nerve in the inner ear.

Lop ear. Lop ear, excessive protrusion of the ear from the side of the head, is a more frequent and less serious deformity of the outer ear. Girls easily conceal the protruding ears by their hair. Boys may desire an operation to bring the ears back to a more normal and less conspicuous position.

THE MIDDLE EAR

Location of middle ear

The air-filled middle-ear cavity and the air cells in the mastoid bone that extend backward from it are located a third of the distance from the side of the head toward its centre. The brain cavity lies just above and behind the middle ear and mastoid air spaces, separated from them only by thin plates of bone. The nerve that supplies the muscles of expression in the face passes through the middle-ear cavity and mastoid bone; it, too, is separated from them by only a thin layer of bone. In some instances this bony covering is incomplete, so that the facial nerve lies directly against the mucous membrane that lines the middle ear and mastoid air cells. This mucous membrane, an extension of a similar mucus-producing membrane that lines the nose and upper part of the throat, extends all the way through the eustachian tube into the middle ear and mastoid. It is subject to the same allergic reactions and infections that afflict the nasal passages. Thus, an acute head cold or other infection of the nose and throat, such as measles or scarlet fever, may extend through the eustachian tube into the middle ear and mastoid air cells. The proximity of the brain cavity to the mastoid air cells is such that an infection, if severe and untreated, may lead to meningitis (inflammation of the covering of the brain) or brain abscess. The large vein that drains blood from the brain passes through the mastoid bone on its way to the jugular vein in the neck. Infection from the middle ear can extend to this vein, resulting in "blood poisoning" (infection of the bloodstream, also called septicemia). Paralysis of the facial nerve and infection extending from the middle ear to the labyrinth of the inner ear are other possible complications of middle-ear infection. All of these possibilities spring from the particular location of the small but important middle-ear cavity.

Acute middle-ear infection. Fortunately, acute middle-ear infections, called acute otitis media, are nearly always due to micro-organisms that respond quickly to antibiotics. As a result, acute infection of the mastoid air cells resulting in a dangerous mastoid abscess with the possibility of meningitis, brain abscess, septicemia, infection of the labyrinth, or facial nerve paralysis, complicating an acute infection of the middle-ear cavity, have become rare. Abscess of the mastoid and the other complications of acute middle-ear infection are seen chiefly in remote regions and countries lacking adequate medical attention.

While serious and life-threatening acute infections of the middle ear and mastoid air cells have become rare, chronic infections, mentioned below, continue to occur,

and another type of middle-ear disease, secretory otitis media, is frequent.

Secretory otitis media. In secretory otitis media the middle-ear cavity becomes filled with a clear, pale-yellowish, noninfected fluid. The disorder is the result of inadequate ventilation of the middle ear through the eustachian tube. The air in the middle ear, when it is no longer replenished through this tube, is gradually absorbed by the mucous membrane, fluid taking its place. Eventually, the middle-ear cavity is completely filled with fluid instead of air. The vibratory movements of the eardrum membrane and the ossicular chain are impeded by the fluid, with a painless impairment of hearing.

The usual causes for secretory otitis media are an acute head cold with swelling of the membranes of the eustachian tube; an allergic reaction of the membranes in the eustachian tube; and an enlarged adenoid (nodule of lymphoid tissue) blocking the entrance to the eustachian tube. The condition is cured by finding and removing the cause and then removing the fluid from the middle-ear cavity, if it does not disappear by itself within a week or two. Removal of the fluid requires puncturing the eardrum membrane and forcing air through the eustachian tube to blow out the fluid. In some cases a tiny plastic tube is inserted through the eardrum membrane to aid in re-establishing normal ventilation of the middle-ear cavity. After a time, when the middle ear and hearing have returned to normal, this plastic tube is removed. The small hole left in the eardrum membrane quickly heals.

Aero-otitis media. The sudden change of altitude by a rapid descent in a nonpressurized or poorly pressurized plane during a head cold or allergic reaction may not permit the normal equalization of air pressure that occurs by the periodic opening of the eustachian tube on swallowing or yawning. The eardrum membrane becomes sharply retracted when the air pressure becomes less within than without, while the opening of the tube into the upper part of the throat becomes pressed tightly together by the increased air pressure in the throat, so that the tube cannot be opened by swallowing. A severe sense of pressure in the ear is accompanied by pain and a decrease in hearing. Sometimes the eardrum membrane ruptures because of the difference in pressure on its two sides. More often, the pain continues until the middle ear fills with fluid, or surgical puncture of the eardrum membrane is done. This condition occurring from airplane flights is called aero-otitis media. Usually, however, pain and hearing loss produced during a flight is of a temporary nature and disappears of its own accord.

Chronic middle-ear infection. Chronic infection of the middle ear occurs when there is a permanent perforation of the eardrum membrane that allows dust, water, and germs from the outer air to gain access to the middle-ear cavity. This results in a chronic drainage from the middle ear through the outer ear canal. There are two distinct types of chronic middle-ear infection, one relatively harmless, the other caused by a dangerous bone-invading process that leads, when neglected, to serious complications.

The harmless type of chronic middle-ear disease is recognized by a stringy, odourless, mucoid discharge that comes from the surface of the mucous membrane that lines the middle ear. Medical treatment with applications of antiseptic solutions and powders is all that is needed to dry up the chronic drainage. The perforation in the eardrum membrane may then be closed, restoring the normal structure and function of the ear with recovery of hearing.

The dangerous type of chronic middle-ear drainage is recognized by its foul-smelling discharge, often scanty in amount, coming from a bone-invading process beneath the mucous membrane. Such cases are usually caused by a condition known as cholesteatoma of the middle ear. This is an ingrowth of skin from the outer ear canal that forms a cyst within the middle ear and mastoid. If untreated, the cyst enlarges slowly but progressively, gradually eroding the bone until the cyst reaches the brain cavity or the nerve that supplies the muscles of the face or a semicircular canal of the inner ear. The infected material within the cyst then produces a complication: meningitis or brain abscess, paralysis of the facial nerve, or infection of the labyrinth of the inner ear with vertigo, often leading to total deafness.

Dangers of neglected cholesteatoma

Fortunately, cholesteatoma of the middle ear is now rarely so neglected as to permit development of a serious complication. By careful examination of the eardrum-membrane perforation and by X-ray studies, the bone-eroding cyst can be diagnosed; it can then be removed surgically before it has caused serious harm. This operation is known as a radical mastoid or a modified radical mastoid operation. If at the same procedure the perforation in the eardrum membrane is closed and the ossicular chain repaired, the operation is known as a tympanoplasty, or plastic reconstruction of the middle-ear cavity.

Ossicular interruption. The ossicular chain of three tiny bones needed to carry sound vibrations from the eardrum membrane to the fluid that fills the inner ear may be disrupted by infection or by a jarring blow on the head. Most often the separation occurs at its weakest point, where the anvil (incus) joins the stirrup (stapes). If the separation is partial there is a mild impairment of hearing; if it is complete there is a severe hearing loss. Testing of the hearing in such a case demonstrates that the nerve of hearing in the inner ear is functioning normally but that sound fails to be conducted from the eardrum membrane to the inner ear. The separated ossicles can be brought back together by repositioning, thus restoring the conduction of sound to the inner ear. This is one of the most successful of hearing-restoring operations.

Fixation of the stirrup by otosclerosis. The commonest cause for progressive hearing loss in early and middle adult life is a disease process of the hard shell of bone that surrounds the labyrinth of the inner ear. This disease of bone is known as otosclerosis, a name that is misleading, for in its early and actively expanding stage the nodule of diseased bone is softer than the ivory-hard bone that it replaces. The more appropriate name otospongiosis is sometimes used, but such is the tenacity of tradition that the older name, applied before the process was well understood, has persisted and is the term generally used.

The cause for the occurrence of the nodule of softened otosclerotic bone is unknown. There is a certain familial tendency, half the cases occurring in families in which one or several relatives have the same condition. It is one-tenth as common among Negroes as among whites, and is twice as common in women as in men. The nodule of softened otosclerotic bone first appears in late childhood or in early adult life. Fortunately, in most cases it remains quite small and harmless, producing no symptoms, and is discoverable only if the ear bones are removed after death and examined under a microscope. Such evidence indicates that approximately one in ten white adult men and one in five white adult women will be found to have such a nodule of otosclerotic bone by middle adult life.

In about 12 percent of cases of otosclerosis the nodule of softened bone becomes large enough to reach the oval window containing the footplate of the stirrup. Increasing pressure caused by the expanding nodule begins to impede its vibratory movements in response to sound striking the eardrum membrane. Gradually and insidiously, such an affected person begins to lose his sharpness of hearing. First he begins to lose the ability to hear faint sounds of low pitch; next he begins to have difficulty hearing the whispered voice; then he has difficulty in hearing conversation from a distance; and finally he can hear and understand the spoken voice only when it is quite loud or close to the ear. One of the characteristics of impaired hearing due to stirrup fixation by otosclerosis is retained ability to hear over the telephone by pressing the receiver against the head so that the sound is carried to the inner ear by bone conduction. Another characteristic of this type of impaired hearing is that hearing seems to be better while one is riding in an automobile,

in a plane, or on a train. The reason is that the low-pitched roar of motors causes normally hearing persons to unconsciously raise their voices, while the individual with stirrup fixation fails to hear the low-pitched roar and thus hears better and enjoys the raised voices around him.

Diagnosis of stirrup fixation

The diagnosis of stirrup fixation by otosclerosis is made on the basis of a history of a gradually increasing impairment of hearing with absence of any chronic infection of the middle ear or of perforation of the eardrum membrane and with hearing tests showing that the nerve of hearing in the inner ear is functioning but that sound fails to be conducted properly to it. The hearing tests demonstrate that the hearing by bone conduction is better than by air conduction.

The final and conclusive diagnosis of otosclerosis is made by surgical exploration and finding that the stirrup bone (stapes) is fixed and unable to be moved because of a nodule of bone that has grown against it. A special type of X-ray of the ear called polytomography is sometimes used to demonstrate that the footplate of the stirrup bone has been invaded by otosclerosis.

Fixation of the stirrup bone can be corrected surgically. This was accomplished formerly by constructing a new window into the inner ear to admit sound to the hearing nerve. This operation originated in Europe in 1924 and in 1937 was brought to the United States, where it was improved and named the fenestration operation. In 1952 it was found possible to mobilize (loosen) the fixed stirrup bone in some cases, thus restoring hearing without the need of constructing a new opening. In 1956 it was found that the fixed stirrup bone could be removed and replaced by a plastic or wire substitute in cases in which it could not be mobilized. Today this operation, known as stapedectomy, is the one most often used to correct fixation of the stapes (stirrup bone) by otosclerosis.

The otosclerotic bone disease in some cases expands as far as the cochlea of the inner ear, causing a gradual deterioration of the hearing nerve. This progressive nerve deafness may precede, accompany, or follow fixation of the stapes. In some cases it may occur without fixation of the stapes.

While the exact cause for the softening of a nodule of bone known as otosclerosis is not known, it may be associated in some cases with lack of fluoride in drinking water. There is evidence, not yet conclusive, that increasing the intake of fluoride may promote hardening of the softened nodule of otosclerotic bone, thus arresting or retarding its expansion. In this way it is possible that the gradual impairment of hearing-nerve function that often occurs with fixation of the stapes may be retarded or arrested.

THE INNER EAR

The labyrinth of the inner ear contains the nerve endings of the vestibular nerve—the nerve of equilibrium—and the auditory nerve, or nerve of hearing. The vestibular-nerve ends supply the semicircular canals and the otolithic membranes in the vestibule. The auditory nerve supplies the cochlea (see above *Structure and function of the ear*). Diseases of the labyrinth of the inner ear may affect both the vestibular nerve and the auditory nerve; or they may affect only the auditory nerve, with loss of hearing, or the vestibular nerve, bringing on vertigo. The commoner inner-ear diseases are touched upon in the following paragraphs.

Congenital nerve deafness. Congenital nerve deafness, a defect of the hearing nerve in the cochlea, may be present at birth or acquired during or soon after birth. Usually both inner ears are affected to a similar degree, and as a rule there is a severe impairment of hearing, although, in some cases of congenital nerve loss the impairment is moderate in degree. Many cases of congenital nerve deafness have been caused by the rubella (German measles) virus in the mother during the first three months of her pregnancy, causing arrest of development of the otocyst. This can happen during a rubella epidemic, even when the mother has no symptoms of the infection. In most cases the vestibular nerve is not affected or is affected to a lesser degree, and in most (but not all) cases the outer- and middle-ear structures are not affected. A vaccine against the rubella virus that has recently been introduced promises to result in a marked reduction in the number of cases of congenital nerve deafness if it is given to prospective mothers who have not had rubella before becoming pregnant and thus have not had an opportunity to build up immunity against infection when they are carrying a child.

Anoxia and kernicterus

Congenital nerve deafness, acquired at or soon after birth, may result from insufficient oxygen (anoxia) during a difficult and prolonged delivery or from the condition known as kernicterus, in which the baby becomes jaundiced because of incompatibility between its blood and that of the mother. In a few cases congenital nerve deafness is an inherited failure of the cochlea to develop properly. There is no medical or surgical treatment that can improve or restore hearing in cases of congenital nerve deafness. When the hearing loss is severe, speech cannot be acquired without special training. Children so afflicted must attend special classes or schools for the severely deafened, where they can be taught lipreading and speech. Electrical hearing aids can be helpful, especially during classes, to utilize the remnants of hearing usually present in such cases.

Viral nerve deafness. Virus infections can cause severe degrees of hearing-nerve loss in one ear and sometimes in both, at any age. The mumps virus is one of the most common causes of severe hearing-nerve loss in one ear. The measles and influenza viruses are less common causes. There is no effective medical or surgical treatment to restore hearing impaired by a virus.

Effect of ototoxic drugs. Ototoxic (ear-poisoning) drugs can cause temporary and sometimes permanent impairment of hearing-nerve function. Salicylates such as aspirin in large enough doses may cause ringing in the ears and then a temporary decrease in hearing that recovers when the person stops taking the drug. Quinine can have a similar effect but with a permanent impairment of hearing-nerve function in some cases. Certain antibiotics, such as streptomycin, dihydrostreptomycin, neomycin, and kanamycin, may cause permanent damage to the hearing-nerve function. The susceptibility to damage to the hearing from ototoxic drugs varies greatly among individuals. In most cases, except when streptomycin is the drug taken, the more durable and less easily damaged vestibular-nerve function is not affected. Streptomycin affects the vestibular nerve more than the auditory nerve.

Skull fracture and concussion. Skull fracture and concussion from a severe blow on the head can impair the functioning of the hearing nerve and of the nerve of balance in varying degrees. The greatest hearing loss arises when a fracture of the skull passes through the labyrinth of the inner ear, totally destroying its function.

Exposure to noise. Noise exposure, when the noise is excessive in loudness and duration, causes deterioration of the auditory nerve. A sudden explosive sound can tear the hair cells from the vibrating membranes in the organ of Corti, with a sudden, severe, permanent loss of hearing in that ear. Prolonged exposure to loud noise causes a gradually progressive impairment of auditory-nerve function, characterized by involvement first of that portion that responds to frequencies around 4,000 cycles per second. As the excessive noise exposure continues, the hearing-nerve loss increases in severity and broadens to include 2,000 and 8,000 cycles per second.

Hearing impairment after exposure to noise is especially common after gunfire explosion, in boilermakers, in stamp-press operators, in rivetters, and in workers around jet engines. The critical noise level that begins to damage the hearing nerve is around 100 decibels of loudness. Below 85 or 90 decibels few if any ears will be damaged. Above 100 decibels continued exposure will cause a significant slow impairment of hearing-nerve function in most persons, with varying degrees, depending on one's susceptibility. At levels of 130 decibels or above, such as are produced by jet engines or modern powerful amplifiers often used by rock and roll bands,

no ears are safe, and the deterioration of hearing is certain and may be rapid.

Noise-exposure deafness is preventable by avoiding exposure or by the wearing of protective earplugs or earmuffs. Cotton worn in the ears is of little help.

Causes of labyrinthitis

Labyrinthitis. Labyrinthitis, an inflammation of the labyrinth of the inner ear, occurs when micro-organisms enter as a result of meningitis, syphilis, acute otitis media and mastoiditis, or chronic otitis media and cholesteatoma. Loss of both equilibrium and hearing occurs in the affected ear. Prompt treatment sometimes arrests the damage with the possibility of partial recovery of the function of the inner ear.

Acoustic neuroma. An acoustic neuroma is a benign tumour that grows on the acoustic nerve near the point where it enters the labyrinth of the inner ear. The tumour causes gradual and progressive loss of auditory- and vestibular-nerve function on one side. Eventually the tumour grows out into the brain cavity, causing headache and paralysis. If it is not removed, blindness and death may result. Fortunately, acoustic neuroma can be diagnosed early and removed before it has serious consequences.

Ménière's disease. Ménière's disease, also called endolymphatic hydrops, is a fairly common involvement of the labyrinth of the inner ear that affects both the vestibular nerve, with resultant attacks of vertigo, and the auditory nerve, with impairment of hearing. It was first described in 1861 by a French physician, Prosper Ménière. It is now known that the symptoms are caused by an excess of endolymphatic fluid in the inner ear. The diagnosis is made from the recurring attacks of vertigo, often with nausea and vomiting, ringing or roaring in the ear, impairment of hearing with a distortion of sound in the affected ear that fluctuates in degree, and a sense of fullness or pressure in the ear. The cause of Ménière's disease is not always known, although in many cases it results from defective functioning of the endolymphatic duct and sac, the structures that normally resorb endolymphatic fluid from the inner ear as fast as it is produced. The treatment of Ménière's disease is directed toward controlling the excess of endolymphatic fluid. If medical treatment does not relieve the repeated attacks of vertigo, surgery may be necessary.

Presbycusis. Presbycusis, the gradual decline of hearing-nerve function of old age, is similar to other aging processes because it affects some people more rapidly and at an earlier age than others. Usually the slow diminishing of hearing does not begin until after age 60. The affected individual notices increasing difficulty in hearing sounds of high pitch and in understanding conversation. There is no medical or surgical treatment that can restore hearing in uncomplicated presbycusis. The physician must make certain that the patient does not have a correctible impairment, such as accumulated earwax, secretory otitis media, or stirrup fixation by otosclerosis, as part of his difficulty. An electrical hearing aid is of limited help to some, while others find that a hearing aid makes voices louder but less clear and therefore is of little help. Lipreading must then be used.

DEAFNESS AND IMPAIRED HEARING

Causes. Impaired hearing is, with rare exception, the result of disease or abnormality of the outer, middle, or inner ear, located in the temporal bone of the skull. One rare exception has already been mentioned, loss of hearing, usually on one side only, caused by a tumour, acoustic neuroma, on the acoustic nerve; an even rarer exception is impairment because of a brain lesion. Since nerve fibres from each ear go to both sides of the brain, a brain tumour or stroke causing paralysis of one side of the body rarely affects hearing.

Serious impairment of hearing (deafness) at birth is nearly always of the nerve type and cannot be improved by medical or surgical treatment. Nerve deafness arising from rubella infection of the mother during the first three months of her pregnancy, from insufficient oxygen during birth, and from kernicterus have already been mentioned.

Causes in childhood

In early and late childhood the most frequent cause for impaired hearing is poor functioning of the eustachian tubes with the accumulation of a clear, pale-yellowish fluid in the middle-ear cavity, known as serous or secretory otitis media. The vibratory movements of the tympanic (eardrum) membrane and ossicular chain in response to sound are impeded by the fluid, causing a moderate loss of hearing. Since this hearing loss is due to impairment of sound conduction to the inner ear, it is known as a conductive loss. In these children, evacuation of the fluid from the middle ear and its replacement with air restores normal sound conduction and normal hearing. The enlarged adenoid or allergic swelling of the membrane lining the eustachian tube needs to be corrected to prevent recurrences of secretory otitis media.

In early and middle adult life the usual cause for progressive impairment of hearing is otosclerosis, the disease process of bone surrounding the inner ear that has been described above. The conductive hearing loss caused by fixation of the stirrup bone (stapes) can be relieved by surgical mobilization of the stirrup or its replacement with a fine stainless-steel wire. In some cases the otosclerotic bone nodule grows inward to the inner ear, resulting in a gradual loss of hearing-nerve function. This may occur at the same time as the fixation of the stapes; it may precede stapes fixation; or it may occur some years after fixation of the stirrup bone, whether or not the stapes has been successfully operated upon. Hearing-nerve deterioration due to otosclerosis or any other cause cannot be improved surgically or by medical treatment. The progressive loss of hearing caused by the enlarging otosclerotic nodule may stop after a time, if the nodule becomes matured and inactive. Treatment of the patient with sodium fluoride tablets taken by mouth is a promising method for promoting inactivation of an active otosclerotic process but has not yet been proved conclusively to be of value.

The usual cause of impaired hearing after the age of 60 is presbycusis, the normal aging of the hearing nerve in the inner ear, for which there is no effective medical or surgical treatment.

Other varieties of ear disease causing impaired hearing have been mentioned above in the sections on diseases of the ear. In most cases when loss of hearing is of the conductive type surgical restoration of useful hearing by correcting the defect in the outer or middle ear is a possibility. When loss of hearing is of the sound-perceiving (nerve) type, surgical restoration cannot be expected. Medical treatment for nerve types of hearing loss is helpful only in rare cases when the loss is due to syphilis and in some early cases of Ménière's disease.

More important than cure for nerve types of hearing loss is prevention. Preventable especially are cases of deafness in the newborn due to rubella in the mother. Excessive and prolonged noise-exposure nerve deafness is preventable by early detection (routine testing of the hearing of persons engaged in noisy occupations), by a change of occupation, or by the wearing of ear protectors, either specially designed earplugs or earmuffs.

Incidence of impaired hearing

The incidence of impaired hearing in the general population depends upon the degree of hearing loss defined as impaired. In 1969 approximately 40,000 children with severe nerve deafness that occurred before speech had been acquired attended preschool classes in the United States.

By age six, 0.2 percent of all children have impaired hearing in one or both ears sufficient to warrant consultation of an ear specialist (otologist). By age 18 the number of children with loss of hearing sufficient to require diagnostic examination reaches 2.5 to 3 percent. By age 65 the number of adults with a recognizable hearing impairment reaches 5 percent. Beyond 65 the incidence of impaired hearing rises rapidly as presbycusis, the normal aging of the hearing nerve takes its toll. In the mid-1970s at least 10,000,000 individuals in the U.S. suffered from a hearing handicap sufficient to interfere with normal communication, and at least an equal number had milder and noticeable impairment in one or both ears.

Comparable figures from Britain show that one in six

persons is estimated to have some hearing difficulty, but only a quarter of these have any real handicap, with a third of this latter group needing hearing aids and one in 20 being deaf to all speech and beyond useful help with a hearing aid. With British children, one in 1,000 is severely deaf and as high as seven per thousand estimated to have sufficient impairment to need some form of help.

Rehabilitation. The child born deaf or with a severe hearing impairment cannot acquire speech by the normal process. He must attend special classes or a school for the deaf to be taught speech and lipreading. Most of these children have remnants of the sense of hearing that can be utilized in their schooling by the use of aids to amplify sound. The child with a moderate or mild hearing impairment is able to acquire speech by himself but a little more slowly than the child with normal hearing, while speech-correction instruction is usually required to improve his diction.

Conductive types of hearing loss are well compensated for by an electrical hearing aid that simply amplifies sound so that it can be transmitted by bone conduction to the inner ear and the normal nerve of hearing. Nerve types of hearing loss are often associated with distortion of sound and loss of intelligibility; hence, amplification with a hearing aid may be of no value or of limited value. In certain cases of the nerve type of hearing impairment, however, the clarity and intelligibility of hearing remains good, and amplification with an electrical hearing aid is satisfactory.

Lipreading

Lipreading, which actually entails observation of the entire facial expression rather than the movements of the lips alone, is utilized even by persons with normal hearing who, in the presence of background noise, need these visual clues to supplement hearing. As hearing begins to be impaired, lipreading, better termed speechreading, becomes increasingly valuable and important.

The hearing-impaired individual can learn lipreading skills by careful observation of the speaker at all times and by watching television with the sound volume turned down so that he needs to use his eyes to supplement his ears. Formal instruction in lipreading by a teacher individually or in classes is of additional help. The greater the loss of hearing, the more essential becomes lipreading, for which good lighting is essential.

Speech-correction instruction, needed for the young with serious degrees of impaired hearing, also becomes necessary for the adult who suddenly loses all hearing in both ears. Without the monitoring effect of hearing his own voice, his speech begins to deteriorate and to acquire the flat, toneless quality of the profoundly deaf.

Social and economic handicaps. The social handicap of severe degrees of hearing impairment is particularly important to the individual and his family. An individual who has a hearing impairment can feel isolated and embarrassed by being unable to join into group conversations. He therefore tends to withdraw within himself and often develops false ideas that others are talking about him and ridiculing him. No one has more poignantly written of this isolation than Beethoven, who is generally believed to have suffered from otosclerosis. Had he lived today, his hearing might have been restored or improved through surgery. As it was, it gradually became worse until, after conducting the first performance of his ninth and last symphony, he had to be told to turn around to face the audience to acknowledge the tremendous applause because he could not hear it.

Small devices, such as installing a buzzer instead of a high-pitched door and telephone bell for the nerve-deafened person who cannot hear tones of high pitch, using an amplifier on the telephone, and being patient when communicating with the hard-of-hearing person, make life easier both for him and for those around him.

The economic handicap of impaired hearing is in proportion to the degree of loss. Nevertheless, persons with a hearing disability, while unable to engage in jobs that require keen hearing, tend to have better records of dependability and fewer days away because of illness than persons with normal hearing.

BIBLIOGRAPHY. G.E. SHAMBAUGH, "A Restudy of the Minute Anatomy of Structures in the Cochlea with Conclusions Bearing on the Solution of the Problem of Tone Perception," *Amer. J. Anat.*, 7:245–257 (1907), the first detailed description of the hearing nerve end-organ in the cochlea, where sound waves are converted into nerve impulses depending upon the pitch of the tone; GEORG VON BEKESY, "The Ear," *Scient. Am.*, 197:66–78 (1957), a description in lay terms of the mechanism of hearing by today's foremost research authority on the ear; G.E. SHAMBAUGH, JR., *Surgery of the Ear*, 2nd ed. (1967), a well-illustrated text on diseases of the ear and their surgical correction; and with A. PETROVIC, "Effects of Sodium Fluoride on Bone," *J.A.M.A.*, 204:969–973 (1968), a summary of recent research on the arrest of progressive deafness due to otosclerosis by means of sodium fluoride; PHILIP H. BEALES, *Noise, Hearing and Deafness* (1965), a useful review, written in lay language, of the problem of deafness and the adverse influence on hearing of excess noise exposure.

(G.E.S.)